JCTC Journal of Chemical Theory and Computation

Transition-State Theory, Dynamics, and Narrow Time Scale Separation in the Rate-Promoting Vibrations Model of Enzyme Catalysis

Baron Peters

Departments of Chemical Engineering and Chemistry and Biochemistry, University of California, Santa Barbara, California 93106

Received January 26, 2010

Abstract: The power of transition-state theory (TST) for understanding enzymes is evidenced by its recent use in the design and synthesis of highly active de novo enzymes. However, dynamics can influence reaction kinetics, and some studies of rate-promoting vibrations even claim that dynamical theories instead of TST are needed to understand enzymatic reaction mechanisms. For the rate-promoting vibration (RPV) model of enzyme catalysis [Antoniou et al., *J. Chem. Phys.* **2004**, *121*, 6442], a reactive flux correlation function analysis shows that dynamical effects do slow the kinetics. However, the RPV model also shows extremely long-lived correlations because the RPV and the bath are not directly coupled. Additionally, earlier studies of the RPV model show a narrow time scale separation due to a small 5kT barrier. Thus earlier findings based on the RPV model may have little bearing on the properties of real enzymes. The intrinsic reaction coordinate (IRC) reveals that the RPV is an important component of the reaction coordinate at early and late stages of the pathway, but the RPV is not an important component of the reaction coordinate direction at the transition state. The unstable eigenmode from harmonic TST (which coincides with the IRC at the saddle point) gives a larger transmission coefficient than the coordinate used in the correlation functions of Antoniou et al. Thus while TST cannot predict the transmission coefficient, the RPV model suggests that TST can provide mechanistic insights on elementary steps in enzyme catalysis. Finally, we propose a method for using the transition-state ensemble as predicted from harmonic TST to distinguish promoting vibrations from other more mundane bath variables.

Introduction

Enzymes are remarkably active and selective catalysts.[1–3] Their catalytic activity at room temperature and mild pH provide exciting alternatives to processes using comparatively harsh conditions.[4] Because enzymes catalyze elementary reactions that break and make bonds much stronger than kT, many investigators use harmonic transition-state theory (TST),[5–8] variational TST,[9,10] and related theories[11–14] to analyze enzymatic catalysis mechanisms. These theories quantify how the enzyme enhances the rate by lowering an activation barrier through electrostatic,[15,16] covalent,[17] or other factors[18–21] that influence the thermodynamics. Mixed quantum mechanics/molecular mechanics (QM/MM)[14,15,22,23] calculations and empirical valence bond (EVB) models[24] are frequently combined with TST to understand these important aspects of enzymatic reaction mechanisms. Tunneling is also important for enzymatic reactions that involve proton or hydrogen transfer.[7,21,25–30] These methods and theories show extraordinary predictive abilities.[31,32] For a striking example, these approaches have recently been used to create highly active *de novo* enzymes.[33,34]

For some activated processes, dynamical effects are also important.[35–41] For many bond-making and -breaking reactions, these effects can be treated as secondary corrections to the TST rate.[32,42–44] However, some recent computational studies note problems with TST[45] and suggest that dynamical theories are essential for understanding enzymatic reaction mechanisms.[46–49] Of particular interest in this work is the

10.1021/ct100051a © 2010 American Chemical Society
Published on Web 04/09/2010

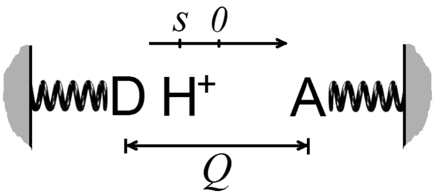


Figure 1. Mechanical model of the rate-promoting vibrations in enzymatic proton transfer reactions. The barrier for proton transfer depends on the donor (D) and acceptor (A) distance.

rate-promoting vibrations (RPV) model of enzyme catalysis.[46] This paper highlights some surprising new findings on dynamics of barrier crossing in the RPV model. We find that parameters used in previous implementations[46] give only a narrow time scale separation and long-lived correlations in the dynamics of the rate-promoting mode. We provide evidence that a lack of direct coupling between the bath and the rate-promoting mode can lead to energy diffusion limitations[5] that may not resemble dynamics in real enzymes. We also compare the unstable eigenmode coordinate from harmonic TST to the reaction coordinates investigated by Antoniou and Schwartz. Our results confirm that TST can be used to understand enzymatic reaction mechanisms that involve promoting vibrations, but perhaps not to the extent which dynamical effects will reduce the rate constant. We conclude by suggesting a simple procedure to identify promoting variables using only the information that is available from a TST analysis.

Dynamics in the Rate-Promoting Vibrations Model

The rate-promoting vibrations (RPV) model of enzyme catalysis[21,46] includes two key variables, s and Q. Q represents donor−acceptor distance, and s corresponds to the position of a proton along the path between donor and acceptor. Figure 1 depicts a mechanical model of the donor and acceptor atoms in the RPV model. The model potential as a function of s and Q includes a quartic bistable barrier of height V_0, a harmonic "promoting vibration" coordinate Q with frequency ω_Q, a term that couples s and Q, and a bilinear coupling between s and bath modes $\boldsymbol{q}$:

$$V = V_0(1 + s^4 - 2s^2) + c(s^2 - 1)Q + \frac{1}{2}m_Q\omega_Q^2Q^2 + \frac{1}{2}\sum_{k=1} m_k\omega_k^2\left(\boldsymbol{q}_k - \frac{c_k s}{m_k\omega_k^2}\right)^2 \quad (1)$$

The harmonic bath has a Debye frequency distribution.[46] The minima of V are at $(s, Q) = (\pm1, 0)$ regardless of ω_Q, but the saddle point $(s^*, Q^*) = (0, c/m_Q\omega_Q^2)$ shifts to higher Q values as ω_Q decreases.

Antoniou et al.[46] used transition-path sampling (TPS)[50–52] to study the RPV model. Their path ensemble includes only those 4 ps trajectories that spend the first 1 ps and last 2 ps entirely within the reactant and product basins, respectively.[46] Antoniou et al. report a dynamical rate-promoting effect that depends nonmonotonically on ω_Q.[46] The parameters used by Antoniou et al. were $V_0 = 6$ kcal/mol, $c = (V_0 m_Q\omega_Q^2)^{1/2}$, $T = 300$ K, $m_Q = 12$ amu, and $m_S = 1$ amu.[46] They investigated promoting vibrational frequencies of $\omega_Q = 100$, 300, and 900 cm^{-1}. Their parameters give a saddle point at $V_0/2$ above the two minima regardless of ω_Q. Their saddle is thus only 5kT above the minima, pushing the limits of the minimal time scale separation that is required for defining a rate constant.[5] Here, we also report new results for the model using a higher 8kT barrier by setting $V_0 = 9.6$ kcal/mol. The masses associated with bath modes were not specified in the original study, but following the convention in later work by Antoniou and Schwartz,[53] we set all bath masses equal to the mass of the promoting vibration. Our analysis of the dynamics is based on the correlation function:[37]

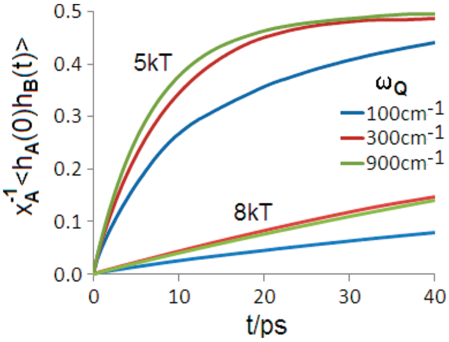


Figure 2. The correlation function $C(t)$ for three values of the rate-promoting vibration frequency at two barrier heights, 5 and 8kT.

$$C(t) = x_A^{-1} < h_A(0)h_B(t) > \quad (2)$$

where $h_A(t) = 1$ if $s(t) > 0$ and $h_A(t) = 0$, otherwise, $h_B(t) = 1$ if $s(t) < 0$ and $h_B(t) = 0$ otherwise, and $x_A = \langle h_A(t)\rangle$. For each condition (six total from two barrier heights and three frequencies), we computed the correlation functions from a single 200 ns trajectory. Figure 2 shows the correlation function $C(t)$.

The derivative dC/dt should decay from the TST rate constant to a plateau at the dynamically correct rate constant after a short a molecular relaxation time.[37] The lack of a plateau indicates a narrow or perhaps nonexistent time scale separation.[5] When time scale separation breaks down, the rate constant no longer exists because the waiting time for the next reactive event becomes dependent on the detailed initial condition in phase space. A theoretically motivated model[37,54] for the reactive flux correlation function is

$$\mathrm{d}C/\mathrm{d}t = x_B[(\tau_{TST}^{-1} - \tau_{rxn}^{-1})\exp(-t/\tau_{mol}^{-1}) + \tau_{rxn}^{-1}\exp(-t/\tau_{rxn})] \quad (3)$$

where $\tau_{TST}^{-1} = k_{TST}/x_B$, $\tau_{rxn}^{-1} = k/x_B$, $x_B = 1 - x_A$, k_{TST} is the TST rate constant from states A to B, and τ_{mol} is the (short) time required to commit to a basin from a typical initial condition on the dividing surface between states A and B. Despite the lack of a proper plateau in dC/dt, in some cases τ_{rxn}^{-1} may still be identified from a best fit of dC/dt to the model in eq 3.

Figure 3 shows dC/dt for the 5kT and the 8kT barriers at each of the promoting vibration frequencies. The 8kT barrier

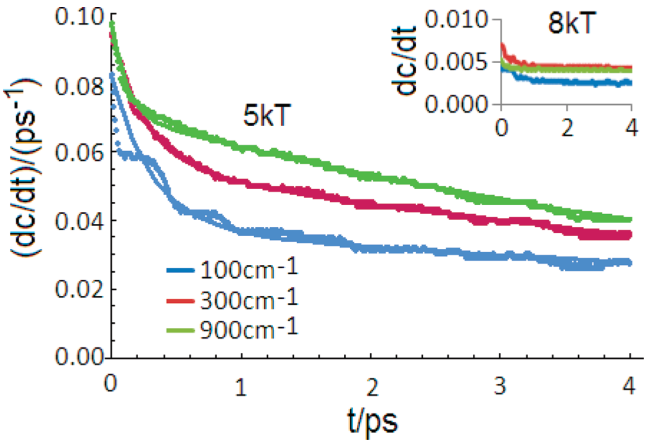


Figure 3. The reactive flux correlation function d*C*/d*t* for a barrier of 5kT and of 8kT at each of the three promoting vibrational frequencies. For the 5kT barrier, the time scales are not sufficiently separated to show a clear plateau, so least-squares fits of eq 2 are also shown as smooth curves behind the data. The fit provides an estimate of the reaction and molecular relaxation time scales.

Table 1. Least-Squares Fit Parameters for the Double Exponential Model of the Reactive Flux Correlation Function[a]

ω_Q	τ_{TST}/ps	τ_{RXN}/ps	τ_{MOL}/ps	test
100 cm^{-1}	6.10	13.38	0.314	0.014
300 cm^{-1}	5.32	8.96	0.272	0.010
900 cm^{-1}	5.15	7.10	0.117	0.003

[a] The last column 'test' shows the left-hand side of the inequality (3), which should be much smaller than unity to interpret τ_{rxn} as an inverse rate constant.

gives a clear plateau in d*C*/d*t*, but the 5kT barrier does not. However, d*C*/d*t* for the 5kT barrier still suggests two time scales: a molecular relaxation time scale shorter than 1 ps and a longer "reaction time scale" that depends strongly on the barrier height. Figure 3 also shows a least-squares fit of the model in eq 3 to the reactive flux correlation function for the 5kT barrier. The reaction time scale is on the order of 100 ps for the 8kT barrier but only about 10 ps for the 5kT barrier. For the 5kT barrier, the molecular relaxation and the reaction time scales are narrowly separated by just over a single order of magnitude. The parameters from a least-squares fit of the correlation function data to eq 3 are given in Table 1.

The fit parameter τ_{rxn}^{-1} will only yield a meaningful rate constant if the flux resulting from the fast molecular relaxation is much less than the flux from the more slowly decaying exponential. Equivalently, for $C(t)$, we require that nearly all of the relaxation from zero to x_{B} occurs after τ_{mol}. For poor dividing surfaces or small barriers where states near the dividing surface are significantly populated at equilibrium, much of the decay in $C(t)$ may occur during the initial time τ_{mol}. Separately integrating the two flux contributions from the model in eq 3 gives the requirement that

$$\left(\frac{\tau_{\mathrm{mol}}}{\tau_{\mathrm{TST}}} - \frac{\tau_{\mathrm{mol}}}{\tau_{\mathrm{rxn}}}\right)x_{\mathrm{B}} << 1 \tag{4}$$

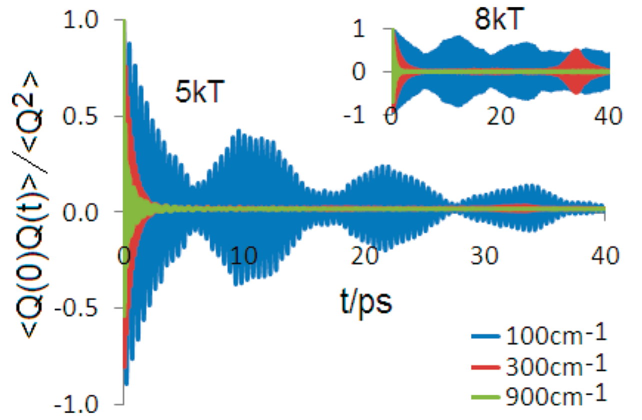


Figure 4. Autocorrelation functions for the rate-promoting vibration *Q*(*t*). The inset shows the autocorrelation function when the barrier is 8kT, and the main plot is for a 5kT barrier.

A subtle feature in the correlation functions is the distinctly nonexponential bump around $t = 10$ ps in $C(t)$ for a 100 cm^{-1} promoting vibration and the 5kT barrier. Figure 4 shows the autocorrelation function for the promoting variable $Q(t)$ to help understand the origin of the nonexponential feature. The autocorrelation function reveals long-lived excitations and surprising "beats" in the dynamics of Q.

For both barrier heights at $\omega_Q = 100$ cm^{-1}, the time for the autocorrelation of $Q(t)$ to decay is similar to the time scale for $C(t)$ to relax. This is consistent with the coupling in the RPV model: Q is coupled only to s, and the coupling is strongest when the (s, Q) subsystem has enough potential energy to cross the barrier. To verify this feature of the dynamics, note that the equation for Q:

$$m_Q\ddot{Q} = -m_Q\omega_Q^2 Q - c(s^2 - 1) \tag{5}$$

can be solved for any initial conditions in Q and for any trajectory $s(t)$:

$$Q(t) = Q_0\cos[\omega_Q t] + \frac{\dot{Q}_0}{\omega_Q}\sin[\omega_Q t] - \frac{c}{m_Q\omega_Q^2}(s^2(t) - 1) + \frac{c}{m_Q\omega_Q^2}\int_0^t \cos[\omega_Q(t-\tau)]2s(\tau)\dot{s}(\tau)\mathrm{d}\tau \tag{6}$$

The solution in eq 6 can give beats when $s(t)$ contains frequencies commensurate with ω_Q. In a real system, Q would be unlikely to show weak coupling artifacts, and thus the RPV model might be improved by directly coupling Q to the bath. In an earlier analysis, Caratzoulas et al.[55] showed how a simple Markovian friction could be added to the dynamics of Q.

Later work by Antoniou and Schwartz[53] used 10 times fewer bath modes with the same individual coupling strengths as in their 2004 study. Again, in the later study, the rate-promoting vibration Q was not directly coupled to a bath. The smaller bath also resulted in significantly weaker coupling to the coordinate s. The section below will show that the coupling in the later study is sufficiently weak that energy diffusion limitations begin to appear.

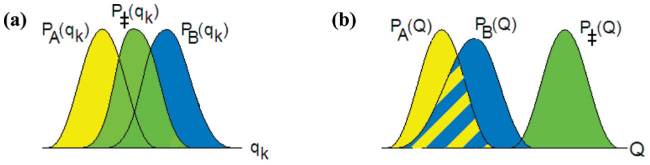


Figure 5. (a) Bath-mode distributions in the transition-state ensemble “interpolate” between the distributions in the reactant and product states. (b) In qualitative contrast, the transition-state ensemble distribution of a promoting variable Q like that of the RPV model does not interpolate between the corresponding reactant and the product distributions. Furthermore, the transition- and stable-state distributions of Q have a small overlap. These characteristics indicate that Q is involved in early and late stages of the reaction coordinate, even when it appears to be uninvolved in the reaction coordinate at the transition state.

Reaction Coordinate

Antoniou and Schwartz[53] performed extensive analyses of possible reaction coordinates for the RPV model. Much of the focus in earlier investigations was to develop strategies to identify special “rate-promoting vibrations” that might participate in the reaction coordinate among more mundane bath variables.[53,55] Two strategies were proposed: one based on the dynamical signature of a rate-promoting vibration in the flux correlation functions[55] and another strategy based on identifying coordinates, whose distributions have a narrow variance in the transition-state ensemble.[53]

The procedure of Caratzoulas et al. is a viable approach for detecting the presence of a rate-promoting vibration, but it does not identify the specific rate-promoting mode.[55] This discussion primarily addresses the later approach of Antoniou and Schwartz. They compute committor probability estimates at points along the harvested transition paths to collect a sample of transition states.[53] They then project that sample onto trial coordinates, e.g., onto s, Q, and various bath modes.[53] Finally, they identify coordinates that give the narrowest projected distribution as coordinates that are important in the reaction coordinate.[53] There are some reasons to suspect this procedure will not work for real systems.

First, it is questionable whether distributions from pairs of coordinates with different units in a real system can be meaningfully compared. Second, Antoniou and Schwartz extensively discuss the *width* of the Q-distribution in the sample of transition states,[53] but it is *the lack of overlap* between the transition-state and equilibrium distributions of Q that most clearly implicates Q in the pre- and reorganization stages. [Antoniou and Schwartz have switched the labels for Figure 1b and c in their paper.][53] Figure 5 provides a schematic example of how overlaps can be used to identify promoting variables that are important at early and late stages of the reaction, but whose involvement may not be clear from the a narrow distribution in the transition-state ensemble.

Third, the strategy of Antoniou and Schwartz finds only the separatrix, i.e., the locus of transition states.[53] Similarly, the approach of Best and Hummer only optimizes the separatrix.[56] However, other strategies discussed in their paper[57–59] optimize the reaction coordinate at all stages of the reaction, including the separatrix.[57–60] An accurate reaction coordinate at all stages[61] is useful for constructing coarse-grained models of the barrier crossing dynamics[62] and for avoiding hysteresis effects[12] that may occur if the reaction coordinate poorly describes early and late stages of the reaction. Figure 6 provides an example of the hysteresis problem from a recent study of methane diffusion in natural gas hydrates.[63]

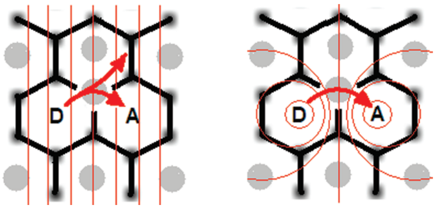


Figure 6. Schematic of methane (gray circles) diffusion by hopping through a water vacancy in a clathrate hydrate (black hexagonal lattice). Left: A calculation of the free energy barrier between donor (D) and acceptor (A) using hyperplane coordinates gave hysteresis because one of the two ways to proceed from the separatrix leads away from the vacant acceptor cage. Right: A bipolar coordinate system *with the same separatrix* gave no hysteresis because it better describes the early and late stages.

Antoniou and Schwartz[53] do not compare the computational cost of their procedure to other approaches. Here, we provide estimates and comparisons to the aimless shooting and likelihood maximization approach using information in their paper.

(1) The footnote in ref 13 of their paper[53] reveals that their implementation of transition-path sampling has an efficiency of less than 1%. The low efficiency results because their implementation of transition-path sampling was not optimized for studying dynamics at sharp saddle-point-type transition states. Aimless shooting can be tuned (by setting $\delta t = 1-2$ fs)[59] to more efficiently sample such sharp barriers.

(2) After computing transition paths, the authors identify 121 transition states by estimating the committor probability at configurations along the paths.[53] Each p_B estimate requires on the order of 100 trajectories.[52,61] Assuming that a small fraction of the points where p_B estimates were computed were found to be transition states, a reasonable estimate for the number of additional trajectories to identify the transition-state ensemble is more than 10 000. Likelihood maximization has been shown to identify a coordinate that is accurate at all stages in model systems[58,62] and in atomistic simulations[64–67] with approximately 1000 trajectories.

(3) Also note that the version of committor analysis used in refs 48 and 43 constrains multiple variables separately. For example, Antoniou and Schwartz[53] applied the two simultaneous constraints $s = 0$ and $Q = c/m_Q\omega_Q^2$. Their version of committor analysis with multiple constraints is less stringent than the usual committor analysis procedure.[52,68,69] Because of the extra constraints, the p_B histogram test does not reflect the accuracy of a true dividing surface through which an observable rate could be computed. We recommend the improved version of committor analysis with the

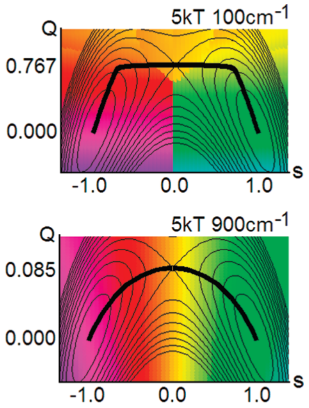


Figure 7. Steepest-descent path (heavy black curve) in mass-weighted coordinates with $m_Q = 12$ and $m_s = 1$ amu (proton). Also shown are contours of $V(s, Q)$ and the scalar arclength coordinate as a color-field background. Results are for $\omega_Q = 100$ cm^{-1} (above) and $\omega_Q = 900$ cm^{-1} (below) with parameters that give a 5kT barrier. Note how ω_Q changes $V(s, Q)$ and the arclength coordinate.

binomial-error deconvolution to isolate reaction coordinate error from committor estimate error.[61]

Aimless shooting and likelihood maximization have some quantitative and qualitative advantages over the approach of Antoniou and Schwartz, but for many enzymatic reactions, an accurate and efficient approach is to identify a saddle point on the energy landscape[70] and then to apply TST.[6] For reactions with sharp barriers that correspond to the breaking and making of chemical bonds, harmonic TST can often provide many mechanistic insights with a minimal computational expense.

Mechanistic Insights from Transition-State Theory

The unstable eigenmode from harmonic TST[5] and more generally the intrinsic reaction coordinate (IRC)[71–75] often provide excellent reaction coordinates for reactions that break and make strong chemical bonds.[32] Furthermore, the reaction path Hamiltonian,[76] constructed from the intrinsic arclength reaction coordinate s and minimum energy path in mass-weighted coordinates (MWC), can be used to understand the dynamics.[41] The potential energy surface in the reaction path Hamiltonian is a harmonic valley[76–80] around the minimum energy path. On a parabolic surface, mass weighted coordinates ($x_k = m_k^{1/2} q_k$ in the current study) transform the multidimensional equation "$F = ma$" into the simpler massless equation: $-\partial V/\partial \mathbf{x} = \mathrm{d}^2\mathbf{x}/\mathrm{d}t^2$. The multidimensional dynamics in mass-weighted coordinates can be simulated using Hamilton's equations from the reaction path Hamiltonian.[41,42]

The arclength coordinate in the (s, Q) subspace is sufficient to understand the role of Q in the reaction coordinate. Figure 7 shows the steepest-descent path, potential energy contours in the (s, Q) subspace, and the arclength coordinate as a colored background. Figure 7 also shows how the rate-promoting vibration frequency changes the role of Q in the arclength reaction coordinate. At very low frequencies, the reaction coordinate increases in the Q-direction, then decreases in the s-direction, and finally increases again as Q decreases back to $Q = 0$. At higher frequencies of the RPV, s retains its primary importance in the reaction coordinate at early and late stages and at the transition state.

For $\omega_Q = 100$ cm^{-1}, the mass weighted arclength coordinate reveals three distinct stages of the reaction: (1) donor−acceptor approach [preorganization], (2) proton transfer at constant Q [instanton], and (3) donor−acceptor recede [reorganization]. In agreement with Antoniou et al.,[46] a low frequency ω_Q may allow several recrossings of the $s = 0$ surface, while the donor−acceptor pair are close. Portions of the reaction coordinate that involve Q also explain additional recrossing at multiples of $(2\pi/\omega_Q)$ after the initial crossing because of slow energy transfer from Q. Note that the three stages are no longer distinguishable at $\omega_Q = 900$ cm^{-1}.

TST cannot predict the extent or importance of dynamical effects, but harmonic TST actually reveals many of the mechanistic details in the RPV model. The Q values on the dividing surface from harmonic TST are markedly different from the values that characterize both the reactant and product states ($Q \approx 0$). For small ω_Q, the saddle-point ($Q = Q^*$) and the minima locations ($Q = 0$) and the real frequencies for displacement along Q (ω_Q in all three locations) reveal a situation at small ω_Q that is much like that in Figure 5b. Thus harmonic TST can identify a role for Q in pre- and re-organization.

For an actual enzyme, saddle-point search algorithms[70] would be needed to find the saddle point. The Hessian matrix at the saddle point could then be computed explicitly or numerically projected onto a few relevant bond lengths,[12] depending on the model chemistry. In the RPV model, the saddle point is trivially identified. The mass-weighted Hessian matrix at the saddle point is

$$\partial^2 V_{\mathrm{mwc}} = \begin{bmatrix} m_s^{-1}(\partial^2 V/\partial s^2)\Big|_{\neq} & 0 & -m_s^{-1/2}\mathbf{c}^{\dagger}\mathbf{M}^{-1/2} \\ 0 & \omega_Q^2 & \mathbf{0} \\ -m_s^{-1/2}\mathbf{M}^{-1/2}\mathbf{c} & \mathbf{0} & \mathbf{\Omega}^2 \end{bmatrix} \tag{7}$$

where

$$\begin{aligned} \mathbf{M} &= \mathrm{diag}[m_1, m_2, \ldots, m_N] \\ \mathbf{\Omega} &= \mathrm{diag}[\omega_1, \omega_2, \ldots, \omega_N] \\ \mathbf{c}^{\dagger} &= (c_1, c_2, \ldots, c_N) \\ \frac{\partial^2 V}{\partial s^2}\Big|_{\neq} &= 2cQ^* - \frac{4V_0}{s_0^2} + \sum_k \frac{c_k^2}{m_k\omega_k^2} \end{aligned} \tag{8}$$

The unstable eigenmode *at the saddle point* has no contribution from the rate-promoting vibration. That is to be expected because of the symmetry of the potential energy surface. However, the unstable eigenmode is not perpendicular to the $s = 0$ plane. Bath modes influence the orientation of the unit normal vector on the dividing surface because of the off-diagonal coupling terms in the mass-weighted Hessian. To verify that the bath-mode contributions from harmonic TST do provide a more accurate reaction coordinate, we numerically diagonalized the mass-weighted Hessian for the exact parameters used by Antoniou and Schwartz.[53]

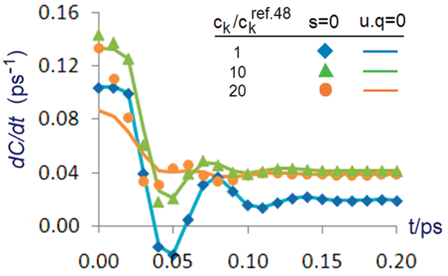


Figure 8. Reactive flux correlation function at different coupling strengths for the two dividing surfaces. The dividing surface **u**.**q** = 0 is always better than or equivalent to $s = 0$ in the variational sense. Parameters are those of Antoniou and Schwartz in ref 48.

In their later work, Antoniou and Schwartz changed the barrier in $V_0(s)$ to 6.3 kcal/mol, ω_Q to 110 cm^{-1}, m_Q to 48 amu, all bath masses to m_k = 48 amu, and the transfer distance from 2 to 1 Å, i.e., $s_0 = 0.5$ instead of $s_0 = 1$ Å.[53] Antoniou and Schwartz also provided the bath frequencies used in their later study.[53] Again, the rate-promoting vibration Q was not directly coupled to a bath. Importantly, Antoniou and Schwartz used 10 times fewer bath modes (100 in the later study[53] vs 1000 in the earlier study)[46] with the same individual coupling strengths (c_k = 0.00025 au.) in both cases. The smaller bath results in significantly weaker coupling to the coordinate s. We used these parameters of Antoniou and Schwartz in a 200 ns trajectory. A time step of 0.04 fs was sufficiently small that the energy drift was less than 0.001kT over the entire 200 ns trajectory.

Transmission coefficients were computed using the hyperplane dividing surface perpendicular to the unstable eigenmode. Denoting the unstable eigenmode as **u** and denoting the mass-weighted position relative to the saddle point as **q**, the dividing surface is **u**.**q** = 0, with positive values of **u**.**q** indicating products. The transmission coefficient from the unstable eigenmode coordinate was compared to the transmission coefficient using $s = 0$ as the dividing surface. Figure 8 shows that at the coupling strength used by Antoniou and Schwartz, the $t = 0^+$ limit of dC/dt (and therefore the transmission coefficients) for $s = 0$ and **u**.**q** = 0 dividing surfaces are indistinguishable. However, as coupling increases the two surfaces become more different, with the unstable eigenmode giving the superior dividing surface. Thus harmonic TST can quantitatively validate the suggestion of Antoniou and Schwartz, who used subtle changes in distribution widths to argue that small components of the 'bath modes' may actually be part of an optimal reaction coordinate.[31,53,81,82]

Note that the plateau at ~0.2 ps is not a true plateau. At longer times, dC/dt continues to decrease, and it becomes difficult to identify a plateau. This behavior is shown for four coupling strengths in Figure 9. At the coupling strength used by Antoniou and Schwartz, dC/dt is becoming too small to accurately compute from the data in our simulations.[83] Interestingly, the value of dC/dt at times beyond 2 ps is approximately a linearly *increasing* function of coupling strength up to c_k/c_k (ref 48) = 10.[36] Then at c_k/c_k (ref 48) = 20, the value of dC/dt begins to decrease again, reminiscent of Kramers' turnover.[36] However, the apparent turnover occurs at coupling strengths that are an order of magnitude stronger than the coupling used by Antoniou and Schwartz.[53] Also note the appearance of discrete steps in dC/dt. The times between steps correspond approximately to $2\pi/\omega_Q$, a single orbit of the weakly coupled rate-promoting vibration. These observations all suggest that action angle variables for energy diffusion-limited kinetics in the promoting variable Q may provide good reaction coordinates in this system. However, we emphasize that energy diffusion limitations may entirely vanish for a real system with direct coupling between Q and the bath.

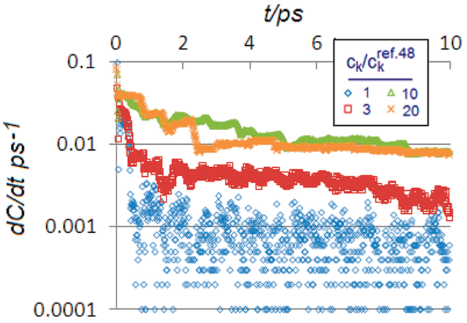


Figure 9. Using the dividing surface **u**.**q** = 0 in each case, dC/dt at longer times for four different coupling strengths. In Figure 8, dC/dt continues to decrease beyond the apparent plateau time. It is not clear whether a plateau in dC/dt can be identified. For the parameters of Antoniou and Schwartz, i.e., for c_k/c_k (ref 48) = 1, dC/dt rapidly decreases to a size where numerical errors exceed the physical value.

Conclusions

In the rate-promoting vibrations (RPV) model of enzyme catalysis, a 'promoting vibration' Q brings the donor and acceptor into proximity and lowers the barrier for motion along a 'proton-transfer coordinate' s. We studied dynamics in the RPV model using the parameters of Antoniou et al.[46] and also those of the later work by Antoniou and Schwartz.[53] The reactive flux correlation function for the RPV model shows that dynamical effects do slow the kinetics. However, the system shows an extremely narrow time scale separation because the parameters of Antoniou et al. give a barrier that is only 5kT high.[46] Additionally, we find strong correlations and surprising beats that persist in the dynamics of Q over times comparable to the reaction time scale. These unusual dynamical features result from a lack of direct coupling between Q and the bath in the RPV model. Near the reactant and product minima, Q also becomes effectively uncoupled from the proton-transfer coordinate s. Our findings suggest that direct coupling between Q and the bath may be needed to damp the beats and long-time correlations in the dynamics of Q.

Using the intrinsic reaction coordinate (IRC),[71–76] we show that Q is an important component of the reaction coordinate direction at early and late stages but not at the separatrix. Interestingly, the unstable eigenmode from harmonic TST (which coincides with the IRC at the saddle point) gives a larger transmission coefficient than the coordinate used in the correlation functions of Antoniou et al.[46,53] Additionally, harmonic TST is an important starting point for methods that compute the energy landscape along the IRC to quantify effects of dynamics[41] and tunneling.[84] Our results thus support the view that promoting modes are a form of preorganization[8,21] and contrast the view that TST cannot provide insight on enzymatic reaction mechanisms.[45]

Finally, the relative merits and efficiencies of the procedure of Antoniou and Schwartz[53] for identifying reaction coordinates were compared to other approaches. We discuss how special 'promoting variables' can be identified by comparing the distribution of coordinate values in the reactant and product states and in the transition-state ensemble. Bath variables show a distribution of values in the transition-state ensemble that approximately interpolates between the reactant and product distributions. The distribution of promoting variable values in the transition-state ensemble will neither overlap nor interpolate between the distributions of promoting variable values in the reactant and product states. In agreement with Antoniou and Schwartz,[53] projecting the transition-state ensemble onto a good reaction coordinate should give a narrow distribution. However, their test only identifies the separatrix and thus is necessary but not sufficient to ensure an accurate reaction coordinate.[61]

Acknowledgment. I thank Gregg Beckham, Michael Hagan, and David Chandler for helpful discussions about this work. I acknowledge support from the Petroleum Research Fund of the American Chemical Society, PRF# 49165-DNI9.

References

(1) Pauling, L. *Nature* **1948**, *161*, 707.

(2) Radzicka, A.; Wolfenden, R. *Science* **1995**, *267*, 5194.

(3) Zhang, X. Y.; Houk, K. N. *Acc. Chem. Res.* **2005**, *38*, 379.

(4) Li, C. H.; Henry, C. S.; Jankowski, M. D.; Ionita, J. A.; Hatzimanikatis, V.; Broadbelt, L. J. *Chem. Eng. Sci.* **2004**, *59*, 5051.

(5) Hanggi, P.; Talkner, P.; Borkovec, M. *Rev. Mod. Phys.* **1990**, *62*, 251.

(6) Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. *J. Phys. Chem.* **1996**, *100*, 12771.

(7) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. *Science* **2004**, *303*, 186.

(8) Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. *Chem. Rev.* **2006**, *106*, 3188.

(9) Keck, J. *Adv. Chem. Phys.* **1967**, *13*, 85.

(10) Garrett, B. C.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **1984**, *35*, 159.

(11) Schenter, G. K.; Garrett, B. C.; Truhlar, D. G. *J. Chem. Phys.* **2003**, *119*, 5828.

(12) Rosta, E.; Woodcock, H. L.; Hummer, G.; Brooks, B. R. *J. Comput. Chem.* **2009**, *30*, 1634.

(13) Aqvist, J.; Warshel, A. *J. Am. Chem. Soc.* **1990**, *112*, 2860.

(14) Hu, H.; Yang, W. *Annu. Rev. Phys. Chem.* **2008**, *59*, 573.

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

(16) Warshel, A. *J. Biol. Chem.* **1998**, *273*, 27035.

(17) Bruice, T. C.; Bruice, P. Y. *J. Am. Chem. Soc.* **2005**, *127*, 12478.

(18) Strajbl, M.; Sham, Y. Y.; Villa, J.; Chu, Z. T.; Warshel, A. *J. Phys. Chem. B* **2000**, *104*, 4578.

(19) Hammes-Schiffer, S.; Benkovic, S. *Annu. Rev. Biochem.* **2006**, *75*, 519.

(20) Hammes-Schiffer, S.; Tully, J. C. *J. Chem. Phys.* **1994**, *101*, 4657.

(21) Cui, Q.; Karplus, M. *J. Phys. Chem. B* **2002**, *106*, 7927.

(22) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2008**, *48*, 1198.

(23) Gao, J. *Rev. Comput. Chem.* **1996**, *7*, 1996.

(24) Warshel, A.; Florian, J. *Empirical Valence Bond and Related Approaches*; Wiley and Sons: New York, NY, 2004.

(25) Agarwal, P. K.; Billeter, S. R.; Hammes-Schiffer, S. *J. Phys. Chem. B* **2002**, *106*, 3283.

(26) Benderskii, V. A.; Goldanskii, V. I.; Makarov, D. E. *Chem. Phys. Lett.* **1991**, *186*, 517.

(27) Billeter, S. R.; Webb, S. P.; Agarwal, P. K.; Iordanov, T.; Hammes-Schiffer, S. *J. Am. Chem. Soc.* **2001**, *123*, 11262.

(28) Pu, J.; Gao, J.; Truhlar, D. G. *Chem. Rev.* **2006**, *106*, 3140.

(29) Fernandez-Ramos, A.; Ellingson, B. A.; Garrett, B. C.; Truhlar, D. G. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Cundari, T. , Eds.; Wiley-VCH: Hoboken, NJ, 2007; Vol. 23.

(30) Dybala-Defratyka, A.; Paneth, P.; Banerjee, R.; Truhlar, D. G. *Proc. Nat. Acad. Sci. U.S.A.* **2007**, *104*, 10774.

(31) Truhlar, D. G.; Gao, J.; Garcia-Viloca, M.; Alhambra, C.; Corchado, J.; Sanchez, M. L.; Poulsen, T. D. *Int. J. Quantum Chem.* **2004**, *100*, 1136.

(32) Garcia-Viloca, M.; Truhlar, D. G.; Gao, J. *Biochemistry* **2003**, *42*, 13558.

(33) Rothlisberger, D.; Khersonsky, D.; Wollacot, A. M.; Jiang, L.; DeChancie, J.; Betker, J. L.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D. *Nature* **2008**, *453*, 190.

(34) Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; Rothlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas, C. F.; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D. *Science* **2008**, *319*, 1387.

(35) Kuharski, R. A.; Chandler, D.; Montgomery, J. A.; Rabii, F.; Singer, S. J. *J. Phys. Chem.* **1988**, *92*, 3261.

(36) Kramers, H. A. *Physica* **1940**, *7*, 284.

(37) Chandler, D. *J. Chem. Phys.* **1978**, *68*, 2959.

(38) Straub, J. E.; Berne, B. *J. Chem. Phys.* **1985**, *83*, 1138.

(39) Vanden-Eijnden, E.; Tal, F. A. *J. Chem. Phys.* **2005**, *123*, 184103.

(40) Berne, B.; Borkovec, M.; Straub, J. E. *J. Phys. Chem.* **1988**, *92*, 3711.

(41) Peters, B.; Bell, A. T.; Chakraborty, A. K. *J. Chem. Phys.* **2004**, *121*, 4453.

(42) Hu, H.; Kobrak, M. N.; Xu, C.; Hammes-Schiffer, S. *J. Phys. Chem. A* **2000**, *104*, 8058.

(43) Hammes-Schiffer, S. *Biochemistry* **2002**, *41*, 13335.

(44) Kim, H. J.; Hynes, J. T. *J. Am. Chem. Soc.* **1992**, *114*, 10528.

(45) Pineda, J. R. E. T.; Schwartz, S. D. *Philos. Trans. R. Soc., B* **2006**, *361*, 1433.

(46) Antoniou, D.; Abolfath, M. R.; Schwartz, S. D. *J. Chem. Phys.* **2004**, *121*, 6442.

(47) Antoniou, D.; Basner, J.; Nunez, S.; Schwartz, S. D. *Chem. Rev.* **2006**, *106*, 3170.

(48) Quaytman, S.; Schwartz, S. D. *Proc. Nat. Acad. Sci. U.S.A.* **2007**, *104*, 12253.

(49) Schwartz, S. D.; Schramm, V. L. *Nat. Chem. Biol.* **2009**, *5*, 552.

(50) Bolhuis, P. G.; Dellago, C.; Chandler, D. *Faraday Discuss.* **1998**, *110*, 421.

(51) Dellago, C.; Bolhuis, P. G.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 9236.

(52) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291.

(53) Antoniou, D.; Schwartz, S. D. *J. Chem. Phys.* **2009**, *130*, 151103.

(54) Liu, F.; Nakaema, M.; Gruebele, M. *J. Chem. Phys.* **2009**, *131*, 195101.

(55) Caratzoulas, S.; Schwartz, S. D. *J. Chem. Phys.* **2001**, *114*, 2910.

(56) Best, R.; Hummer, G. *Proc. Nat. Acad. Sci. U.S.A.* **2005**, *102*, 6732.

(57) Ma, A.; Dinner, A. R. *J. Phys. Chem. B* **2005**, *109*, 6769.

(58) Peters, B.; Trout, B. L. *J. Chem. Phys.* **2006**, *125*, 054108.

(59) Peters, B.; Beckham, G. T.; Trout, B. L. *J. Chem. Phys.* **2007**, *127*, 1.

(60) Borrero, E. E.; Escobedo, F. A. *J. Chem. Phys.* **2007**, *127*, 164101.

(61) Peters, B. *J. Chem. Phys.* **2006**, *125*, 241101.

(62) Knott, B.; Duff, N. C.; Doherty, M. F.; Peters, B. *J. Chem. Phys.* **2009**, *131*, 224112.

(63) Peters, B.; Zimmerman, N. E. R.; Beckham, G. T.; Tester, J. W.; Trout, B. L. *J. Am. Chem. Soc.* **2008**, *130*, 17342.

(64) Juraszek, J.; Bolhuis, P. G. *Biophys. J.* **2008**, *95*, 4246.

(65) Beckham, G. T.; Peters, B.; Starbuck, C.; Variankaval, N.; Trout, B. L. *J. Am. Chem. Soc.* **2007**, *129*, 4714.

(66) Beckham, G. T.; Peters, B.; Trout, B. L. *J. Phys. Chem. B* **2008**, *112*, 7460.

(67) Vreede, J.; Juraszek, J.; Bolhuis, P. G. *Proc. Nat. Acad. Sci. U.S.A.* **2010**, *107*, 2397.

(68) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334.

(69) Geissler, P. G.; Dellago, C.; Chandler, D. *J. Phys. Chem. B* **1999**, *103*, 3706.

(70) Schlegel, H. B. *J. Comput. Chem.* **2003**, *24*, 1514.

(71) Marcus, R. A. *J. Chem. Phys.* **1966**, *45*, 4500.

(72) Marcus, R. A. *J. Chem. Phys.* **1968**, *49*, 2610.

(73) Truhlar, D. G.; Kuppermann, A. *J. Am. Chem. Soc.* **1971**, *93*, 1840.

(74) Shavitt, I. *J. Chem. Phys.* **1968**, *49*, 4048.

(75) Fukui, K. *Acc. Chem. Res.* **1981**, *14*, 363.

(76) Miller, W. H.; Handy, N. C.; Adams, J. E. *J. Chem. Phys.* **1980**, *72*, 99.

(77) Garrett, B. C.; Truhlar, D. G. *J. Phys. Chem.* **1979**, *83*, 1052.

(78) Garrett, B. C.; Truhlar, D. G. *J. Phys. Chem.* **1979**, *83*, 1079.

(79) Garrett, B. C.; Truhlar, D. G. *Proc. Nat. Acad. Sci. U.S.A.* **1979**, *76*, 4755.

(80) Truhlar, D. G.; Garrett, B. C. *Acc. Chem. Res.* **1980**, *13*, 440.

(81) Truhlar, D. G.; Garrett, B. C. *J. Phys. Chem. B* **2000**, *104*, 1069.

(82) Pollak, E. *J. Chem. Phys.* **1986**, *85*, 865.

(83) Frenkel, D.; Smit, B. *Understanding molecular simulation: from algorithms to applications*; Academic Press: San Diego, CA, 2002.

(84) Peters, B.; Bell, A. T.; Chakraborty, A. K. *J. Chem. Phys.* **2004**, *121*, 4461.

CT100051A

JCTC Journal of Chemical Theory and Computation

Proton Transfer Dynamics in Crystalline Maleic Acid from Molecular Dynamics Calculations

Przemyslaw D. Dopieralski,*,† Zdzislaw Latajka,† and Ivar Olovsson‡

*University of Wroclaw, Faculty of Chemistry, 14 Joliot-Curie Str.
50-383 Wroclaw, Poland, and Department of Materials Chemistry, Ångström
Laboratory, SE-751 21 Uppsala, Sweden*

Received February 9, 2010

Abstract: The crystal structure of maleic acid, the cis conformer of HOOC−CH=CH−COOH has been investigated by Car−Parrinello molecular dynamics (CPMD) and path integral molecular dynamics (PIMD) simulations. The interesting feature of this compound, compared to the trans conformer, fumaric acid, is that both intra- and intermolecular hydrogen bonds are present. CPMD simulations at 100 K indicate that the energy barrier height for proton transfer is too high for thermal jumps over the barrier in both the intra- and intermolecular hydrogen bonds. Dynamics at 295 K reveal that the occupancy ratio of the proton distribution in both the intra- and intermolecular hydrogen bonds is 0.96/0.04. The time lag between the proton transfers in the intra- and intermolecular hydrogen bonds is in the range of 2−9 fs. This is slightly shorter than the time lag obtained previously for fumaric acid, where only intermolecular hydrogen bonds are present. It is also interesting to notice that in most cases the proton transfer process starts in the intramolecular hydrogen bond and subsequently follows in the intermolecular hydrogen bond. Vibrational spectra of the investigated system and its deuterated analogs HOOC−CH=CH−COOD and DOOC−CH=CH−COOD have been calculated and compared with experimental data.

Introduction

Carboxylic acids are typical examples of molecular systems with double hydrogen bonds. Several experimental[1−5] as well as theoretical[6−9] studies of the proton transfer dynamics in such systems have been done recently. But there is still a lack of experimental and theoretical data in the literature for the proton transfer dynamics in dicarboxylic acids. Among the dicarboxylic acids, the cis and trans conformers of butenedioic acid, HOOC−CH=CH−COOH, maleic and fumaric acid, respectively, play a very important role. The physical properties of maleic acid are very different from those of fumaric acid, so it is of interest to study this system and to compare it with our previous results on crystalline fumaric acid.[10] Maleic acid is used in organic synthesis, in the polymer industry, and in oil conservation.[11−14] It is also an inhibitor of fumarate dehydrogenase. Salts of maleic acid are also used in the pharmaceutical industry for drug preparation. Polymorphism is of crucial importance in this context as different crystal structures of the same material may have markedly different physicochemical properties.

Although maleic acid has a wide range of applications and is of biological significance, the literature data are rather limited. The trans isomer, fumaric acid, forms infinite chains of double hydrogen bonds (H-bonds), whereas the cis isomer, maleic acid, forms infinite chains of single H-bonds, as one of the protons is involved in an intramolecular H-bond (Figure 1). The O···O distances, 2.64 Å for the intermolecular and 2.50 Å for the intramolecular H-bond,[15] are shorter than the O···O distance in fumaric acid, 2.67 Å,[16] which suggests that it may be possible to observe similar proton transfer processes as seen in fumaric acid.[10] In fumaric acid, benzoic acid,[17] and $KHCO_3$,[18] the protons are disordered, and the occupation of two possible positions varies with temperature. The situation in maleic acid is complicated as there are two different types of H-bonds. In this particular situation, theoretical calculations are able to provide informa-

* Corresponding author e-mail: mclar@elrond.chem.uni.wroc.pl.
† University of Wroclaw.
‡ Ångström Laboratory.

10.1021/ct100078w
Published on Web 04/14/2010

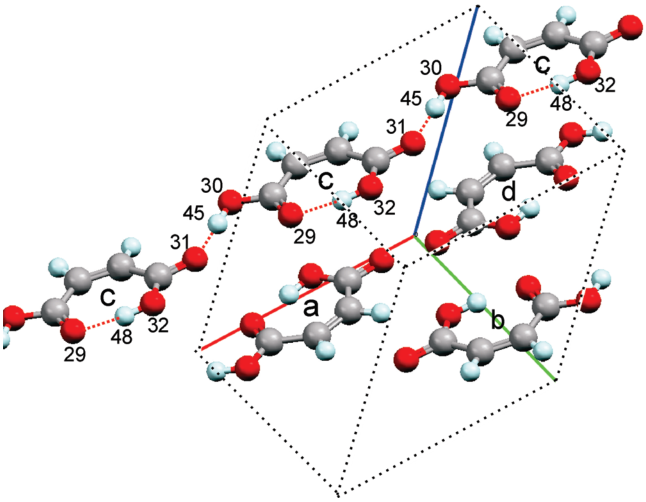


Figure 1. Crystal structure of maleic acid.

tion about the occupation ratio by integration of the distribution functions. In previous X-ray studies, there are no reports of experimentally determined occupation ratios, therefore our investigations are initially restricted to the atomic positions determined by the X-rays structure. The spectroscopic study of isolated maleic acid was essentially limited to the matrix-isolated structure.[19]

The proton transfer reaction in proteins and in solution has been studied extensively.[20−22] In the literature, there are also numerous theoretical papers which deal with double proton transfer, most of which involve the isolated formic acid dimer.[23] The real turning point in study of formic acid dimer dynamics was work by Miura et al.,[6] where authors investigated double proton transfer reaction. Two types of ab initio simulations were carried out: one type of nuclei were treated classically, while in the other, they were quantized via the path integral. In several recent papers, the double proton transfer (DPT) reaction in the cyclic dimer of chloroacetic acid has also been studied using Car−Parrinello molecular dynamics (CPMD) and path integral molecular dynamics (PIMD) techniques, cf. Durlak et al.[9] According to these studies the proton transfers in the isolated system are asynchronous (the two protons do not pass the midpoints of their respective H-bonds at exactly the same time). The two-step mechanism proposed by Ushiyama and Takatsuka for DPT[7] was consistent with the CPMD results. These authors also agree that there is a coupling between the O−H stretching motions and the low-frequency vibrational modes.[2,3,24−26] In the present work, we are taking these studies one step further. To the best of our knowledge intra- and intermolecular H-bond couplings obtained by the CPMD and PIMD methods in the "solid state" have not been compared previously.

Calculations

This work involves CPMD and PIMD calculations of crystalline maleic acid based on the code CPMD,[27] version 3.11.1.[28] The crystal data from the X-ray study by James and Williams[15] have been selected as starting point (cf. also refs 29 and 30). The crystal is monoclinic ($P2_1/c$) with cell dimensions $a = 7.473$, $b = 10.098$, $c = 7.627$ Å, and $\beta =$ 123.59° with four formula units in the unit cell ($Z = 4$).[15] The unit cell of the crystal contains four maleic acid molecules, and each of them forms separate chains (in Figure 1 the content of one unit cell is shown, but the H-bonds in only one chain are illustrated for clarity). The molecules in the unit cell have been optimized with periodic boundary conditions (PBC), and the proton transfer has been studied at 100 and 295 K. A kinetic energy cutoff of 100 Ry was used for the electron plane-wave basis. Troullier and Martins pseudopotentials[31] and Perdew et al. exchange and correlation functional[32] were applied. To control the temperature of the system, the Nosé-Hoover chain thermostat[33,34] was set for the whole system at a target temperature of 100 and 295 K and a coupling frequency of 3000 cm^{-1}.

In the PIMD case, a separate thermostat was used for each degree of freedom.[35] We used time steps of 3 au. The PIMD simulation[36−38] explores the quantum behavior of both the nuclear and electronic degrees of freedom. It maps the problem of a quantum particle onto one of a classical ring polymer model with beads that interact through temperature- and mass-dependent spring forces. Such mapping is known in the literature as quantum-classical isomorphism.[39−41] The path integral simulations in the present study used eight beads and normal mode variable transformation.[38] Eight beads were used in our previous study on fumaric acid[10] and thus proved that such a number of beads is appropriate for the studied system. Dynamics runs of around 20 ps were performed for the crystalline maleic acid. A simulation with a box size of four unit cells, which we employed for fumaric acid[10] and $KHCO_3$[42] was not performed. The size and number of molecules inside one unit cell of maleic acid was deemed to be sufficient.

To study the proton transfer process, we introduce the reaction coordinate δ, defined as the difference between the $\mathrm{r}_{\mathrm{O-H}}$ and $r_{\mathrm{H\cdots O}}$ bond lengths. We use the same procedure to calculate the reaction coordinate for the intra- and the intermolecular H-bonds.

Vibrational spectra have been generated using the program by Forbert,[43] which calculates the spectrum using the inverse fast Fourier transform of the classical autocorrelation function of the total dipole moment, including all contributions—nuclear and electronic. The so-called high-temperature (or harmonic) quantum correction factors to the classical line shape functions were used to approximate the true quantum line shape function and thus the IR spectra.[44] This method is found to work well for anharmonic vibrational spectra and H-bonded systems.[45−48] The visual molecular dynamics program, VMD,[49] has been used for data visualization.

The applicability of density functional theory (DFT)-based methods to describe H-bonded systems depends on the nature of the interaction (these methods are known to have deficiencies in accounting for dispersion interactions).[50] The strong H-bonds in the present system are mainly dominated by electrostatic interactions,[51,52] which supports our choice of the computational methodology. Furthermore, it is important to note that the accuracy of the potential energy surface in Car−Parrinello simulations is determined by the exchange and correlation functionals used. In the CPMD studies of the acetic acid dimer in the gas phase, we have

Table 1. Selected Average Bond Distances (Å) in Crystalline Maleic Acid[a]

		crystal				
		CPMD	CPMD	CPMD	PIMD8	
	bond	opt.	100 K	295 K	295 K	expt[15]
c	O−H (32−48)	1.04	1.05	1.08 (1.03)	1.12 (1.07)	0.91
	O···H (29−48)	1.46	1.47	1.45 (1.48)	1.39 (1.41)	1.59
	O···O (29−32)	2.50	2.51	2.51	2.48	2.50
	O−H (30−45)	1.04	1.04	1.08 (1.03)	1.09 (1.05)	0.98
	O···H (31−45)	1.53	1.55	1.54 (1.53)	1.48 (1.49)	1.66
	O···O (30−31)	2.57	2.58	2.61	2.55	2.64

[a] PIMD8 - Path Integral with 8 beads; data in parentheses are the most probable values. CPMD opt: data from geometry optimization with PBC; most probable values: maximum value of the distribution function for the studied bond; the numbers in parentheses in the column "bond" refer to the atoms numbering in Figure 1.

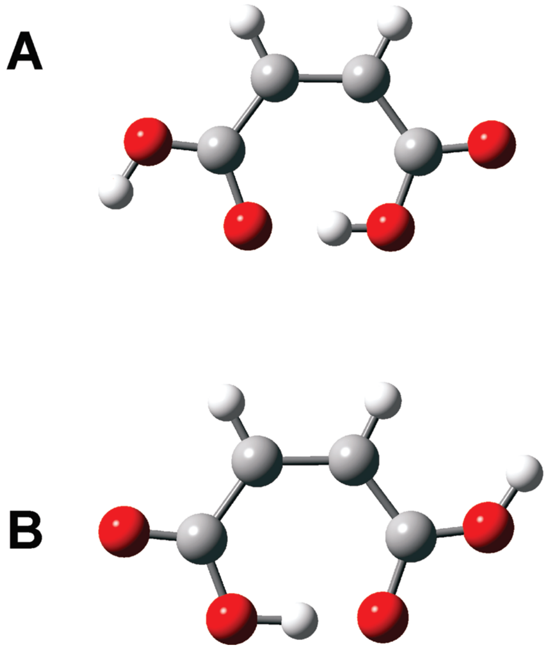


Figure 2. Two conformers studied in present work: (A) the most stable conformer and (B) the second conformer observed in our simulation. The illustrated structures differ by a 180° rotation of the −OH group.

shown that underestimation of the barrier height for proton transfer employing the DFT formalism is approximately compensated by neglecting the zero-point vibrational contribution in the simulations.[9]

Results and Discussion

Structural Parameters. Average bond lengths of the CPMD-optimized structure are compared with the results from the PIMD calculations and the X-ray study[15] in Table 1. Data are only shown for one maleic acid molecule. Bond lengths for other maleic acid molecules in the unit cell do not differ more than 0.01 Å (for CPMD at 100 K). For CPMD and PIMD simulations at 295 K, the differences are larger but still do not exceed 0.04 Å. It is worth pointing out that the conformer of maleic acid shown in Figure 2A was found by Macoas et al.[19] in IR matrix isolation and computational studies to be the most stable in the case of the isolated molecule, and this conformer is also present in the crystalline state. However, during the dynamics simulation, the conformer in Figure 2B was also observed. The energy gap between this conformer and the most stable one (Figure 2A) was predicted by Macoas et al. to be approximately 5.9 kcal/mol (in the case of the isolated system) at the B3LYP/6-31G(d,p) level. Figure 3 shows the difference between the intra- and intermolecular H-bond distribution functions at the two temperatures from the CPMD simulations. At 100 K, no population is observed around a reaction coordinate value of +0.5 Å, which corresponds to a situation where the proton in both the intra- and intermolecular H-bonds has been transferred. When the temperature is increased to 295 K, the distribution becomes broader. The question now arises whether there is a possibility that proton transfer may take place in only one of the two bonds and not in the second? In our case, we did not observe such a situation. For the different molecules in the same chain, something close to a collective mechanism, like in a Grotthuss model,[53,54] was observed. From the present simulations, we may conclude that the proton transfers in the four chains are independent of each other (cf. Figure 1). It is possible that when more than one cell is included in the simulations, new features will appear, but a long enough simulation with at least four unit cells is at present too expensive. This is one of possible future lines of investigations. The proton behavior in crystalline maleic acid is

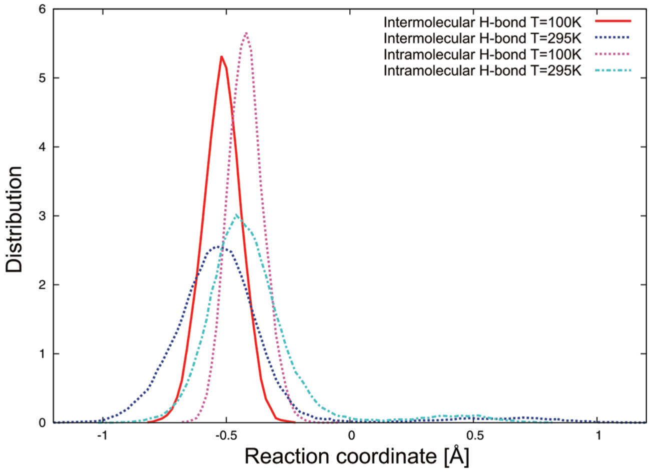


Figure 3. Distribution function (δ distribution) from CPMD at temperatures 100 and 295 K for the intra- and intermolecular H-bonds. Reaction coordinate (δ) is defined as the difference between $r_{O32-H48}$ and $r_{H48-O29}$ bond lengths for the intramolecular H-bond and as the difference between $r_{O30-H45} - r_{H45-O31}$ bond lengths for the intermolecular H-bond.

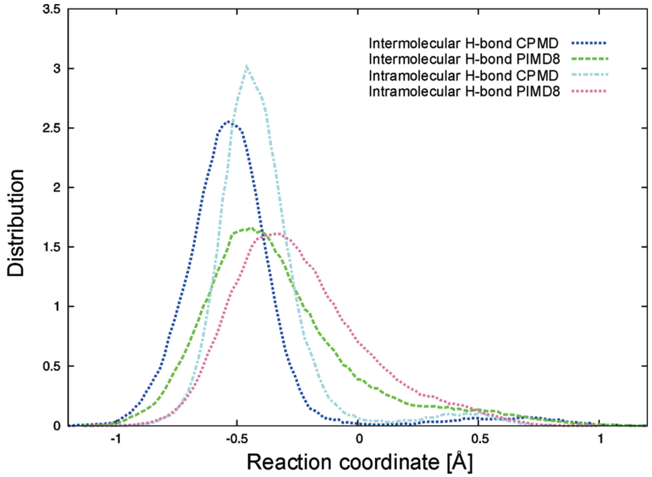


Fig. 4

Figure 4. Distribution function at 295 K from CPMD and PIMD8 simulations for the intra- and intermolecular H-bonds.

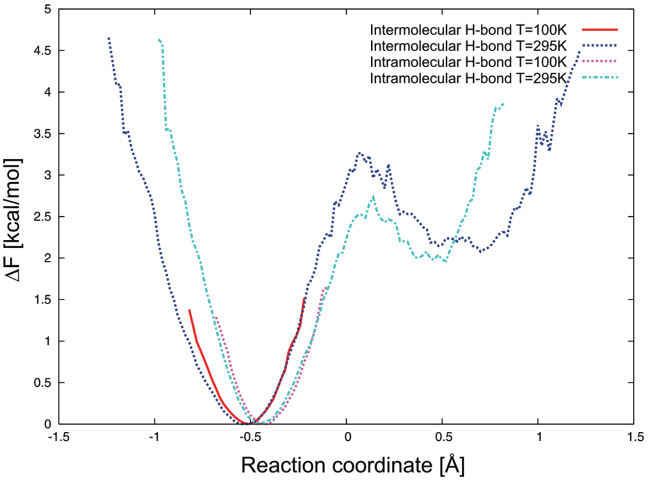


Figure 5. Single proton transfer free energy Δ*F* profile in crystalline maleic acid at 295 and 100 K from CPMD simulations for the intra- and intermolecular H-bonds.

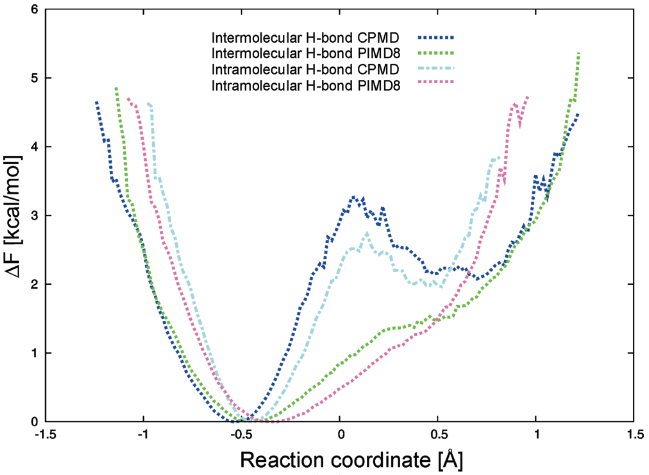


Fig. 6

Figure 6. Single proton transfer free energy Δ*F* profile in crystalline maleic acid at 295 K from CPMD and PIMD8 simulations for the intra- and intermolecular H-bonds.

different depending on which of the standard CPMD and PIMD methods is used. No proton transfer is observed at 100 K using CPMD. PIMD calculations were not performed at 100 K as the strength of the H-bonds is about the same as in fumaric acid, and we accordingly expect similar results, namely that the thermal energy will be not sufficient to push protons across hydrogen bridges.

At 295 K proton transfer occurs with both the CPMD and PIMD methods (Figure 4), but in the PIMD simulation, proton transfers occur more often; the population at reaction coordinate = 0 Å is higher for PIMD than for that of CPMD. The next question is: do the protons transfer simultaneously or successively in the intra- and intermolecular H-bonds, and what is the time lag between the proton jumps? We performed a procedure similar to the one used by Ushiyama and Takatsuka,[7] where the relative coordinates n_1 and n_2 are defined to specify the position of the protons in the hydrogen bonds. Coordinate n_1 describes the proton in the intermolecular H-bond, and coordinate n_2 describes the proton in the intramolecular H-bond. When $n = 0.5$, the proton is exactly in the middle between the oxygen atoms. To the best of our knowledge, no one has earlier performed such a comparison.

$$\begin{aligned} n_1 &= r_{\mathrm{O31-H45}} \cos \theta_{\mathrm{O30-O31-H45}} / r_{\mathrm{O31-O30}} \\ n_2 &= r_{\mathrm{O32-H48}} \cos \theta_{\mathrm{O29-O32-H48}} / r_{\mathrm{O32-O29}} \end{aligned} \qquad (1)$$

This procedure was applied for all four molecules in the unit cell. Analysis of these parameters shows that the proton transfers do not occur exactly simultaneously (with no time lag) or successively (with a large time lag between the transfers)—the observed situation is in between these two scenarios. However, based on the time lags, the proton transfers occur nearly simultaneously. The time series of the parameters n_1 and n_2 from the CPMD simulations allows us to calculate the time lag between the proton transfer in the intra- and intermolecular H-bonds. For the molecule marked (a) in Figure 1 there were four proton jumps with time lags 5, 7, 7, and 7 fs. For molecule (b) there were two proton transfers with time lags 2 and 9 fs. Four proton transfers were observed for molecule (c) with the time lags 8, 5, 7, and 5 fs. For molecule (d) no proton transfer was observed. The whole process of proton transfers in all of the molecules is completed within 40 fs. These values are influenced by the fact that simulations based on the DFT approach can underestimate the barrier, but an important factor is also the O•••O distance. From an analysis of the n_1 and n_2, coordinates we may conclude that the intramolecular proton starts the process, and after that, the intermolecular proton is transferred. In the opposite proton transfer reaction, the intermolecular proton initiates the process. In almost all cases, such a situation was observed for all four molecules.

From our simulations of crystalline maleic acid, it is evident that conformer 2B is easily obtained by proton transfer. It was observed once for every 24 molecules of the most stable conformer 2A, as the CPMD occupancy ratio is 0.96/0.04. The free energy profiles obtained from each simulation type are shown in Figures 5 and 6. The profiles were calculated from the equation:

$$\Delta F = -kT \ln(P[\delta]) \qquad (2)$$

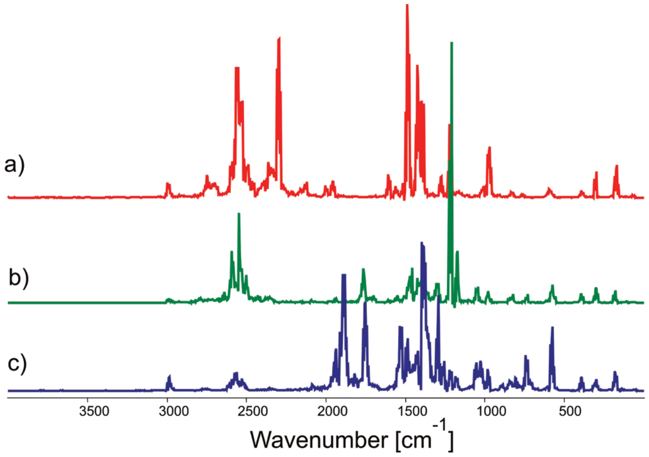


Figure 7. Comparison between theoretical spectra with selectively deuterated hydrogen atoms; upper (red) is without deuteration; middle (green) is only intramolecular H-bond deuterated; lower (blue) is with both intra- and intermolecular H-bonds deuterated. All spectra from autocorrelation of total dipole moments at 100 K.

Where k is the Boltzmann constant, T is the temperature, and $P[\delta]$ is the distribution profile for δ, and the reaction coordinate (δ) is defined as the difference $r_{O32-H48} - r_{H48-O29}$ between the bond lengths in the intramolecular H-bond and as $r_{O30-H45} - r_{H45-O31}$ in the intermolecular H-bond.

As demonstrated by the profiles in Figures 4 and 6, inclusion of quantum effects drastically changes the effective potential shape. The classical limit represented by the 295 K curve exhibits two minima separated by a free energy barrier of around 3.2 kcal/mol. Quantum effects introduced by the PIMD method tend to decrease the barrier. In the next step, we defined two new coordinates: $\rho_1 = r_{O32-H48} - r_{H48-O29}$ and $\rho_2 = r_{O29-O32}$ and then correlated them (data not shown). From this correlation, it was found that proton transfer occurs only when the O•••O distance is shorter than the distance observed in the optimized structure. This is in agreement with our observations for fumaric acid and $KHCO_3$.[10,42] At 295 K, proton transfer occurs only when the O•••O distance (for both the intra- and intermolecular H-bonds) is shorter than 2.5 Å.

Vibrational Spectra

The line shape of the O−H stretching modes of OH groups involved in H-bonds is very complex.[4] Many earlier attempts have been made to analyze the line shape of the O−H/O−D stretching band, mainly in acetic acid dimers in the liquid and gas phases.[55−57] Mechanisms such as Davydov splitting, Franck−Condon combinations with low frequency H-bond modes, multiple Fermi resonances,[58−60] hot bands, exchange tunnelling, predissociation, or breakdown of Born−Oppenheimer approximation have been considered.[55,56,58,59,61−63] Their relative influence on the line shape cannot be determined from the absorption spectrum alone.

In Figure 7, spectra are compared from three different CPMD simulations (from autocorrelation of total dipole moments) with different degrees of deuteration at 100 K. In the first case, only hydrogen in the intramolecular H-bond was deuterated, in the second case hydrogen in both the intra- and intermolecular H-bonds. According to the theoretical results, at 100 K the spectroscopic isotope effect $\nu_{O-H}/\nu_{O-D} = 2556/1890 = 1.35$ for the intermolecular H-bond and $2298/1762 = 1.30$ for the intramolecular H-bond. A large isotopic ratio indicates a large anharmonicity in the H-bond. It is assumed that an isotopic ratio of the order of 1.3 is typical for medium to strong H-bonds.[64] The assignment of bands in the infrared spectrum for crystalline maleic acid is presented in Table 2.

Table 2. Bands Observed in the IR Spectrum of Crystalline Maleic Acid from CPMD Calculations at 100 K

	T = 100 K	
Band (cm^{-1})	H	D
O−H strech. inter.	2556	1890
O−H strech. intra.	2298	1762
O·· ·O strech. inter.	173	184
O•••O strech. intra.	308	299
C═O strech.	1487	1467
C−O strech.	1424, 1218	1303, 1218
C−O strech. intra.	1280	1303
C−H strech.	2995	2987
C═C strech.	1606	1554
C−C strech.	833	827

Here, the description of molecular vibrations was obtained from a classical approach to the nuclear motions, which is the basis of CPMD. Further studies of the vibrational character of maleic acid crystals require the use of a quantum approach to nuclear motion. There are approaches aimed at recovering at least some quantum effects from the CPMD trajectory at a small additional computational cost.[65,66] We would like, however, to remain at the classical level of describing nuclear motion. The main aim of our study is an analysis of the correlation between intra- and intermolecular events, and very accurate description of molecular vibrations is of secondary importance. Solving the multidimensional vibrational Schrödinger equation would be too costly for the purpose of this work.

A comparison between the theoretical spectrum of crystalline maleic acid and the experimental data in the spectral data base system (SDBS)[67] is shown in Figure 8 (the 295 K spectrum from the CPMD simulation with autocorrelation of total dipole moments is not shown here). In this spectrum, a broad absorption band is observed in the range 3100−1900 cm^{-1} due to the proton transfer process. A clearer spectrum was obtained from autocorrelating the velocities—but one must note that the intensities in this spectrum do not represent the true intensities, in contrast to the spectrum obtained by autocorrelating the total dipole moments. Nevertheless, from the spectrum at 295 K one can notice a broad absorption (OH stretching in the intra- and intermolecular bonds) in the range 3100−1900 cm^{-1} due to the proton transfer process. In order to obtain a more resolved spectrum, we have also used the data from the 100 K simulations, where proton transfer was not observed. In Figure 8B, we notice that the theoretical spectrum at 100 K differs significantly in the −OH stretching mode range in comparison with the experimental spectrum. Better agreement is achieved when we compare with the theoretical spectrum at 295 K (where proton transfer

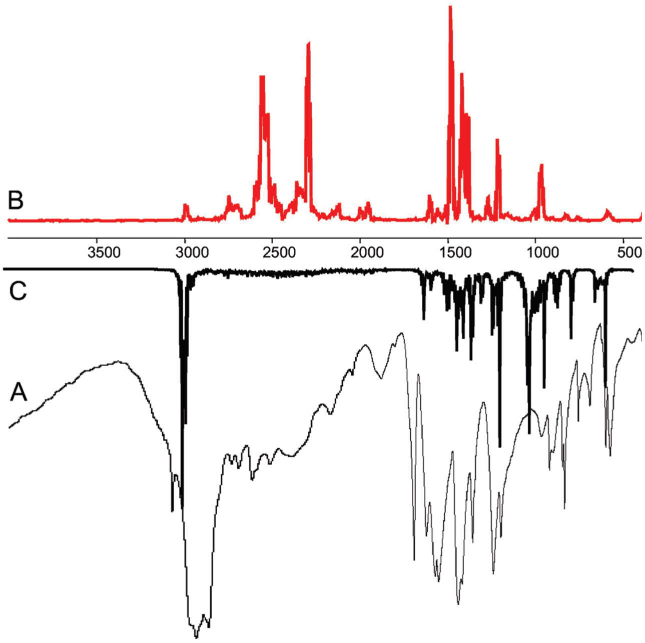


Figure 8. Comparison between experimental A (lower black, IR Nujol)[62] and theoretical spectra B (upper red): theoretical spectrum from dynamics with autocorrelation of total dipole moments at 100 K (no proton transfer) and C (middle black): from dynamics with autocorrelation of atomic velocities at 295 K (with proton transfer).

was observed), Figure 8C. This supports our earlier conclusions that proton transfer occurs at 295 K.

Conclusions

A less stable conformer 2B has been observed besides conformer 2A in our molecular dynamics simulations of crystalline maleic acid. The occupancy ratio (2A/2B) is around 0.96/0.04, which suggests that conformer 2B was observed once for every 24 molecules of the more stable conformer 2A. This ratio is quite small, and as the structure of 2B is not very different from 2A, it is probably very difficult to identify conformer 2B experimentally. Our calculations have shown that the whole proton transfer process starts in the intramolecular hydrogen bond (H-bond) and after that in the intermolecular H-bond. The opposite proton transfer process starts in the intermolecular H-bond. The stronger intramolecular H-bond may accordingly be considered as a switch, which makes the double proton transfer possible. But this is not true for the opposite process! The time gap between the proton transfers is in the range of 2−9 fs, which is more limited when compared to our previous observations in fumaric acid (1−24 fs)[10] and $KHCO_3$ (1−20 fs).[42] This limited range suggests that the proton motion in the intra- and intermolecular H-bonds has to be highly correlated and much more correlated than in the fumaric acid and $(HCO_3)_2^{2-}$ dimers.

Acknowledgment. The authors gratefully acknowledge the Wroclaw Supercomputer Center (WCSS) for providing computer time. We acknowledge the use of the Galera-ACTION Cluster and the support of The Academic Computer Center in Gdansk (CI TASK). The authors thank the Ministry of Science and the Higher Education of Poland (MNiSzW) for funding under grant no. N N204 0958 33. The authors thank Dr. Jaroslaw Panek and Dr. Rachel Glaves for fruitful discussion and comments.

References

(1) Madeja, F.; Havenith, M. *J. Chem. Phys.* **2002**, *117*, 7162–7168.

(2) Heyne, K.; Huse, N.; Nibbering, E. T. J.; Elsaesser, T. *Chem. Phys. Lett.* **2003**, *382*, 19–25.

(3) Heyne, H.; Huse, N.; Dreye, J.; Nibbering, E. T. J.; Elsaesser, T.; Mukamel, S. *J. Chem. Phys.* **2004**, *121*, 902–913.

(4) Nibbering, E. T. J.; Elsaesser, T. *Chem. Rev.* **2004**, *104*, 1887–1914.

(5) Ortlieb, M.; Havenith, M. *J. Phys. Chem.* **2007**, *A111*, 7355–7363.

(6) Miura, Y. S.; Tuckerman, M. E.; Klein, M. L. *J. Chem. Phys.* **1998**, *109*, 5290–5299.

(7) Ushiyama, H.; Takatsuka, H. *J. Chem. Phys.* **2001**, *115*, 5903–5912.

(8) Emmeluth, C.; Suhm, M. A.; Luckhaus, D. *J. Chem. Phys.* **2003**, *118*, 2242–2255.

(9) Durlak, P.; Morrison, C. A.; Middlemiss, D. S.; Latajka, Z. *J. Chem. Phys.* **2007**, *127*, 064304−064311.

(10) Dopieralski, P.; Panek, J.; Latajka, Z. *J. Chem. Phys.* **2009**, *130*, 164517164517−9.

(11) Kawaguchi, S.; Kitano, T.; Ito, K. *Macromolecules* **1992**, *25*, 1294–1299.

(12) Muller, B.; Schmelich, T. *Corros. Sci.* **1995**, *37*, 877–892.

(13) Wang, F. C.; Green, J. G.; Gerhart, B. B. *Anal. Chem.* **1996**, *68*, 2477–2481.

(14) Solich, M.; Krol, W.; Skirmuntt, K. *Pol. J. Chem.* **1993**, *67*, 433–443.

(15) James, M. N. G.; Williams, G. J. B. *Acta Crystallogr.* **1974**, *B30*, 1249–1257.

(16) Bednowitz, A. L.; Post, B. *Acta Crystallogr.* **1966**, *21*, 566–571.

(17) Wilson, C. C.; Shankland, N.; Florence, A. J. *Chem. Phys. Lett.* **1996**, *253*, 103–107.

(18) Thomas, J. O.; Tellgren, R.; Olovsson, I. *Acta Crystallogr.* **1974**, *B30*, 2540–2549.

(19) Macoas, E. M. S.; Fausto, R.; Lundell, J.; Pettersson, M.; Khriachtchev, L.; Rasanen, M. *J. Phys. Chem.* **2001**, *A105*, 3922–3933.

(20) Braun-Sand, S.; Olsson, M. H. M.; Mavri, J.; Warshel, A. Computer Simulation of Proton Transfer in Proteins and Solutions. In *Hydrogen Transfer Reactions*; Hynes, J. T., Klinman, J. P., Limbach, H.-H., Schowen, R. L., Eds.; Wiley-VCH Verlag GmbH: Weinhein, Germany, 2007; pp 1171−1205.

(21) Warshel, A. Calculations of Enzymic Reactions: Calculations of pKa, Proton Transfer Reactions, and General Acid Catalysis Reactions in Enzymes. *Biochemistry* **1981**, *20*, 3167.

(22) Warshel, A. Molecular Dynamics Simulations of Biological Reactions. *Acc. Chem. Res.* **2002**, *35*, 385–395.

(23) Kim, Y. *J. Am. Chem. Soc.* **1996**, *118*, 1522–1528.

(24) Dreyer, J. *Int. J. Quantum Chem.* **2005**, *104*, 782–793.

(25) Fernando-Ramos, A.; Smedarchina, A.; Rodriges-Otero, J. *J. Chem. Phys.* **2001**, *114*, 1567–1574.

(26) Stenger, J.; Madsen, D.; Dreyer, J.; Nibbering, E. T. J.; Hamm, P.; Elsaesser, T. *J. Phys. Chem.* **2001**, *A105*, 2929–2932.

(27) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.

(28) *CPMD*, version 3.11.1; IBM Research Division and Max Planck Institute: Stuttgart, Germany, 2008; www.cpmd.org.

(29) Gupta, M. P.; Mahata, A. P. *Indian J. Phys.* **1975**, *49*, 74–80.

(30) Shahat, M. *Acta Crystallogr.* **1952**, *5*, 763–768.

(31) Troullier, N.; Martins, J. L. *Phys. Rev.* **1991**, *B43*, 1993–2006.

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

(33) Nose, S. *Mol. Phys.* **1984**, *52*, 255–268.

(34) Martyna, G. J.; Tuckerman, M. E.; Klein, M. L. *J. Chem. Phys.* **1992**, *97*, 2635–2643.

(35) Tuckerman, M. E. Path Integration via Molecular Dynamics. In *Quantum Simulation of Complex Many-Body Systems: From Theory to Algorithms*; Grotendorst, J., Marx, D., Muramatsu, A., Eds.; John von Neumann Instiitute for Computing (NIC): Juelich, Germany, 2002, pp 269−298.

(36) Marx, D.; Parrinello, M. *Z. Phys.* **1994**, *B95*, 143–144.

(37) Marx, D.; Parrinello, M. *J. Chem. Phys.* **1996**, *104*, 4077–4082.

(38) Tuckerman, M. E.; Marx, D.; Klein, M. L.; Parrinello, M. *J. Chem. Phys.* **1996**, *104*, 5579–5588.

(39) Feynman, R. P.; Hibbs, A. R. *Quantum Mechanics, Path Integrals*; McGraw-Hill: New York, NY, 1965; pp 280−286.

(40) Schweizer, K. S.; Stratt, R. M.; Chandler, D.; Wolynes, P. G. *J. Chem. Phys.* **1981**, *75*, 1347–1363.

(41) Chandler, D.; Wolynes, P. G. *J. Chem. Phys.* **1981**, *74*, 4078–4095.

(42) Dopieralski, P.; Latajka, Z.; Olovsson, I. *Chem. Phys. Lett.* **2009**, *476*, 223–226.

(43) Forbert, H.;version 30.04.2002(harald.forbert@theochem.ruhr-uni-bochum.de) rev. A. Kohlmayer 04.05.2005, Lehrstuhl fuer Theoretische Chemie, Ruhr-University Bochum: Bochum, 2002.

(44) Ramirez, R. P.; Lopez-Ciudad, T.; Kumar, P.; Marx, D. *J. Chem. Phys.* **2004**, *121*, 3973–3983.

(45) Mathias, G.; Marx, D *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 6980–6985.

(46) Kumar, P.; Marx, D. *Phys. Chem. Chem. Phys.* **2006**, *8*, 573–586.

(47) Rousseau, R.; Kleinschmidt, V.; Schmitt, U. W.; Marx, D. *Angew. Chem., Int. Ed.* **2004**, *43*, 4804–4807.

(48) Asvany, O.; Kumar, P.; Redlich, P. B.; Hegemann, I.; Schlemmer, S.; Marx, D. *Science* **2005**, *309*, 1219–1222.

(49) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.

(50) Zimmerli, U.; Parrinello, M.; Koumouutsakos, P. *J. Chem. Phys.* **2004**, *120*, 2693–2699.

(51) Steiner, T. *Angew. Chem.* **2002**, *41*, 48–76.

(52) Jeffrey, G. A. Nature and Properties. Strong hydrogen bonds. In *An Introduction to Hydrogen Bonding*; Oxford University Press: New York, NY, 1997; pp 11−55.

(53) Grotthuss, C. *Ann. Chim.* **1806**, *LVIII*, 54.

(54) Marx, D. *Chem. Phys. Chem.* **2006**, *7*, 1848–1870.

(55) Marechal, Y.; Witkowski, A. *J. Chem. Phys.* **1968**, *48*, 3697–3705.

(56) Marechal, Y. *J. Chem. Phys.* **1987**, *87*, 6344–6353.

(57) Bratos, S.; Hadzi, D. *J. Chem. Phys.* **1957**, *27*, 991–998.

(58) Chamma, D.; Henri-Rousseau, O. *Chem. Phys.* **1999**, *248*, 53–70.

(59) Chamma, D.; Henri-Rousseau, O. *Chem. Phys.* **1999**, *248*, 71–89.

(60) Dreyer, J. *J. Chem. Phys.* **2005**, *122*, 184306184306−10.

(61) Florio, G. M.; Zwier, T. S.; Myshakin, E. M.; Jordan, K. D.; Sibert, E. J., III *J. Chem. Phys.* **2003**, *118*, 1735–1746.

(62) Emmeluth, C.; Suhm, M. A. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3094–3099.

(63) Chamma, D.; Henri-Rousseau, O. *Chem. Phys.* **1999**, *248*, 91–104.

(64) Bratos, S.; Leicknam, J.-Cl.; Gallot, G.; Ratajczak, H. Ultrafast Hydrogen Bonding Dynamics and Proton transfer Processes. In *The Condensed Phase*, Elsaesser, T., Bakkaer, H. J., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2002; pp 5.

(65) Giannopoulou, A.; Aletras, A. J.; Papatheodorou, G. N.; Yannopoulos, N. *J. Chem. Phys.* **2007**, *126*, 205101–205109.

(66) Stare, J.; Panek, J.; Eckert, J.; Grdadolnik, J.; Mavri, J.; Hadzi, D. *J. Phys. Chem. A* **2008**, *112*, 1576–1586.

(67) SDBSWeb; National Institute of Advanced Industrial Science and Technology: Tokyo, Japan; http://riodb01.ibase.aist.go.jp/sdbs/. Accessed July 19, 2009.

CT100078W

JCTC Journal of Chemical Theory and Computation

Trends in R−X Bond Dissociation Energies ($R^\bullet$ = Me, Et, *i*-Pr, *t*-Bu, $X^\bullet$ = H, Me, Cl, OH)

Igor Ying Zhang,[†,‡] Jianming Wu,[†] Yi Luo,[‡] and Xin Xu*,[†]

State Key Laboratory of Physical Chemistry of Solid Surfaces, College for Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China, and Department of Theoretical Chemistry, School of Biotechnology, Royal Institute of Technology, KTH, Sweden

Received January 7, 2010

Abstract: Trends for R−X bond dissociation energies have been examined with density functional methods of B3LYP, BMK, M06-2X, MC3MPW, B2PLYP, MCG3-MPW, and XYG3, as well as G3, MCG3/3, G3X, and G4 theories as functions of alkylation (i.e., $R^\bullet$ = Me, Et, *i*-Pr, *t*-Bu) and $X^\bullet$ substitution (i.e., $X^\bullet$ = H, Me, Cl, OH). The results highlight the physical origin of success or failure of each method and demonstrate the good agreement with experimental results for G4, MCG3-MPW, and XYG3. The last holds great promise as a reliable method that is applicable to larger systems.

1. Introduction

Chemistry is fundamentally about making and breaking bonds. For a basic understanding of a chemical process, one has to account for how much energy is required for the cleavage of the old bonds in the reactants and how much energy will be released to form new bonds in the products. Accurate bond dissociation energies (BDEs) are mandatory, as the equilibrium constant K_{eq} is very sensitive to any error in the BDEs. An error of 1, 2, or 3 kcal/mol can lead to an error of a factor of 5, 29, or 158, respectively, in the equilibrium constant K_{eq} at 298 K.[1] Finding trends of relative BDEs can be even more important both practically and fundamentally. Such trends may highlight the active sites most accessible to a reagent,[2] and the knowledge of substitution effects on relative BDEs can be very helpful for the rational synthesis of a target structure.[3]

The homolytic BDE of R−X is defined as the enthalpy change of the dissociation reaction in the gas phase at 298 K and 1 atm:[1]

$$R{-}X(g) = R^\bullet(g) + X^\bullet(g) \qquad (1)$$

where the indicated bond is broken and the products are radicals (•). BDE may be obtained by supplying the experimental or calculated heat of formation (HOF, $\Delta_f H°$) for each species as in eq 2.

$$\mathrm{BDE}(R{-}X) = \Delta_r H° \text{ (eq 1)} = \Delta_f H°(R^\bullet) + \Delta_f H°(X^\bullet) - \Delta_f H°(RX) \qquad (2)$$

People have varied $X^\bullet$ in accordance with their electronegativities (e.g., H, CH_3, Cl, OH) and $R^\bullet$ in accordance with their degree of alkylation (i.e., $R^\bullet = (CH_3)_nCH_{3-n}$ where n = 0, 1, 2, and 3 for Me = CH_3, Et = CH_3CH_2, *i*-Pr = $(CH_3)_2CH$, and *t*-Bu = $(CH_3)_3C$, respectively.[4−6] Relative BDEs to the CH_3−X bond are effectively the enthalpy changes for the X-transfer reaction between CH_3 and $R^\bullet = (CH_3)_nCH_{3-n}$:

$$CH_3{-}X(g) + R^\bullet(g) = CH_3(g) + R{-}X(g) \qquad (3)$$

$$\Delta\mathrm{BDE} = \Delta_r H° \text{ (eq 3)} = \mathrm{BDE}(CH_3{-}X) - \mathrm{BDE}(R{-}X) \qquad (4)$$

In particular, when $X^\bullet$ = H, the enthalpy change for eq 3 is usually defined as the radical stabilization energy (RSE) for the radical $R^\bullet$.[7,8] Such a concept of the ordering of alkyl radical stabilities has proven to be extremely useful in explaining the kinetics and thermodynamics of many chemical reactions.

The majority of experimental BDE data suffer from an uncertainty of 1 to 2 kcal/mol,[1] and this is a continuing source

* Corresponding author e-mail: xinxu@xmu.edu.cn.

[†] Xiamen University.

[‡] Royal Institute of Technology.

10.1021/ct100010d
Published on Web 04/20/2010

of debate and controversy. The high-level theoretical calculations, on the other hand, should produce reasonable absolute BDEs, but more significantly, they should produce even more accurate trends based on the assumption that the uniform treatment of a series of BDEs to calculate ΔBDEs should enhance the prospect of a cancellation of errors.[5−7] In this work, we show that this is not always true with ΔBDE trend calculations for the widely used composite method such as the G3 method[9] and some popular density functional theory (DFT) methods such as B3LYP[10−13] and BMK.[14] We demonstrate that the recently introduced G4,[15] although at higher expense, is a significant improvement over G3, while the multicoefficient extrapolated density functional theory, MCG3-MPW,[16] shows very good agreement with the experiment at a comparable cost of G3. The newly developed fifth-rung[17] XYG3 functional[18] offers particular promise as a reliable method that is applicable to larger systems than do G4 and MCG3-MPW.

It is challenging to calculate accurate BDE via eq 2 for many theoretical methods. As the unpaired electron does not have a partner electron sharing the same space, the open-shell radicals are inherently different from their closed-shell parent molecule.[17] Accurate BDE calculations demand treatment of the open-shell and closed-shell species on an equivalent footing. The widely used B3LYP fails badly in this context, whose BDE errors accumulate as the molecules become large along with alkylation.[5−7,20−23] The situation is likely to be improved in calculating ΔBDE, as eq 3 is isodesmic,[24,25] holding radicals, C−X bonds, and other bond types in both sides. However, Coote and co-workers[6] showed that many popular density functional methods (BLYP,[11,12] PBE,[26] B3LYP,[10−13] B3P86,[12,27] BB1K,[28] MPW1K,[29] KMLYP,[30] BMK,[14] etc.) overestimate the stabilizing effect on BDEs in going from $R^\bullet$ = Me to $R^\bullet$ = *t*-Bu, leading in some cases to incorrect qualitative behavior. Among the DFT methods they examined, they claimed that BMK showed the smallest systematic errors in ΔBDE and provided very reasonable predictions of BDE.[6] The present work confirms the poor behavior of B3LYP but reveals some inherent weaknesses of BMK by doing energy decomposition analysis of the contributions to ΔBDEs from the exchange and correlation functionals.

A recent important development in DFT is the M06 family of functionals,[31] which currently provides the highest accuracy with a broad applicability for chemistry. Similar to that for its predecessor M05,[32] the development of these functionals involved using four alkyl bond dissociation energies (i.e., the ABDE4 set including $R-CH_3$ and $R-OCH_3$ with R = Me and *i*-Pr) as the training set. We test M06-2X here for its applicability to the R−X bond series as functions of alkylation (i.e., $R^\bullet$ = Me, Et, *i*-Pr, *t*-Bu) and $X^\bullet$ substitution (i.e., $X^\bullet$ = H, Me, Cl, OH).

2. Computational Details

Unless otherwise stated or defined by the method per se, the equilibrium geometry of each molecule or radical was optimized at the level of B3LYP/6-311+G(d,p).[10−13,33,34] Analytical harmonic frequency was calculated at the same level to give zero-point energy (ZPE, with scaling factor 0.9877[18]) and thermo-corrections and to ensure that each geometry corresponded to a true local minimum. The final electronic energy was obtained by single point calculation at the level of 6-311+G(3df,2p)[33,34] for all DFTs other than MC3MPW[35] and MCG3-MPW.[16]

B3LYP[10−13] is currently the most popular DFT method. It is one of the first hybrid functionals that replaces some portion of the local exchange energy with the exact exchange energy E_x^{exact}.

$$E_{xc}^{\text{B3LYP}}[\rho] = E_{xc}^{\text{SVWN}} + c_1(E_x^{\text{exact}} - E_x^{\text{S}}) + c_2\Delta E_x^{\text{B}} + c_3\Delta E_c^{\text{LYP}} \quad (5)$$

Here, ΔE_x^{B} [11] and ΔE_c^{LYP} [12] are the generalized gradient approximation (GGA) correction terms to the local density approximation (LDA) exchange-correlation SVWN.[36,37] The three mixing parameters are $\{c_1, c_2, c_3\} = \{0.20, 0.72, 0.81\}$. BMK[14] and M06-2X[31] improve B3LYP by also including the ingredient of kinetic energy density (i.e., hybrid meta-GGAs), whose functional forms can be found in the original papers.[14,31]

The MC3MPW method is one of the first doubly hybrid DFTs,[35] whose total energy was defined as

$$E_{\text{tot}}^{\text{MC3MPW}} = d_2E_{\text{tot}}(\text{SAC/DIDZ}) + (1 - d_2)E_{\text{tot}}(\text{MPWX/MG3S}) \quad (6)$$

Here, MPWX is a one-mixing-parameter hybrid DFT[29] using mPW exchange[38] and PW91 correlation[39] functionals, and MG3S is a 6-311+G(3d2f,2df,2p) basis set,[34,35,40] which uses 3d2f polarization on the second-row, 2df polarization on the first row, and 2p on hydrogen. SAC/DIDZ can be expressed as

$$E(\text{SAC/DIDZ}) = E(\text{HF/DIDZ}) + d_1\Delta E(\text{MP2/DIDZ}) \quad (7)$$

where DIDZ stands for the 6-31+G(d,p) basis set.[34,41] The name MC3 suggests that this is a multicoefficient method that contains three parameters.[35] The mixing parameters are $d_1 = 1.339$ and $d_2 = 0.266$. The percentage of the exact exchange in MPWX is 38%. MC3MPW scales formally as N^5 where N is the number of atoms. Since the scaling-all-correlation (SAC) method[42,43] was employed to extrapolate MP2/DIDZ calculations to the limit of full dynamic correlation of the valence electrons and a complete one-electron basis set for the valence electrons, MC3MPW was found to be more accurate than the conventional hybrid method without an appreciable increase of computational cost.[35]

The MCG3-MPW method is also a doubly hybrid DFT.[16] It combines the G3-like component calculations such as HF/6-31G(d), MP2/MG3S, MP4SDQ/6-31G(2df,p), QCISD/6-31G(d), etc. with hybrid DFT of MPWX/MG3S using eight mixing parameters. MCG3-MPW scales formally as N^7. Its cost is smaller than that of the G3 theory, as full MP4 is replaced with MP4SDQ.

B2PLYP is a widely recognized doubly hybrid functional.[44,45] It employs a hybrid GGA functional defined in eq 8, which may be called B2LYP,[46] as it contains two mixing parameters (i.e., $\{a_x, a_c\} = \{0.53, 0.73\}$).

$$E_{\mathrm{xc}}^{\mathrm{B2LYP}} = a_{\mathrm{x}} E_{\mathrm{x}}^{\mathrm{exact}} + (1 - a_{\mathrm{x}}) E_{\mathrm{x}}^{\mathrm{B}} + a_{\mathrm{c}} E_{\mathrm{c}}^{\mathrm{LYP}} \tag{8}$$

This B2LYP functional cannot be used alone, whose mere purpose is to generate the Kohn−Sham (KS) orbitals and orbital eigenvalues for the MP2-like perturbative correlation energy evaluation. The final form of B2PLYP is completed as

$$E_{\mathrm{xc}}^{\mathrm{B2PLYP}}[\rho] = E_{\mathrm{xc}}^{\mathrm{B2LYP}} + (1 - a_{\mathrm{c}}) E_{\mathrm{c}}^{\mathrm{MP2}} \tag{9}$$

Unlike that in MC3MPW,[35] where Hartree−Fock orbitals are used for a conventional MP2 calculation, the reference wave function in B2PLYP does not satisfy the Brillouin theorem. Nevertheless, single contributions are neglected in B2PLYP.[44,45] There are several new functionals (i.e., B2T-PLYP,[47] B2K-PLYP,[47] B2GP-PLYP,[48] B2π-PLYP,[46] ROB2-PLYP,[49] UB2-PLYP[49]), which are constructed in the same way as B2PLYP but use different $\{a_{\mathrm{x}}, a_{\mathrm{c}}\}$ parameters.

XYG3[18] is a new version of doubly hybrid functional. It has the form as

$$\begin{aligned} E_{\mathrm{xc}}^{\mathrm{XYG3}}[\rho] = E_{\mathrm{xc}}^{\mathrm{SVWN}} + e_1(E_{\mathrm{x}}^{\mathrm{exact}} - E_{\mathrm{x}}^{\mathrm{S}}) + e_2 \Delta E_{\mathrm{x}}^{\mathrm{B}} + \\ e_3(E_{\mathrm{c}}^{\mathrm{MP2}} - E_{\mathrm{c}}^{\mathrm{LYP}}) + \Delta E_{\mathrm{c}}^{\mathrm{LYP}} \end{aligned} \tag{10}$$

It was developed on the basis the adiabatic connection formalism[50−55] using initio slope of the exchange-correlation potential energy defined rigorously as the second-order correlation energy in the Görling−Levy theory (GL2)[56] of coupling-constant perturbation expansion, which demands[56,57]

$$E_{\mathrm{c}}^{\mathrm{GL2}} = E_{\mathrm{c}}[\rho_{1/\lambda}]|_{\lambda=0} \tag{11}$$

The validity of eq 11 is critically dependent on the quality of density and orbitals generated by the corresponding exchange-correlation functional, such that eq 11 was suggested to be used as a check for the accuracy of an approximate correlation functional.[56] XYG3 omits the single contribution as B2PLYP does to approximate $E_{\mathrm{c}}^{\mathrm{GL2}}$ as $E_{\mathrm{c}}^{\mathrm{MP2}}$, while it differs from B2PLYP by using B3LYP to generate the density used to calculate the DFT energy and orbitals from which the PT2 term is computed. We assume, on the basis of the generally good performance of B3LYP, that the B3LYP orbitals are a reasonably good approximation to the real (unknown) KS orbitals to better fulfill the requirement of the Görling−Levy theory than do the B2LYP orbitals used in B2PLYP.

G3 is a widely used composite method.[9] It uses MP2(full)/6-31G(d) geometries for energy evaluations and scaled HF/6-31G(d) frequencies for ZPE and $H_{0\rightarrow 298}$. The G3 energy is effectively at the QCISD(T, Full)/G3Large level through a series of calculations at lower levels. The G3Large basis set is similar to 6-311+G(3d2f,2df,2p). G3X is a modification of G3 theory.[58] The new features include (1) B3LYP/6-31G(2df,p) geometry, (2) B3LYP/6-31G(2df,p) ZPE (scaling factor 0.9854), (3) addition of a g polarization function to the G3Large basis set for second-row atoms at the Hartree−Fock level, and (4) revised empirical higher-level correction (HLC). G4 is the latest successor to G3,[15] which modifies G3 in five ways: (1) B3LYP geometry and ZPE as those in G3X, (2) G3LargeXP (with XP standing for extra polarization functions) as the replacement of G3Large basis set, (3) extrapolated HF limit energies, (4) CCSD(T), in substitution for QCISD(T), and (5) revised HLC.

Table 1. Mean Absolute Deviations (MADs, kcal/mol) for Heats of Formation ($\Delta_f H°$) of $R^•$ Radicals and RX Compounds, As Well As for Bond Dissociation Energies (BDEs) of R−X Bonds[a,b]

no.	methods	$\Delta_f H°(R^•)$	$\Delta_f H°$(RX)	BDE(R−X)[c]
1	B3LYP	1.91 (2.57)	4.25 (11.52)	5.98 (11.91)
2	BMK	**0.73** (1.39)	1.37 (3.97)	1.44 (3.93)
3	M06-2X	0.95 (1.25)	1.50 (3.33)	1.21 (2.51)
4	MC3MPW	2.13 (3.70)	**1.04** (2.75)	1.47 (2.90)
5	B2PLYP	3.81 (7.50)	7.04 (11.69)	2.93 (6.11)
5*	B2PLYP*[d]	1.07 (2.80)	3.47 (7.98)	3.05 (5.94)
6	XYG3	**0.72** (1.05)	**0.84** (2.12)	**1.00** (2.01)
7	MCG3-MPW	0.84 (1.15)	1.33 (2.42)	**0.61** (1.69)
8	G3	0.51 (1.07)	0.38 (1.23)	0.83 (1.95)
9	MCG3/3	**0.29** (0.49)	1.84 (3.59)	1.33 (3.49)
10	G3X	**0.42** (0.86)	**0.23** (0.71)	**0.64** (1.50)
11	G4	0.43 (0.59)	**0.17** (0.50)	**0.76** (1.33)

[a] For a given entry, the maximum absolute error is given in parentheses. The best two values based on the smallest MADs are in boldface, which are resulted from the methods with (nos. 1−7) or without (nos. 8−11) using DFT components. [b] See Tables S1−S3 for details, Supporting Information. [c] BDE(R−X) = $\Delta_f H°(R^•)$ + $\Delta_f H°(X^•)$ − $\Delta_f H°$(RX). [d] B2PLYP* results are taken from ref 45, using a very large CQZV3P basis set including core-polarization functions as originally recommended by Grimme.

Besides the Gn series developed by Curtiss and co-workers,[9,15,40,58] Truhlar's group developed a series of multicoefficient correlation methods (MCCMs[59]). Here, we examine the MCG3/3 method, the best performer of this family, being a G3 analogy of MCCM-version 3. As the full MP4 energy component in G3 is removed, MCG3/3 is faster than G3. The HLC scheme is not adopted in MCG3/3. Instead, six parameters are used to linearly mix the energy components from wave function theory (WFT) methods of different levels and basis sets. The salient difference between MCG3/3,[59] MC3MPW,[35] and MCG3-MPW[16] is that there is no DFT component in the first method.

After being armed with total electronic energy, we calculate the standard HOF in the same manner as Curtiss et al.[9,15,40,58] by first subtracting the calculated atomization enthalpies, using a scaled ZPE for the molecule, from the known experimental HOFs of the isolated atoms at 0 K and then adding the calculated thermo-corrections ($H_{0\rightarrow 298\ \mathrm{K}}$) for the molecule, as well as $H_{0\rightarrow 298\ \mathrm{K}}$ for elements in their standard states from experiments. Spin−orbit (SO) corrections are also included in the present work.[9] Having HOFs on hand, we calculate BDEs and ΔBDEs as enthalpy changes, as defined in eqs 2 and 4.

Calculations were performed by using the Gaussian 03 suite of programs.[34] For Truhlar's M06 suite of functionals,[31] calculations were carried out by using Jaguar.[60]

3. Results and Discussion

Table 1 summarizes the mean absolute deviations (MADs) and maximum absolute deviations (MAXs) for HOFs[61−63] of radicals ($R^•$ = Me, Et, *i*-Pr, and *t*-Bu) and RX compounds with $X^•$ = H, Me, Cl, and OH. MADs and MAXs for R−X BDEs are also presented in Table 1 (the detailed results can be found in Tables S1−S3, Supporting Information). The experimental uncertainties cited in the present work for HOFs are all below 0.5 kcal/mol.[61−63] These are good representative systems, showing how alkylation and X$^\bullet$ substitution will affect the R−X bond strengths.[5−7,20−23] In addition to B3LYP and BMK, we focus on some "new generation" functionals M06-2X,[31] MC3MPW,[35] B2PLYP,[44,45] XYG3,[18] and MCG3-MPW[16] in the present work. Results from G3,[9] G3X,[58] and G4,[15] as well as those from MCG3/3,[59] are also included in Table 1 for comparison.

B3LYP leads to MADs of 1.91 and 4.25 kcal/mol for HOFs of R$^\bullet$ radicals and RX compounds, respectively. These errors unfortunately accumulate,[20−23] increasing MAD to 5.98 kcal/mol for BDE prediction. On the other hand, B2PLYP leads to 3.81 and 7.04 kcal/mol for HOFs of radicals R$^\bullet$ and RX compounds, respectively. MAD is decreased, due to error cancellation, to 2.93 kcal/mol for BDE calculations. It is noteworthy that the present B2PLYP results are obtained at the 6-311+G(3df,2p) level. Using a very large CQZV3P basis set including core-polarization functions as originally recommended by Grimme[44,45] will reduce MADs of HOFs to 1.07 for R$^\bullet$ and 3.47 for RX. Nevertheless, we find that the quadruple-ζ-quality basis sets slightly degrade the R−X BDE calculations (MAD = 3.05 kcal/mol, see Tables S1−S3 for details). BMK, M06-2X, MC3MPW, XYG3, and MCG3-MPW present the best DFT methods currently available for HOFs and BDEs calculations (c.f. Table 1 and Tables S1−S3). Their MADs for BDE predictions are all below 1.50 kcal/mol, approaching "chemical accuracy". G3, G3X, and G4 all display MADs below 0.51 kcal/mol for HOFs and below 0.83 kcal/mol for BDEs, being most satisfactory. Notably, MCG3-MPW displays the lowest MAD (0.61 kcal/mol) for BDEs, even surpassing that of the G4 method.

ΔBDEs defined in eqs 3 and 4 examine how alkylation by successive replacement of H in CH_3X with methyl group affects the C−X bond. In terms of HOFs, ΔBDE may also be written as

$$\Delta BDE = \Delta\Delta_f H°(R^\bullet) - \Delta\Delta_f H°(RX) \qquad (12)$$

where $\Delta\Delta_f H°(R^\bullet) = [\Delta_f H°(CH_3) - \Delta_f H°(R^\bullet)]$ and $\Delta\Delta_f H°(RX) = [\Delta_f H°(CH_3X) - \Delta_f H°(RX)]$. Equation 12 clearly demonstrates that the ΔBDE trend is a composite effect of the stability change of R$^\bullet$ and that of RX along with alkylation, while the latter depends on the electronic nature of X$^\bullet$ ligands.[4−8] If $\Delta\Delta_f H°(R^\bullet) > \Delta\Delta_f H°(RX)$ and $\Delta BDE > 0$, ΔBDE goes up with increasing alkylation. On the other hand, if $\Delta\Delta_f H°(R^\bullet) < \Delta\Delta_f H°(RX)$ and $\Delta BDE < 0$, ΔBDE goes down with increasing alkylation. Experimentally,[61−63] it was found that, when X$^\bullet$ = H and Me, one has $\Delta\Delta_f H°(R^\bullet) > \Delta\Delta_f H°(RX)$. This has been attributed mainly to the increasingly stronger hyperconjugation[64] in radicals than in RX from R$^\bullet$ = Me to R$^\bullet$ = *t*-Bu. Thus, we see that ΔBDE goes up with increasing alkylation, as shown in Figure 1. When X$^\bullet$ = Cl and OH, one has $\Delta\Delta_f H°(R^\bullet) < \Delta\Delta_f H°(RX)$. This effect of X$^\bullet$ has been attributed to the increasing contribution of the ionic R^+X^- configuration for electronegative X$^\bullet$ substituents.[4−8] Such stabilization of R−X increases with increasing alkylation and leads to an increase in the R−X BDEs, despite the accompanying increase of R$^\bullet$ stability.[4−8] Hence, in these cases, we see that ΔBDE goes down with increasing alkylation (Figure 1). B3LYP results clearly violate such trends qualitatively for X$^\bullet$ = Cl or OH. Instead of going down from R$^\bullet$ = Me to R$^\bullet$ = *t*-Bu, it erroneously goes up. It was claimed that the incorrect qualitative behavior of B3LYP ΔBDE for electronegative X$^\bullet$ substituents is a result from overestimation of the stabilizing effect on BDEs, giving $\Delta\Delta_f H°(R^\bullet)$ too large in going from R$^\bullet$ = Me to R$^\bullet$ = *t*-Bu.[4−8] Instead, we find that B3LYP actually underestimates the increasing rates with alkylation for $\Delta\Delta_f H°(R^\bullet)$. The experiments give $\Delta\Delta_f H°(R^\bullet)$ = 22.76 kcal/mol from R$^\bullet$ = Me to R$^\bullet$ = *t*-Bu (See Table S4, Supporting Information). The corresponding B3LYP values are 18.02, being too low by 4.74 kcal/mol. On the other hand, B3LYP also underestimates the increasing rates with alkylation for $\Delta\Delta_f H°(RX)$. The experimental value for $\Delta\Delta_f H°(ROH)$ from R$^\bullet$ = Me to R$^\bullet$ = *t*-Bu is 26.65 kcal/mol (See Table S4), while the corresponding B3LYP value is only 17.13, falling short by 9.52 kcal/mol. Thus, we conclude that the erroneous ΔBDE trend is in fact due to a more severe underestimation tendency for $\Delta\Delta_f H°(RX)$ than for $\Delta\Delta_f H°(R^\bullet)$.

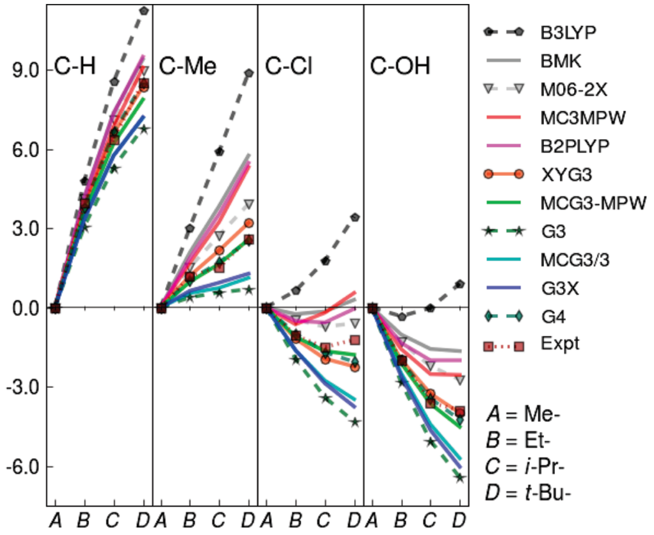


Figure 1. Trends of ΔBDE (kcal/mol) for R−X (R$^\bullet$ = Me, Et, *i*-Pr, and *t*-Bu; X$^\bullet$ = H, Me, Cl, and OH).

Most surprisingly, Figure 1 uncovers the quantitative failure of the G3 method for predicting the ΔBDE trend, despite its good performance for BDE prediction (see Table 1). It is indisputable that the G3 ΔBDE slopes are too gentle when X$^\bullet$ = H and Me, whereas ΔBDEs decrease too fast when X$^\bullet$ = Cl and OH. Table 2 shows that G3 has an error as high as 1.24 kcal/mol for $\Delta\Delta_f H°(R^\bullet)$, and the error in $\Delta\Delta_f H°(RX)$ adds up, leading to MAD = 1.50 and MAX = 3.10 kcal/mol for ΔBDE! Figure 1 illustrates that G3X and MCG3/3 only marginally improve over G3, whereas G4, MCG3-MPW, and XYG3 are very satisfactory in predicting the ΔBDE trend. Table 2 shows that MADs for ΔBDE associated with G4, MCG3-MPW, and XYG3 are 0.21, 0.26, and 0.32 kcal/mol, respectively (See Tables S4−S6 for more details, Supporting Information). We recall that, as eq 3 is isodesmic, holding radicals, C−X bonds, and other bond types in both sides, errors for ΔBDE should be even smaller

Table 2. Mean Absolute Deviations (MADs, kcal/mol) for Relative Heats of Formation ($\Delta\Delta_f H°$) and Relative Bond Dissociation Energies (ΔBDEs)[a,b]

no.	methods	$\Delta\Delta_f H°(R^\bullet)$[c]	$\Delta\Delta_f H°(RX)$[c]	ΔBDE(R−X)[d]
1	B3LYP	2.33 (4.74)	5.48 (11.02)	3.16 (6.28)
2	BMK	1.16 (1.67)	0.59 (2.25)	1.45 (3.14)
3	M06-2X	0.31 (0.62)	0.56 (1.43)	0.79 (1.38)
4	MC3MPW	2.77 (3.65)	1.75 (3.08)	1.04 (2.70)
5	B2PLYP	4.10 (6.76)	5.30 (9.62)	1.21 (2.86)
5*	B2PLYP*[e]	1.86 (3.56)	3.54 (6.98)	1.68 (3.42)
6	XYG3	**0.23** (0.33)	**0.30** (0.92)	**0.32** (1.01)
7	MCG3-MPW	**0.14** (0.41)	**0.37** (0.64)	**0.26** (0.65)
8	G3	1.24 (1.80)	0.31 (1.30)	1.50 (3.10)
9	MCG3/3	**0.27** (0.39)	0.78 (1.86)	**1.05** (2.25)
10	G3X	1.00 (1.55)	**0.30** (0.97)	1.10 (2.53)
11	G4	**0.19** (0.27)	**0.21** (0.57)	**0.21** (0.84)

[a] For a given entry, the maximum absolute error is given in parentheses. The best two values based on the smallest MADs are in boldface, which are resulted from the methods with (nos. 1−7) or without (nos. 8−11) using DFT components. [b] See Tables S4−S6 for details. [c] $\Delta\Delta_f H°(R^\bullet) = \Delta_f H°(CH_3) - \Delta_f H°(R^\bullet)$ and $\Delta\Delta_f H°(RX) = \Delta_f H°(CH_3X) - \Delta_f H°(RX)$. [d] ΔBDE = BD”(CH$_3$-X) − BDE(R-X) = $\Delta\Delta_f H°(R^\bullet) - \Delta\Delta_f H°(RX)$. [e] B2PLYP* results are taken from ref 45, using a very large CQZV3P basis set including core-polarization functions as originally recommended by Grimme.

Table 3. Mean Absolute Deviations (MADs, kcal/mol) for Hartree−Fock and Correlation Contributions to the Relative Bond Dissociation Energies ΔBDEs, Using the Corresponding G4 Values As References[a,b,c]

no.	methods	ΔBDE(HF)	ΔBDE(corr.)	ΔBDE(ZPE+)
1	B3LYP	2.23	1.17	0.05
2	BMK	0.87	2.10	0.05
3	M06-2X	2.01	1.17	0.05
4	MC3MPW	1.12	0.66	0.00
5	B2PLYP	1.21	**0.40**	0.05
6	XYG3	**0.49**	0.52	0.05
7	MCG3-MPW	**0.36**	**0.50**	0.00
8	G3	**0.12**	**0.93**	0.41
9	MCG3/3	**0.16**	0.98	0.00
10	G3X	0.17	**0.85**	0.00
11	G4	0.00	0.00	0.00

[a] ΔBDEs are decomposed in terms of Hartree−Fock (HF) contributions, correlation contributions (corr.), and zero-point-energy plus thermo-contributions (ZPE+). [b] See Tables S7−S9 for details. [c] The best two values based on the smallest MADs are in boldface, which are resulted from the methods with (nos. 1−7) or without (nos. 8−11) using DFT components.

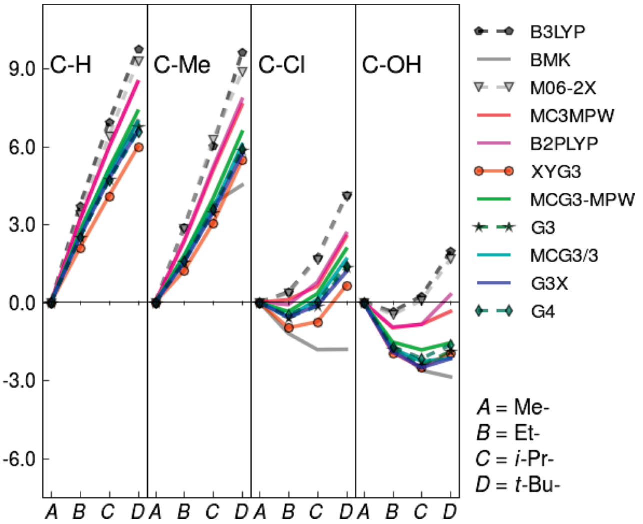


Figure 2. Hartree−Fock contribution to ΔBDE trends of R−X (kcal/mol).

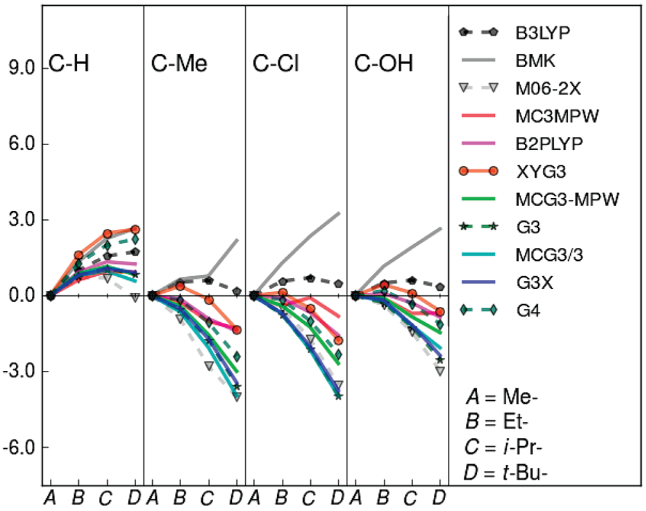


Figure 3. Correlation contribution to ΔBDE trends of R−X (kcal/mol).

than that of BDE for a theoretical method to be satisfactory. Indeed, MADs decrease from BDE to ΔBDE for all methods other than BMK, G3, and G3X.

In order to better understand the physical origin of success or failure of each method, we use G4 energies as references and decompose ΔBDE in terms of Hartree−Fock (HF) contributions, correlation contributions (corr.), and zero-point-energy plus thermo-contributions (ZPE+).

$$\Delta\text{BDE} = \Delta\text{BDE(HF)} + \Delta\text{BDE(corr.)} + \Delta\text{BDE(ZPE+)} \tag{13}$$

The data are summarized in Table 3 with details in Tables S7−S9 (Supporting Information), which are depicted in Figures 2 and 3.

Let us first address ΔBDE(ZPE+). G4, G3X, and Truhlar's MC methods all use B3LYP/6-31G(2df,p) geometries and the corresponding scaled ZPE.[15] All other DFT calculations employ B3LYP/6-311+G(d,p) geometries and the corresponding scaled ZPE,[18] which lead, on average, to ΔBDE(ZPE+) divergence from the G4 results by only 0.05 kcal/mol. G3, on the other hand, uses MP2(full)/6-31G(d) geometries and scaled HF/6-31G(d) ZPE,[9] giving a higher MAD (0.41 kcal/mol) in this quantity. Figure 2 demonstrates that the G3 HF contribution, obtained with HF/G3Large, is quite close to G4 HF results extrapolated to the basis set limit. G3 MAD for ΔBDE(HF) with respect to the G4 data is only 0.12 kcal/mol. Figure 3 and Table 3 clearly disclose that it is the deficiency in G3 correlation contribution, i.e., ΔBDE(corr.), that results in the unsatisfactory performance in quantitatively predicting ΔBDE trends. This holds true for G3X and MCG3/3 methods.

It would be interesting to compare DFT results with those of G4 in terms of ΔBDE(HF) and ΔBDE(corr.). Although it should be kept in mind that the distinction between exchange and correlation is different in DFT and WFT methods, and ΔBDE(HF) in fact contains more information than just that of exchange, we do believe that the trends should be in parallel. Figure 2 demonstrates that B3LYP and

M06-2X display very similar behavior for ΔBDE(HF). They diverge from the G4 values most significantly with MADs = 2.23 and 2.01 kcal/mol, respectively, but they both have a similar trend to that of G4. BMK, on average, seems to be in reasonable agreement with G4 ΔBDE(HF) with MAD = 0.87 kcal/mol. Figure 2, on the other hand, reveals that BMK may have a problem with its exchange functional. For X$^\bullet$ = H, there is a strong increase of ΔBDE(HF) with increasing alkylation, and BMK faithfully reproduces this trend. When X$^\bullet$ = Me, BMK shows a reduced rate of increase for ΔBDE(HF) from *i*-Pr to *t*-Bu, which is not shown in the G4 data. For X$^\bullet$ = Cl, the BMK trend for ΔBDE(HF) is in sharp contrast to the G4 trend. BMK suggests that it is all the way downward from Me to *i*-Pr with a plateau from *i*-Pr to *t*-Bu, whereas G4 shows an initio decrease from Me to Et, followed by a steady increase from Et to *t*-Bu. Even though BMK's trend for X$^\bullet$ = OH is good from Me to *i*-Pr, it differs from G4 ΔBDE(HF) by 1.22 kcal/mol from *i*-Pr to *t*-Bu. We notice that MC3MPW and B2PLYP give a qualitatively good ΔBDE(HF) trend, although its MAD (1.12 and 1.21 kcal/mol, respectively) is less satisfactory from a quantitative point of view. We suggest that updating *E*(HF/DIDZ) in eq 7 to *E*(HF/MG3S) would improve MC3MPW prediction of the ΔBDE(HF) trend.

Figure 3 displays the correlation contribution to the ΔBDE trend. As compared to Figures 1 and 2, it is immediately clear that the ΔBDE trend is generally HF-dominant. Interestingly, while ΔBDE(HF) leads to similar uprising trends for both X$^\bullet$ = H and Me, it is ΔBDE(corr.) that distinguishes them, leading to an attenuated trend for X$^\bullet$ = Me. Indeed, as X$^\bullet$ becomes more electronegative and R$^\bullet$ becomes more alkylated, the correlation effect eventually turns ΔBDE curve downward.

Figure 3 shows that the BMK results are anomalous when G4 ΔBDE(corr.) data are taken as references. This functional clearly displays an opposite trend for electronegative substituents X$^\bullet$. Figure 3 also shows that the B3LYP ΔBDE(corr.) is too mild, being unable to pull the overall ΔBDE trend downward for electronegative X$^\bullet$ along with alkylation. On the other hand, even though M06-2X correlation makes little contribution to ΔBDE for X$^\bullet$ = H, it starts to overshoot the G4 correlation, resulting in a deep decrease of ΔBDE(corr.) with increasing alkylation (Figure 3). Such a correlation contribution is compensated by the sharp increase of ΔBDE(HF) in the opposite direction (Figure 2), giving a net reasonable performance for BDE and ΔBDE (MAD = 1.21 for BDE and 0.79 kcal/mol for ΔBDE). Interestingly, ΔBDE(corr.) from all doubly hybrid functionals (nos. 4−7, see Table 3) are in good agreement with that of G4, suggesting the importance of its second-order perturbation term. Even though B2PLYP ΔBDE(corr.) is most satisfactory, it cannot be balanced with the B2PLYP ΔBDE(HF), making its overall performance (i.e., MAD = 2.93 for BDE and 1.21 kcal/mol for ΔBDE) less satisfactory. Figures 2 and 3 demonstrate that MCG3-MPW's ΔBDE(HF) and ΔBDE(corr.) are both in close parallel with the G4 counterparts, with MADs of 0.36 and 0.50 kcal/mol, respectively (see Table 3). G4 is the most accurate and most expensive Gaussian-n type method with explicit extrapolation of Hartree−Fock energy and using correlation methods up to CCSD(T), while the less expensive MCG3-MPW by taking linear combinations of DFT method with WFT single-level methods to inexplicitly extrapolate toward complete configuration interaction has achieved similar (if not better) accuracy. Most significantly, Figures 2 and 3 demonstrate that XYG3's ΔBDE(HF) and ΔBDE(corr.) are also in good parallel with the G4 counterparts, with MADs of 0.49 and 0.52 kcal/mol, respectively. The comprehensively good performance for BDE and ΔBDE trends suggests that XYG3 gets the right answer for the right reason with a correct description of the fundamental physics. As XYG3 formally scales as N^5, it offers a valuable alternative to the N^7 methods, especially when the latter become prohibitively expensive.

4. Conclusion

Recently, we proposed a doubly hybrid functional, XYG3,[18] which shows very good performance for the calculations of the standard heats of formation, reaction barrier heights, as well as nonbonded interaction. In the present work, we examined the XYG3 performance to calculate bond dissociation energies using R−X (R$^\bullet$ = Me, Et, *i*-Pr, *t*-Bu; X$^\bullet$ = H, Me, Cl, OH) as a representative system. We compared the XYG3 results with those of other state-of-art DFT methods such as doubly hybrid functionals MC3MPW, B2PLYP, MCG3-MPW, and hybrid meta-GGAs (M06-2X and BMK), and the most widely used hybrid GGAs (B3LYP), as well as those from WFT-based composite methods of G3, G3X, MCG3/3, and G4. We conclude the following:

(1) BDEs can be calculated as the energy difference of HOFs (atomization energies), and ΔBDEs are defined here as BDE trends as functions of alkylation and X$^\bullet$ substitution. Our calculations show that errors can accumulate or cancel out in calculating energy differences, such that BDE and ΔBDE carry additional information as compared to the widely used HOF that is important for the judgment of functional performance for "real" chemistry.

(2) Jointly with others,[5−7,20−23] the present work confirms B3LYP's poor performance for BDE calculations, leading to a MAD of 5.98 kcal/mol. B2PLYP yields a MAD of around 3 kcal/mol regardless of the standard 6-311+G(3df,2p) basis set or the very large CQZV3P basis set including core-polarization. MADs for BDE predictions are 1.47 (MC3MPW), 1.44 (BMK), 1.21 (M06-2X), and 1.00 kcal/mol (XYG3), approaching the G3 and G4 results (MAD = 0.8 kcal/mol). MCG3-MPW leads to the smallest MAD (0.61) for this set of BDEs.

(3) B3LYP fails in qualitative prediction of ΔBDE trends for electronegative substituents X$^\bullet$ = Cl or OH, leading to MAD = 3.16 with MAX = 6.28 kcal/mol. Our calculations show that even G3 falls short quantitatively in these cases, giving MAD = 1.50 with MAX = 3.10 kcal/mol. XYG3, MCG3-MPW, and G4 are all satisfactory with MADs = 0.32, 0.26, and 0.21 and MAXs = 1.01, 0.65, and 0.84 kcal/mol, respectively.

(4) Using G4 energy terms as references, our calculations display the anomalous behaviors of BMK in ΔBDE(HF) and

ΔBDE(corr.), downplaying its role as a reliable tool for ΔBDE trend calculations. ΔBDE(HF) and ΔBDE(corr.) from XYG3 and MCG3-MPW are all in close parallel with the G4 counterparts, suggesting that these methods have captured the physical essence of the R−X bond as functions of alkylation and $X^{\bullet}$ substitution, and demonstrating the power of combining DFT and WFT methods as an efficient way of doing accurate electronic structure calculations.

Acknowledgment. This work is supported by NSFC (10774126, 20973138, 20923004) and the Ministry of Science and Technology (2007CB815206).

Supporting Information Available: Experimental and calculated heats of formation; experimental and calculated bond dissociation energies; Hartree−Fock and correlation contributions to the relative bond dissociation energies; optimized geometries at the levels of B3LYP/6-311+G(d,p), B3LYP/6-31G(2df,p), UMP2/6-31G(d), and UHF/6-31G(d). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Luo, Y.-R. *Bond Dissociation Energies in Organic Compounds*; CRC press LLC: Boca Raton, FL, 2002.

(2) Wright, J. S.; Carpenter, D. J.; McKay, D. J.; Ingold, K. U. *J. Am. Chem. Soc.* **1997**, *119*, 4245–4252.

(3) Grimme, S. *Angew. Chem., Int. Ed.* **2006**, *45*, 4460–4464.

(4) Matsunaga, N.; Rogers, D. W.; Zavitsas, A. A. *J. Org. Chem.* **2003**, *68*, 3158–3172.

(5) Coote, M. L.; Pross, A.; Radom, L. *Org. Lett.* **2003**, *5*, 4689–4692.

(6) Izgorodina, E. I.; Coote, M. L.; Radom, L. *J. Phys. Chem. A* **2005**, *109*, 7558–7566.

(7) Henry, D. J.; Parkinson, C. J.; Mayer, P. M.; Radom, L. *J. Phys. Chem. A* **2001**, *105*, 6750–6756.

(8) Poutsman, M. L. *J. Org. Chem.* **2008**, *73*, 8921–8928.

(9) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764–7776.

(10) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(11) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(12) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

(13) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(14) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405–3416.

(15) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2007**, *126*, 084108−1–12.

(16) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 43–52.

(17) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401−1–4.

(18) Zhang, Y.; Xu, X.; Goddard III, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 4963–4968.

(19) Bachrach, S. M. *Computational Organic Chemistry*; John Wiley & Sons, Inc.: Hoboken, NJ, 2007.

(20) Gilbert, T. M. *J. Phys. Chem. A* **2004**, *108*, 2550–2554.

(21) Wu, J. M.; Xu, X. *J. Chem. Phys.* **2008**, *129*, 164103−1–11.

(22) Wu, J. M.; Xu, X. *J. Comput. Chem.* **2009**, *30*, 1424–1444.

(23) Wu, J. M.; Xu, X. *J. Chem. Phys.* **2007**, *127*, 214105−1–8.

(24) Hehre, W. J.; Ditchfie, R.; Radom, L.; Pople, J. A. *J. Am. Chem. Soc.* **1970**, *92*, 4796–4801.

(25) Pople, J. A.; Radom, L.; Hehre, W. J. *J. Am. Chem. Soc.* **1971**, *93*, 289–300.

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

(27) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.

(28) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715–2719.

(29) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 4811–4815.

(30) Kang, J. K.; Musgrave, C. B. *J. Chem. Phys.* **2001**, *115*, 11040–11051.

(31) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(32) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364–382.

(33) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650–654.

(34) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.01; Gaussian, Inc.: Wallingford, CT, 2004.

(35) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 4786–4791.

(36) Slater, J. C. *The Self-Consistent Field for Molecular and Solids, Quantum Theory of Molecular and Solids*; McGraw-Hill: New York, 1974; Vol. 4.

(37) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.

(38) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664–675.

(39) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244–13249.

(40) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1999**, *110*, 4703–4709.

(41) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257–2261.

(42) Gordon, M. S.; Truhlar, D. G. *J. Am. Chem. Soc.* **1986**, *108*, 5412–5419.

(43) Fast, P. L.; Corchado, J.; Sanchez, M. L.; Truhlar, D. G. *J. Phys. Chem. A* **1999**, *103*, 3139–3143.

(44) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108−1–16.

(45) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397–3406.

(46) Sancho-García, J. C.; Pérez-Jiménez, A. P. *J. Chem. Phys.* **2009**, *131*, 084108−1–11.

(47) Tarnopolsky, A.; Karton, A.; Sertchook, R.; Vuzman, D.; Martin, J. M. L. *J. Phys. Chem, A* **2008**, *112*, 3–8.

(48) Karton, A.; Tarnopolsky, A.; Lamère, J.-F.; Schatz, G. C.; Martin, J. M. L. *J. Phys. Chem. A* **2008**, *112*, 12868–12886.

(49) Gramham, D. C.; Menon, A. S.; Goerigk, L.; Grimme, S.; Radom, L. *J. Phys. Chem. A* **2009**, *113*, 9861–9873.

(50) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(51) Langreth, D. C.; Perdew, J. P. *Phys. Rev. B* **1977**, *15*, 2884–2901.

(52) Gunnarsson, O.; Lundqvist, B. *Phys. Rev. B* **1976**, *13*, 4274–4298.

(53) Kurth, S.; Perdew, J. P. *Phys. Rev. B* **1999**, *59*, 10461–10468.

(54) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982–9985.

(55) Yang, W. T. *J. Chem. Phys.* **1998**, *109*, 10107–10110.

(56) Görling, A.; Levy, M. *Phys. Rev. B* **1993**, *47*, 13105–13113.

(57) Mori-Sanchez, P.; Cohen, S. J.; Yang, W. T. *J. Chem. Phys.* **2006**, *124*, 091102−1–4.

(58) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **2001**, *114*, 108–117.

(59) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 3898–3906.

(60) *JAGUAR 7.5*; Schrödinger Inc: Portland, OR, 2008.

(61) Pedley, J. B.; Naylor, R. D.; Kirby, S. P. *Thermochemical Data of Organic Compounds*, 2nd ed.; Chapman and Hall: New York, 1986.

(62) Ruscic, B.; Boggs, J. E.; Burcat, A.; Császár, A. G.; Demaison, J.; Janoschek, R.; Martin, J. M. L.; Morton, M. L.; Rossi, M. J.; Stanton, J. F.; Szalay, P. G.; Westmoreland, P. R.; Zabel, F.; Bérces, T. *J. Phys. Chem. Ref. Data* **2005**, *34*, 573–656.

(63) Berkowitz, J.; Ellison, G. B.; Gutman, D. *J. Phys. Chem.* **1994**, *78*, 2744–2765.

(64) Ingold, K. U.; DiLabio, G. A. *Org. Lett.* **2006**, *8*, 5923–5925.

CT100010D

JCTC Journal of Chemical Theory and Computation

Topological Analysis of the Fukui Function

Patricio Fuentealba,*,† Elizabeth Florez,†,‡ and William Tiznado*,§

Departamento de Física, Universidad de Chile, Las Palmeras 3425, Santiago-Chile, Instituto de Química, Universidad de Antioquia, A.A. 1226, Medellín, Colombia, and Departamento de Ciencias Químicas, Facultad de Ecología y Recursos Naturales, Universidad Andres Bello, Av. República 275, Santiago-Chile

Received January 12, 2010

Abstract: In this work an alternative to the analysis of the Fukui function will be presented and compared with the traditional condensed function. The topological analysis allows us to define basins corresponding to different regions of the space, and the numerical integration of the density over those volumes gives a number amenable of a chemical interpretation in line with the Fukui function applications. Various examples are shown, a series of small molecules, a couple of clusters, and aromatic molecules. They are discussed in comparison with other methodologies and with the experimental evidence.

Introduction

In the last 25 years the development of the density functional theory of chemical reactivity has allowed introducing in a formal framework many empirical chemical concepts like electronegativity,[1] hardness,[2] Fukui function,[3] electrophilicity,[4] and others. Most of them were early defined by Parr and co-workers and are well described in ref 5. New developments were recently reviewed.[6,7] In this work, the focus will be in the Fukui function which was defined as

$$f(\vec{r}) = \left(\frac{\delta\mu}{\delta v(\vec{r})}\right)_N \qquad (1)$$

where μ is the chemical potential and $v(r)$ is the external potential, and the derivative is taken at constant number of electrons N. Very early the impossibility of an exact evaluation of this derivative was realized. The energy presents a discontinuity at an integer number of electrons.[8] Therefore, one has a chemical potential from the left and another one from the right. The first corresponds to the situation where the molecule will lose charge, μ^-, and the later to the situation where the molecule will gain charge, μ^+. In the limit of zero temperature, they are exactly the ionization potential, I, and the electronaffinity, A, respectively. Working further at that limit and using $I = E(N) - E(N-1)$ and $A = E(N+1) - E(N)$ in eq 1, we can deduce that

$$f^-(\vec{r}) = \rho_N(\vec{r}) - \rho_{N-1}(\vec{r}) \qquad (2)$$

for the derivative taken from the left side, and

$$f^+(\vec{r}) = \rho_{N+1}(\vec{r}) - \rho_N(\vec{r}) \qquad (3)$$

for the derivative taken from the right side. In this way, the mathematical discontinuity acquires a chemical meaning. The derivative from the left side corresponds to the capability of the molecule to donate an electron, and the derivative for the right side corresponds to the capability of accepting an electron. One further approximation has been usually done. Under the frozen orbital approximation, these equations transform into

$$f^-(r) = |\phi_H(r)|^2 \qquad (4)$$

and

$$f^+(r) = |\phi_L(r)|^2 \qquad (5)$$

where $\phi_H(r)$ and $\phi_L(r)$ stand for the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), respectively. The last approximation has the practical advantage of their simplicity to allow

* Corresponding author's e-mail: wtiznado@unab.cl and pfuentea@uchile.cl.

† Universidad de Chile.

‡ Universidad de Antioquia.

§ Universidad Andres Bello.

10.1021/ct100022w © 2010 American Chemical Society
Published on Web 04/21/2010

the calculation of the Fukui function in just one calculation to the neutral species without the necessity of calculating the charged species, especially the anions, as it is the case using eqs 2 and 3. However, their lack of relaxation effects has been lately criticized.[9,10] In the last few years, an option to calculate the Fukui function directly from eq 1 without differentiating with respect to the electron number has been put forward,[11,12] and the exact derivative with respect to the number of electrons has been also implemented.[13,14] However, as long as the currently used exchange−correlation functionals are more accurate for integer N than for fractional N,[15–17] the numerical results seem to be better for the models based in the quadratic expansion of the energy with respect to the number of electrons. Independent of the approximations used to calculate the Fukui function, all of them follow the exact equation:

$$\int f(r)\mathrm{d}r = 1 \tag{6}$$

which is important in the use of the Fukui function as an intramolecular reactivity index. In this work, the aim is not to describe a new strategy to calculate the Fukui function but more in the way one can analyze the results of such a calculation. Very soon after the Fukui function was proposed it was realized that the analysis of a three-dimensional function is not trivial and is many times just a number for a molecule or better for a region in a molecule that is more desirable than a number for each value of $\vec{r}$. Yang and Mortier[18] coined the term "condensed" to the approximation of assigning to the Fukui function a number for each atom in the molecule. Hence, under the Mulliken population analysis approach, they proposed to approximate the Fukui function at the atom k as

$$f_k^+ = q_k(N) - q_k(N+1) \tag{7}$$

and

$$f_k^- = q_k(N-1) - q_k(N) \tag{8}$$

where $q(N+1)$, $q(N)$ and $q(N-1)$ are the charges at atom k on the anion, neutral, and cations species, respectively. Under the frozen orbital approximation, these expressions depend only on the electronic structure of the neutral species.[19,20] A variety of forms to calculate the charges have been presented. Most of them are based in some sort of population analysis. The arbitrariness in the way of choosing the charges has been one of the principal criticisms to the condensed Fukui function approximation.

In this work, we propose a methodologically different way to analyze the Fukui function, a topological analysis of the Fukui function, and its comparison with the most related condensation approaches.

Topological Elements to Interpret The Fukui Function. The topological analysis of the electron density was done almost three decades ago by Bader, who investigated the gradient field of the electron density to give a definition of an atom in a molecule.[21] After that, this type of analysis has been used for different functions in chemistry. Especially, Silvi and Savin[22] used it to analyze the electron localization function (ELF). The Fukui function, like the electron density and the ELF, is a scalar function in a three-dimensional (3D) space. Therefore, the analysis of its gradient field allows us to locate the critical points. The critical points of a 3D scalar function can be maximum, minimum, or saddle points. The maximum are called attractors, which are many times amenable of a physical interpretation. For instance, because of Kato's cusp condition,[23] the electron density has a maximum at the nuclei position and, therefore, the electron density has always an attractor associated to each nuclei position.[24] The cusp condition on the Fukui function was first stated by Chattaraj et al,[25] they proposed a gradient expansion for the Fukui function, lately it was derived by Ayers and Levy.[26] In the frontier molecular orbital (FMO) approximation, there is no cusp condition if the orbital has a node only at the atomic position. In this case, there is a "generalized cusp condition" for the density[27,28] depending on how many spatial nodes intersect at the atomic position. Other qualitative difference to be seen in the topological analysis is that the Fukui function calculated as the square of the frontier orbital has, of course, only the symmetries of the irreducible representation of the frontier orbital, whereas the Fukui function calculated as the density differences has all the molecule symmetries for nondegenerate states. It is also useful to define the f-localization domains, which are defined as the volume enclosed by the isosurface $f(r) = f$. It encloses all the points for which $f(r) > f$. They are reducible when they contain more than one attractor and irreducible when they contain one attractor. Each attractor is characterized by its basin, which is the set of points lying on the trajectories ending in this attractor. Since two trajectories cannot cross each other, the basins are irreducible domains, they do not overlap, and the set of all basins fills the complete space. Hence, the whole molecular space is partitioned into basins of attractors, and any physical observable can be defined into this regions. In particular, for a basin labeled, Ω_k, one can calculate the average number of electrons contained into this basin as

$$N_k = \int_{\Omega_k} \rho(\vec{r})\mathrm{d}\vec{r} \tag{9}$$

The sum of the N_k overall basins gives, of course, the total number of electrons. Since we have a donor, f^-, and an acceptor, f^+, Fukui function, we will have two different sets of basins, $\Omega_k^\pm$, and the corresponding chemical interpretation of the resulting numbers, $N_k^\pm$, will be different, accordingly. The site with the greatest N_k^- value should be the site susceptible to donate charge, and inversely, the site with the lowest N_k^+ should be the site susceptible of accepting charge. Note that the N_k index normalizes to the number of electrons, whereas the Fukui function normalizes to one (it is an extensive quantity opposite of the condensed Fukui function). Hence, it can be used for intra- or intermolecular reactivity. However, the chemical interpretation of the N_k values needs further studies to verify its quantitative capability.

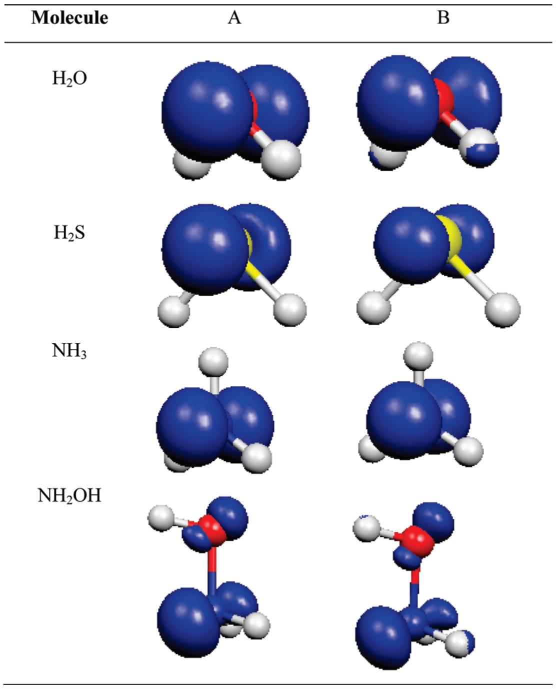


Figure 1. Isosurfaces of the donor Fukui function as (A) the square of the HOMO and (B) as the density differences.

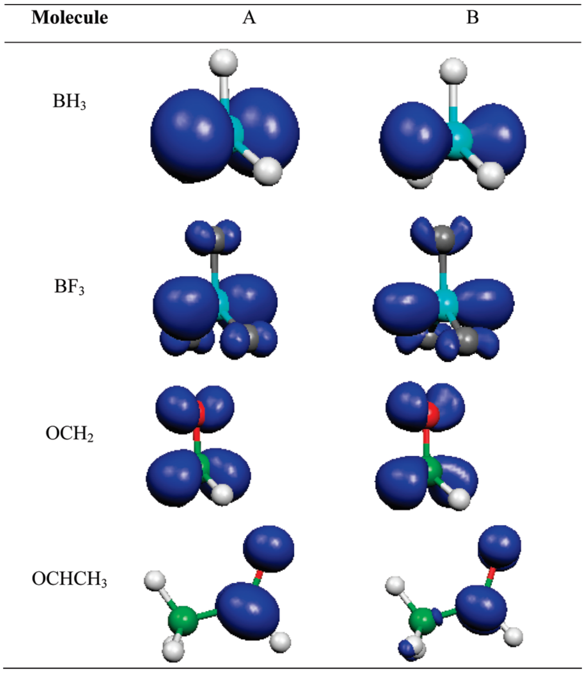


Figure 2. Isosurfaces of the acceptor Fukui function as (A) the square of the HOMO and (B) as the density differences.

One should also like to make contact with the condensed Fukui function. Hence, one can define the quantity:

$$f_k^{\pm} = \frac{N_k}{N} \tag{10}$$

to be compared with any form of condensed Fukui function. It is to notice that eq 10 corresponds to the integration of the shape function, $\sigma(\vec{r}) = \rho(\vec{r})/N$, over the basins of the Fukui function.[29] To avoid the basis set dependence of most population analysis, we condense the Fukui function integrating over its own basins:[30]

Table 1. Average Number of Electrons on the Basins of the Donor Fukui Function Calculated as the Square of the HOMO, A, and as the Density Differences, B, at B3LYP/6-311g** and B3LYP/6-311++g**

molecule	atomic basin	A		B	
		BS1[a]	BS2[b]	BS1[a]	BS2[b]
H_2O	O	10.0	10.0	8.46	8.46
H_2S	S	18.0	18.0	16.0	16.0
HCN	C	6.6	6.6	5.6	5.6
	N	7.4	7.4	7.2	7.2
CO	C	5.7	5.7	5.2	5.2
NH_2^-	N	10.0	10.0	8.4	8.4
NH_3	N	10.0	10.0	7.8	7.8
NH_2OH	N	8.69	8.72	6.86	6.90
	O	7.93	7.94	7.66	7.61
NH_2F	N	9.08	9.09	7.19	7.25
NHF_2	N	8.45	8.45	6.86	6.96
NF_3	N	8.02	8.02	6.56	6.66

[a] BS1 is B3LYP/6-311g**. [b] BS2 is B3LYP/6-311++g**.

Table 2. Average Number of Electrons on the Basins of the Acceptor Fukui Function Calculated As the Square of the LUMO, A, and as the Density Differences, B, at B3LYP/6-311g** and B3LYP/6-311++g**

molecule	atomic basin	A		B	
		BS1[a]	BS2[b]	BS1[a]	BS2[b]
BH_3	B	8.0	8.0	5.8	5.7
BH_2F	B	7.37	7.34	5.45	5.56
BHF_2	B	6.84	6.79	5.24	5.78
BF_3	B	6.51	9.8	7.7	7.1
BCl_3	B	8.10	8.16	5.67	5.88
CH_3^+	C	8.03	8.03	5.29	5.64
CF_3^+	C	6.94	6.90	5.21	5.37
CCl_3^+	C	7.89	7.89	5.96	5.93
CBr_3^+	C	7.85	7.85	5.93	5.93
CI_3^+	C	6.28	6.26	5.75	5.75
CO	C	6.28	6.26	5.93	5.92
OCH_2	C	7.87	7.83	6.08	5.92
$OCHCH_3$	C1	8.94	8.68	6.45	6.51

[a] BS1 is B3LYP/6-311g**. [b] BS2 is B3LYP/6-311++g**.

$$f_{k,C}^{\pm} = \int_{\Omega_k} f_k^{\pm}(\vec{r})\mathrm{d}\vec{r} \tag{11}$$

Assuming one has the topology of the basins to give them a chemical interpretation, the average number of electrons on them, to quantify the capability of a site to accept or donate charge and is analogous to a condensed Fukui function to compare with.

Results and Discussion

The molecular geometries have been optimized using the B3LYP density functional method[31] and two different basis sets,[32,33] namely 6-311G** and 6-311++G**, denoted in the tables as BS1 and BS2, respectively. The two different basis sets have been used to prove that the results are not sensitive to the use of diffuse function. All electronic structure calculations were done using the Gaussian 03 program,[34] the topological analysis of the scalar functions,

Table 3. Condensed Donor Fukui Function, using eqs 10 and 11, for the Two Methodologies[a]

		A		B	
molecule	atom	eq 10	eq 11	eq 10	eq 11
H_2O	O	1.00	1.00	0.84	0.78
H_2S	S	1.00	1.0	0.9	0.9
HCN	C	0.47	0.44	0.40	0.43
	N	0.53	0.56	0.51	0.46
CO	C	0.41	0.86	0.37	0.67
NH_2^-	N	1.0	1.0	0.84	0.82
NH_3	N	1.00	1.00	0.78	0.77
NH_2OH	N	0.48	0.76	0.38	0.47
	O	0.44	0.19	0.42	0.21
NH_2F	N	0.51	0.80	0.40	0.59
NHF_2	N	0.325	0.54	0.26	0.50
NF_3	N	0.23	0.64	0.19	0.41

[a] The square of the HOMO, A, and the density difference, B, at the B3LYP/6-311++g** level of calculation.

Table 4. Condensed Acceptor Fukui Function, using eqs 10 and 11, for the Two Methodologies[a]

		A		B	
molecule	atom	eq 10	eq 11	eq 10	eq 11
BH_3	B	1.00	1.00	0.72	0.8
BH_2F	B	0.46	0.89	0.34	0.86
BHF_2	B	0.28	0.82	0.22	0.88
BF_3	B	0.3	0.9	0.24	0.9
BCl_3	B	0.14	0.64	0.10	0.35
CH_3^+	C	1.00	1.00	0.70	0.71
CF_3^+	C	0.21	0.62	0.17	0.40
CCl_3^+	C	0.14	0.50	0.11	0.22
CBr_3^+	C	0.071	0.49	0.054	0.19
CI_3^+	C	0.24	0.50	0.22	0.15
CO	C	0.45	0.75	0.42	0.83
OCH_2	C	0.49	0.67	0.37	0.68
$OCHCH_3$	C1	0.36	0.50	0.27	0.39

[a] The square of the HOMO, A, and density difference, B, at the B3LYP/6-311++g** level of calculation.

and the calculations of the condensed Fukui function were done with the DGrid 4.4 set of programs.[35]

Figures 1 and 2 show for some molecules the representative isosurfaces of the donor and acceptor Fukui function, respectively. One can see that qualitatively in all cases the functions are very similar, presenting high similarity with the frontier orbital shapes. It is to note that this result is valid only when one analyzes the chemically meaningful Fukui function, i.e., the f^+ and f^- for the acceptor and the donor molecules, respectively. However, there are many cases where the differences can be significant.[36,37] Therefore, for a more exhaustive evaluation, it is important to have a methodology to quantify the Fukui function at the reactive sites of a molecule. This is the main point of this work, to introduce the topological analysis of the Fukui function.

In Table 1, one can see the average electron population of each basin of the donor Fukui function, eq 9, for a series of donor molecules calculated with the different methodologies and with two different basis sets. One notes that, contrary to the calculations based on population analysis, the results are almost independent of the basis set. In general, the numbers calculated integrating over the basins associated to the square of the HOMO, A in Figures 1 and 2, are greater than the ones calculated integrating over the basins associated to the density differences, B in Figures 1 and 2. In particular, using the square of the HOMO, the hydrogen atoms give all the charge to the basin associated to the heteroatom. Hence, for molecules, like H_2O and NH_3, the basin associated to the heteroatom has the total electron numbers of the molecule. However, the qualitative trends are the same. For example, comparing a series of molecules, like NH_2F, NHF_2, and NF_3, the tendency is the same. Only one exception exists in this series of molecules. For the NH_2OH molecule, the Fukui function calculated as the square of the HOMO predicts the nitrogen atom as the most reactive, whereas the Fukui function calculated as the density differences predicts

Table 5. Condensed Donor Fukui function, using eqs 9 and 11, for the Two Methodologies[a]

		A		B		f_k^- [b]	observed products, % per site[c]		
compound C_6H_5X	position (*k*)	eq 9	eq 11	eq 9	eq 11		nit.	benz.	brom.
	o	6.65	0.08	4.93	0.06	0.12	28.5	43.5	19.9
CH_3	*m*	6.20	0.05	4.69	0.04	0.05	1.5	4.5	0
	p	8.40	0.34	6.75	0.20	0.03	40	52	60.3
	o	7.54	0.12	5.84	0.51	0.13			
	m	5.20	0.001	5.12	0.02	0.03			
NH_2	*p*	9.48	0.28	4.98	0.20	0.23			
	o	7.32	0.12	5.64	0.09	0.12	20		4.9
OH	*m*	5.34	0.02	5,2	0.03	0.06	0		0
	p	8.92	0.32	6.9	0.22	0.27	6090.2		
	o	7.66	0.14	6.0	0.10	0.14			
OCH3	*m*	4.26	0.01	4.9	0.01	0.02			
	p	9.60	0.3	7.02	0.19	0.25			
	o	8.40	0.26	4.61	0.04	0.16			
CF_3	*m*	7.00	0.24	4.90	0.05	0.17			
	p	0.00	0.00	6.70	0.20	0.16			
	o	6.53	0.06	4.86	0.05	0.16			
C≡N	*m*	6.2	0.05	4.77	0.04	0.15			
	p	8.24	0.29	6.62	0.19	0.13			
	o	10.3	0.25	5.98	0.14	0.15	0.3		0
NO_2	*m*	10.38	0.24	6.34	0.15	0.17	93.2		100
	p	0.00	0.00	3.06	0.00	0.16	6.40		

[a] The square of the HOMO, A, and density difference, B, at the B3LYP/6-311++g** level of calculation. [b] See ref 39 and 40. [c] See ref 41.

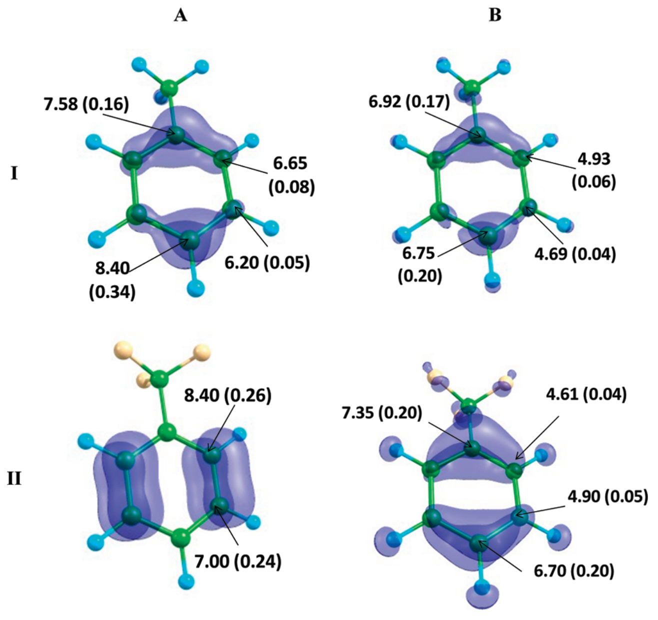


Figure 3. Isosurfaces of the donor Fukui function (A) as the square of the HOMO and (B) as the density differences. Condensed Fukui functions using eqs 9 and 11 in parentheses for (I) $C_6H_5CH_3$ and (II) $C_6H_5CF_3$ molecules at the B3LYP/6-311++G** level of calculation.

the oxygen atom as the most reactive. Experimentally, there is evidence that the nitrogen atom is the most reactive, at least, for protonation, an addition of a carbocation.[38] In Table 2, a similar analysis is shown for the acceptor Fukui function in a series of Lewis acids. Again, the results are very independent of the basis set, and the values calculated using the square of the LUMO are always greater than the values calculated using the density differences. There is one exception for the BF_3 molecule, the values show a great dependence on the basis set, especially the ones calculated using the square of the LUMO. The reason is simple. Changing the basis set changes the order of the virtual orbitals, and the LUMO's are not the same. Hence, the effect is more pronounced in the Fukui function calculated with the LUMO. However, also in this case, looking at any family of molecules the trends are similar. It is important to observe that the interpretation of the numbers is now different. It seems that the most reactive molecule or site is not the one with the biggest number, as it is the case with the condensed Fukui function. Now, the most reactive molecule or site seems to be the one with the smallest number, as it can be seen in the series BH_3, BH_2F, and BHF_2. The apparent failure of this rule in the BHF_2 molecule using the 6-311++G** basis set is due to the numerical difficulty to find a basin associated to the hydrogen atom. Hence, all the charge is added to the basin associated to the boron atom. The BF_3 molecule is ruled out because of the explained change in the frontier orbital. It is important to mention that, independent of the procedure used to calculate the Fukui function, the acceptor Fukui function will be always more difficult to calculate accurately. This is due to the dependence on the virtual frontier orbital and to the complications of correctly calculating the density of an anion. The empirical rule presented here needs further study.

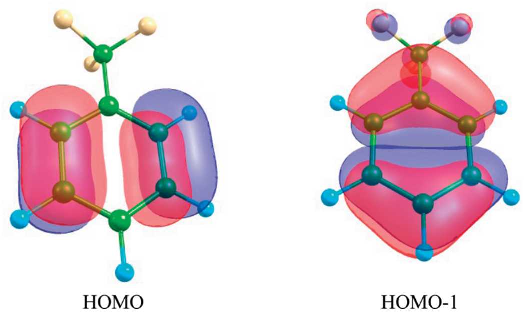


Figure 4. Isosurfaces of the HOMO and HOMO-1 for $C_6H_5CF_3$ at the B3LYP/6-311++G** level of calculation.

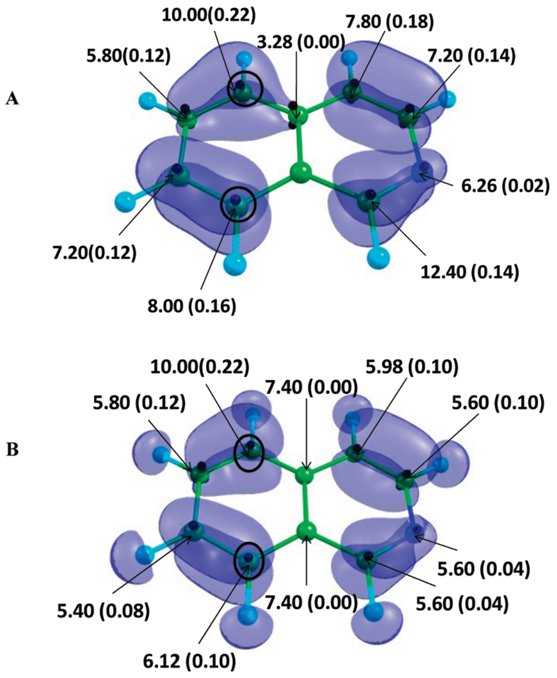


Figure 5. Isosurfaces of the donor Fukui function as (A) the square of the HOMO and (B) as the density differences. Condensed Fukui functions using eqs 9 and 11 in parentheses, enclosed by circles are the experimentally more reactive carbons.

To make contact with the commonly used condensed Fukui function, we have calculated a related quantity given by eq 10 to compare with. The results are in Tables 3 and 4 for the donor and acceptor Fukui functions, respectively. Regarding the donor Fukui function, one can see that the numbers are different, but the trends are similar with the exceptions of CO and NH_2OH molecules where even the trends are different. Remember, however, that the prediction of the correct polarization in the CO molecule is difficult for any methodology. The comparison of the acceptor Fukui function is more difficult because the interpretation of the numbers is different. The condensed Fukui function of eq 11 predicts the most reactive site as the one with the greatest value. Whereas the number associated to eq 10 is the opposite one, with the smallest value for the most reactive site.

In Table 5, the models presented in this work are compared with the condensed Fukui function according to eqs 9 and 11. A set of monosubstituted benzenes have been chosen, C_6H_5X, X = CH_3, NH_2, OH, and OCH_3 (electron-releasing

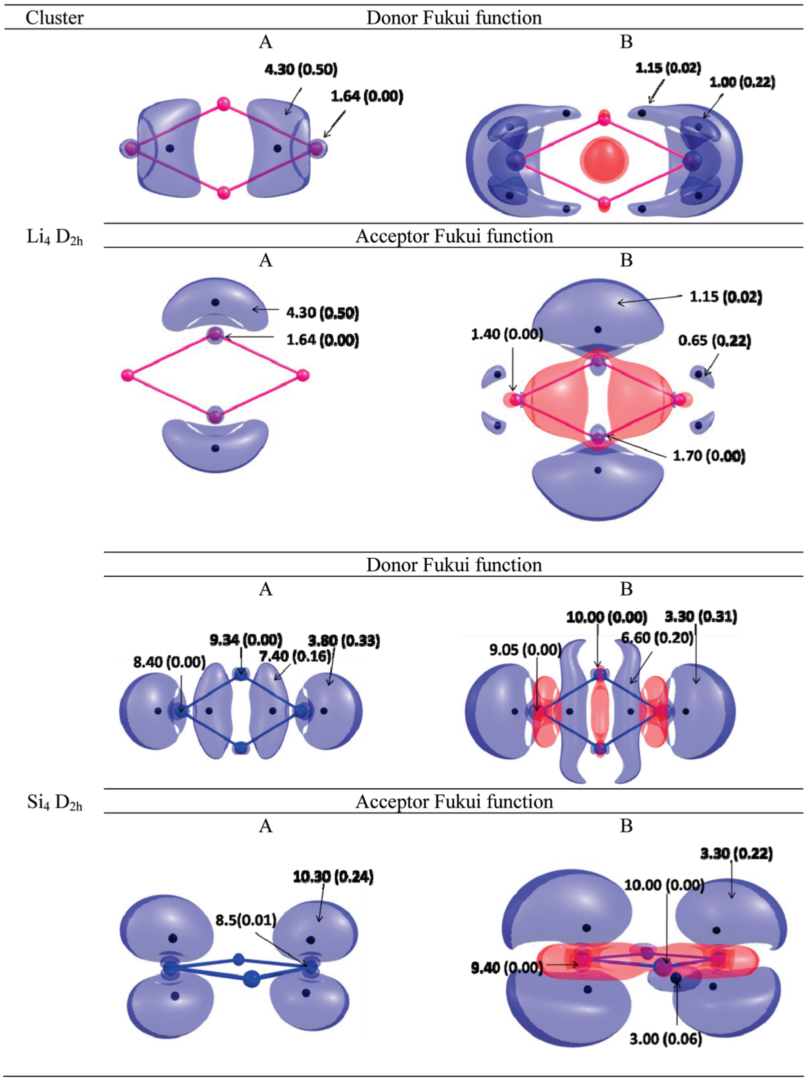


Figure 6. Isosurfaces of the donor Fukui function (A) as the square of the HOMO and (B) as the density differences. Condensed Fukui functions using eqs 9 and 11 in parentheses for Li_4 and Si_4 clusters at the B3LYP/6-311++G** level of calculation.

substituent) and X = CF_3, CN, and NO_2 (electron-attracting substituent). In general, all values show the same qualitative trends, which have also been studied before.[39–41] Note, that in the cases the HOMO is degenerated, it is necessary to consider an average among them.

The isosurfaces and condensed values of the Fukui functions (A and B approximations) to the $C_6H_5CH_3$ (*ortho*−*para* reactivity) and $C_6H_5CF_3$ (*meta* reactivity) are shown in parts I and II of the Figure 3; the condensed values were obtained using eqs 9 and 11. Excluding the *ipso* position in the $C_6H_5CH_3$ reactivity rank, carbons in position *ortho* and *para* are the next more reactive positions, with the *para* position as the most reactive between them, which is in agreement with the experimental observations. When we analyze the results to the $C_6H_5CF_3$ (*meta* reactivity), there are clearly differences between both methods. To evaluate whether this difference is due to an inadequate energy ordering of the orbitals or not, we have explored the lowest energy orbitals. Figure 4 shows the isosurfaces of the HOMO and HOMO-1 orbitals, and comparing with the Figure 3, we can see that the topology of the HOMO-1 orbital is similar to the Fukui function isosurface obtained by the finite differences approximation (method B). Hence, it is important to remark that to obtain an adequate description of reactivity in agreement with the experimental observations, it is necessary the use the HOMO and HOMO-1 orbitals in the Fukui function calculation at the frozen orbital approximation, even when they are not strictly degenerates. The use of the HOMO-1 orbital has been earlier empirically used[41] and recently formally justified.[42]

Sometime ago, Dewar discussed some examples of molecules where the FMO approximation fails to adequately describe regioselectivity.[43] Those molecules were recently studied by Ayers et al[37] using density functional chemical reactivity concepts like the ones analyzed in this work. Therefore, one of these molecules, isoquinoline, has been taken as an example, and the results are shown in Figure 5. The most reactive carbon atoms are enclosed by a circle. Leaving aside the ipso reactions, which are energetically unfavorable, one can see that the finite difference approximation gives results in qualitative agreement with the experimentally observed reactivity. The model based in the square of the HOMO fails as it was predicted and explained by Dewar[43] and lately by Ayers.[37] Qualitatively, one can try to

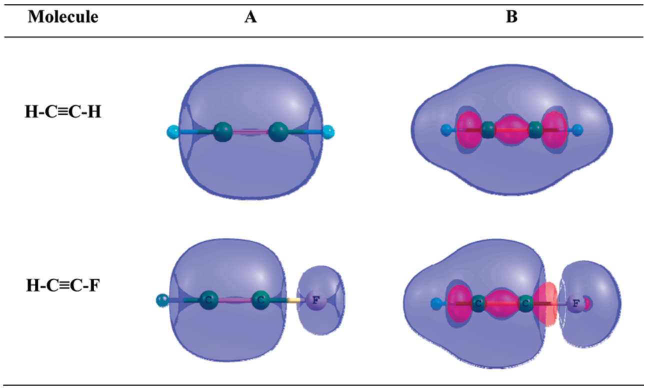


Figure 7. Isosurfaces of the donor Fukui function (A) as the square of the HOMO and (B) as the density differences.

understand the difference because the models based on the finite difference approximation take into account the covalent effect, through the topology of the Fukui function, and the electrostatic effects through the integration of the total density over the basins.

Other important chemical species which have been extensively studied in the last few years are the atomic clusters.[44] The local reactivity of some clusters has been theoretically addressed using different methodologies to calculate the Fukui functions.[30,45–49] Figure 6 shows the Fukui function isosurfaces for the Li_4 and Si_4 clusters obtained by A and B approximations and the corresponding condensed values calculated by eqs 9 and 11. One can see that the principal difference between A and B approximations are the number of basins, but in general, the qualitative reactivity information is the same.

Another important point to study through a topological analysis is the existence of regions with a negative value of the Fukui function.[9,50] Of course, under the approximation of frozen orbitals, this is not possible. However, going beyond this approximation, there is in fact regions of negative values.[10] Figure 7 shows some isosurfaces of the Fukui function for the acetylene molecule and one nonsymmetric derivative. The color now indicates the sign of the Fukui function. Blue means positive, and red means negative. Model A which is the frozen orbital approximation shows, as expected, only positive values. However, in all molecules, the Fukui function in model B, with relaxation effects, presents a region of negative values. There is always a plane containing the atoms into the region of the triple bond which has a negative value of the Fukui function. It seems that in the core region is always probable to find negative values because of the orthogonalization restrictions. A more difficult question is the existence of a basin completely contained in the negative region. Unfortunately, the numerical accuracy

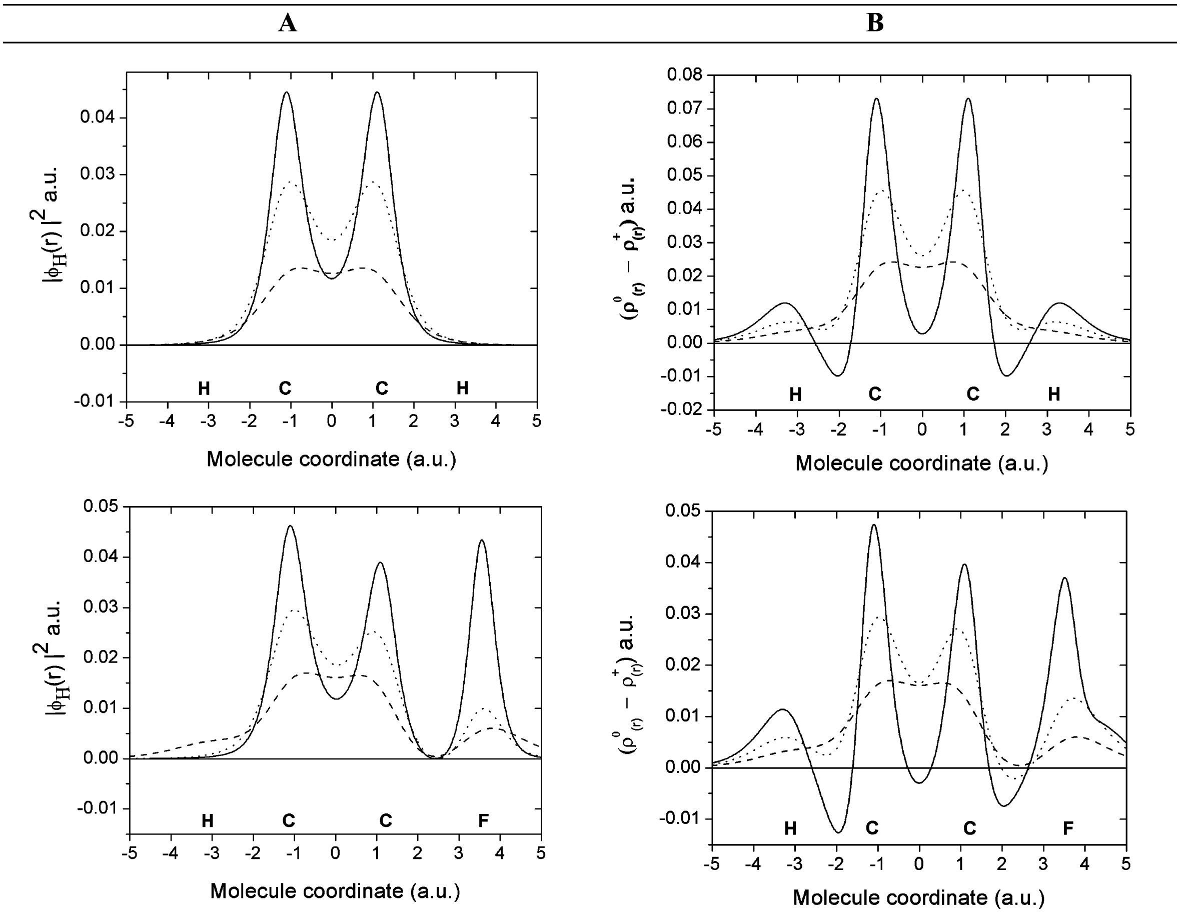


Figure 8. Values of the Fukui function, in atomic units, calculated (A) as the square of the HOMO and (B) as the density differences. Continuous, dot, and dash curves stand for calculated values at distance of 0.5, 1.0, and 1.50 au from the molecular axes.

Table 6. Average Number of Electrons on the Basins of the Donor Fukui Function Calculated as the Square of the HOMO, A, and as the Density Differences, B, at B3LYP/6-311g** and B3LYP/6-311++g**

		A		B	
molecule	atom	BS1[a]	BS2[b]	BS1[a]	BS2[b]
H−C≡C−H	C_1	7.00	7.00	5.83	5.83
	C_2	7.00	7.00	5.83	5.83
H−C≡C−F	C_1	7.18	7.16	6.09	6.05
	C_2	5.96	5.97	5.50	5.55

[a] BS1is B3LYP/6-311g**. [b] BS2 is B3LYP/6-311++g**.

to answer the question is too high. Figure 8 shows the values of the Fukui function calculated at distance of 0.5, 1.0, and 1.50 au from the molecular axes. One can see that, beyond the frozen orbital approximation, model B, the Fukui function shows various minima at distance of 0.5 au which, however, disappears rapidly when the function is evaluated at greater distances with respect to the molecular axes. More specific, there are critical points of different ranges and not precisely an attractor. It is also interesting to note that the attempts to reduce the function to a collection of numbers lost information. Table 6 has the values of the integration of the density over the basins. One can see that the integrated numbers cannot distinguish between both models, A and B. Even though Figure 8 shows clearly that both functions are different.

Resuming, the topological analysis of the Fukui function seems to be an alternative to the condensed version of the Fukui function, and it has the advantage of being mathematically clearly defined, avoiding the ambiguities in the form of condensing the Fukui function.

Acknowledgment. Part of this work has been supported by Fondecyt, Grants 1080184 and 1109043, MECESUP, Grand FSM0605 and “Progama de Sostenilibidad 2009-2010 of Universidad de Antioquia”. We thank discussions with Prof. Paul Ayers and Dr. C. Cardenas about the issues of degenerate states.

References

(1) Parr, R. G.; Donnelly, R. A.; Levy, M.; Palke, W. E. *J. Chem. Phys.* **1978**, *68*, 3801–3807.

(2) Parr, R. G.; Pearson, R. G. *J. Am. Chem. Soc.* **1983**, *105*, 7512–7516.

(3) Parr, R. G.; Yang, W. T. *J. Am. Chem. Soc.* **1984**, *106*, 4049–4050.

(4) Parr, R. G.; Von Szentpaly, L.; Liu, S. B. *J. Am. Chem. Soc.* **1999**, *121*, 1922–1924.

(5) Parr, R. G.; Yang, W. *Density Functional Theory of atoms and Molecules*; Oxford University Press: Oxford, 1989.

(6) Geerlings, P.; De Proft, F.; Langenaeker, W. *Chem. Rev.* **2003**, *103*, 1793–1873.

(7) Chattaraj, P. K.; Sarkar, U.; Roy, D. R. *Chem. Rev.* **2006**, *106*, 2065–2091.

(8) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L. *Phys. Rev. Lett.* **1982**, *49*, 1691–1694.

(9) Ayers, P. W. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3387–3390.

(10) Melin, J.; Ayers, P. W.; Ortiz, J. V. *J. Phys. Chem. A* **2007**, *111*, 10017–10019.

(11) Michalak, A.; De Proft, F.; Geerlings, P.; Nalewajski, R. F. *J. Phys. Chem. A* **1999**, *103*, 762–771.

(12) Ayers, P. W.; De Proft, F.; Borgoo, A.; Geerlings, P. *J. Chem. Phys.* **2007**, 126.

(13) Balawender, R.; Komorowski, L. *J. Chem. Phys.* **1998**, *109*, 5203–5211.

(14) Flores-Moreno, R.; Melin, J.; Ortiz, J. V.; Merino, G. *J. Chem. Phys.* **2008**, 129.

(15) Cardenas, C.; Chamorro, E.; Galvan, M.; Fuentealba, P. *Int. J. Quantum Chem.* **2007**, *107*, 807–815.

(16) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. *Science* **2008**, *321*, 792–794.

(17) Mori, Sánchez, P.; Cohen, A. J.; Yang, W. *J. Chem. Phys.* **2006**, *125*, 201102.

(18) Yang, W.; Mortier, W. J. *J. Am. Chem. Soc.* **1986**, *108*, 5708–5711.

(19) Contreras, R. R.; Fuentealba, P.; Galvan, M.; Perez, P. *Chem. Phys. Lett.* **1999**, *304*, 405–413.

(20) Fuentealba, P.; Perez, P.; Contreras, R. *J. Chem. Phys.* **2000**, *113*, 2544–2551.

(21) Bader, R. *Atoms in Molecules: A Quantum Theory*; Oxford University Press: Oxford, 1990.

(22) Silvi, B.; Savin, A. *Nature* **1994**, *371*, 683–686.

(23) Kato, T. *Comm. Pure Appl. Math.* **1957**, *10*, 151–177.

(24) Steiner, E. *J. Chem. Phys.* **1963**, *39*, 2365–2366.

(25) Chattaraj, P. K.; Cedillo, A.; Parr, R. G. *J. Chem. Phys.* **1995**, *103*, 10621–10626.

(26) Ayers, P. W.; Levy, M. *Theor. Chem. Acc.* **2000**, *103*, 353–360.

(27) Ayers, P. W. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 1959–1964.

(28) Nagy, A.; Sen, K. D. *J. Phys. B: At., Mol. Opt. Phys.* **2000**, *33*, 1745–1751.

(29) Baekelandt, B. G.; Cedillo, A.; Parr, R. G. *J. Chem. Phys.* **1995**, *103*, 8548–8556.

(30) Tiznado, W.; Chamorro, E.; Contreras, R.; Fuentealba, P. *J. Phys. Chem. A* **2005**, *109*, 3220–3224.

(31) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(32) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650–654.

(33) McLean, A. D.; Chandler, G. S. *J. Chem. Phys.* **1980**, *72*, 5639–5648.

(34) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.;O. Kitao; 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.; JohnsonB.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A.; *Gaussian 03*, revision E.01 ed.; Gaussian, Inc.: Wallingford, CT, 2004.

(35) Kohout, M.; Dgrid, version 4.4; *Radebeul*, 2008.

(36) Garza, J.; Vargas, R.; Cedillo, A.; Galvan, M.; Chattaraj, P. K. *Theo. Chem. Acc.* **2006**, *115*, 257–265.

(37) Anderson, J. S. M.; Melin, J.; Ayers, P. W. *J. Chem. Theory Comput.* **2007**, *3*, 375–389.

(38) Angelelli, F.; Aschi, M.; Cacace, F.; Pepi, F.; Depetris, G. *J. Phys. Chem.* **1995**, *99*, 6551–6556.

(39) Ingold, C. K. Electrophilic aromatic Substitution. In *Structure and Mechanism in Organic Chemistry*; Cornell University Press: Ithaca, New York, 1953; pp*223−243*, 278.

(40) Markovnikov, V. *Liebigs Ann. Chem.* **1870**, *153*, 228–259.

(41) Meneses, L.; Tiznado, W.; Contreras, R.; Fuentealba, P. *Chem. Phys. Lett.* **2004**, *383*, 181–187.

(42) Flores-Moreno, R. *J. Chem. Theory Comput.* , *6*, 48–54.

(43) Dewar, M. J. S. *THEOCHEM* **1989**, *59*, 301–323.

(44) Michael, D.; Mingos, P.; Wales, D. J. A Survey of Cluster Chemistry. In *Introduction to cluster Chemistry*; Grimes, R. N. , Ed.; Prentice Hall: Upper Saddle River, NJ, 1990; pp 1−64.

(45) Florez, E.; Tiznado, W.; Mondragon, F.; Fuentealba, P. *J. Phys. Chem. A* **2005**, *109*, 7815–7821.

(46) Fuentealba, P.; Savin, A. *J. Phys. Chem. A* **2001**, *105*, 11531–11533.

(47) Mañanes, A.; Duque, F.; Mendez, F.; Lopez, M. J.; Alonso, J. A. *J. Chem. Phys.* **2003**, *119*, 5128–5141.

(48) Tiznado, W.; Oña, O. B.; Bazterra, V. E.; Caputo, M. C.; Facelli, J. C.; Ferraro, M. B.; Fuentealba, P. *J. Chem. Phys.* **2005**, 123.

(49) Tiznado, W.; Oña, O. B.; Caputo, M. C.; Ferraro, M. B.; Fuentealba, P. *J. Chem. Theory Comput.* **2009**, *5*, 2265–2273.

(50) Bultinck, P.; Carbo-Dorca, R.; Langenaeker, W. *J. Chem. Phys.* **2003**, *118*, 4349–4356.

CT100022W

JCTC Journal of Chemical Theory and Computation

Bound Triplet Pairs in the Highest Spin States of Coinage Metal Clusters

David Danovich and Sason Shaik*

The Institute of Chemistry and The Lise-Meitner Minerva Center for Computational Quantum Chemistry, The Hebrew University, Jerusalem 91904, Israel

Received February 14, 2010

Abstract: The work discusses bonding in coinage metal clusters, $^{n+1}M_n$ (M = Cu, Ag, Au), that have maximum spin without a single electron pair. It is shown that the bonding energy per atom, D_e/n, exhibits a strong nonadditive behavior; it grows rapidly with the cluster size and converges to values as large as 16−19 kcal/mol for Au and Cu. A valence bond (VB) analysis shows that this no-pair ferromagnetic bonding arises from *bound triplet electron pairs* that spread over all the close neighbors of a given atom in the clusters. The bound triplet pair owes its stabilization to the resonance energy provided by the mixing of the local ionic configurations, $^3M(\uparrow\uparrow)^- M^+$ and $M^+ {}^3M(\uparrow\uparrow)^-$, and by the various excited covalent configurations (involving p_z and d_{z^2} atomic orbitals) into the fundamental covalent structure $^3(M\uparrow\uparrow M)$ with a s^1s^1 electronic configuration. The VB model shows that a weak interaction in the dimer can become a remarkably strong binding force that holds together monovalent atoms without a single electron pair.

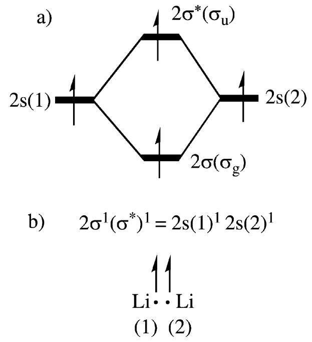
Introduction

No-pair ferromagnetic bonding involves no electron pairing, and the bonding interaction, curiously as it may sound at this point, originates from *triplet electron pairs,* as found, for example, in high-spin alkali metal clusters.[1−6]

To clarify the term no-pair ferromagnetic bonding, consider in Scheme 1a the Li_2 case, where the 2s atomic orbitals form a set of bonding (2σ, i.e., σ_g) and antibonding (2σ* i.e., σ_u) orbitals. In the singlet ground state, the two electrons occupy the bonding orbital to form a Li−Li molecule bound by an electron pair. By contrast, in the triplet $^3\Sigma_u^+$ state, where the electron occupancy is $2\sigma^1 2\sigma^{*1}$, the bond order is formally zero. Indeed, the $2\sigma^1 2\sigma^{*1}$ triplet configuration (Scheme 1a) is equivalent to the purely covalent triplet $2s(1)^1 2s(2)^1$ configuration, in Scheme 1b, where each Li possesses a single electron localized in the respective 2s orbital. This configuration is repulsive and should cause the dissociation of $^3Li_2$. However, the triplet $^3\Sigma_u^+$ state of Li_2 is actually bound,[3] albeit weakly, and the same is true for the other no-pair alkali dimers, which form weakly bonded triplet $^3\Sigma_u^+$ states.[7−10]

As was shown by means of high-level ab initio calculations and valence bond (VB) theory,[3] the weak bonding arises due to the mixing of higher lying ionic and covalent configurations, $2s(1)^1 2p_z(1)^1$, $2s(2)^1 2p_z(2)^1$, and $2p_z(1)^1 2p_z(2)^1$, into the repulsive $2s(1)^1 2s(2)^1$ configuration. Thus, while these additional configurations are high lying, their mixing is still sufficient to overcome the 2s−2s triplet repulsion and to produce a shallow minimum.[3,5,11] As the cluster grows to

Scheme 1. (a) Orbital Mixing of the Pure 2s Atomic Orbitals in $^3Li_2$ and (b) The Equivalence Between $2\sigma^1$ $2\sigma^{*1}$ and $2s(1)^1 2s(2)^1$ Configuration Representations[a]



[a] The symmetry labels of 2σ and 2σ* are indicated in parentheses. The 1 and 2 in parentheses are atom numbers. The Li···Li moiety lies on the *z*-axis.

* Corresponding author e-mail: sason@yfaat.ch.huji.ac.il.

10.1021/ct100088u © 2010 American Chemical Society
Published on Web 04/14/2010

${}^{n+1}Li_n$ with $n > 2$, the number of high-spin ionic and excited covalent configurations increases steeply, so does the binding energy of the cluster, which grows and converges to 12 kcal mol^{-1} per atom, without having a single electron pair.[5] Henceforth, we refer to this type of bonding by the term *no-pair ferromagnetic bonding* (NPFM) bonding. An alternative representation of NPFM bonding was described by McAdon and Goddard,[1] using interstitial orbitals. The two representations are ultimately equivalent.[3,4]

Except for the intellectual interest aroused by this unusual bonding form, some no-pair clusters are real molecular entities, which have actually been made and probed by experimental techniques. Thus, laser-induced emission spectroscopy of the triplet lithium, sodium, potassium, rubidium, and cesium dimers showed a weakly bound ${}^3\Sigma_u^+$ state.[7−9] In fact, there exists spectroscopic evidence also for the no-pair alkali trimer species (${}^4A'$), ${}^4Li_3$, ${}^4Na_3$, and ${}^4K_3$.[12−16] But not only alkali clusters, copper seems also capable of this form of bonding, as may be deduced from the characterization of the ${}^3\Sigma_u^+$ state of ${}^3Cu_2$.[17] As such, these no-pair clusters are real entities, which enrich the scope of chemical bonding, and are, therefore, of wide general interest to chemists. An additional interest in this kind of clusters is their relationship to Bose−Einstein condensates in which the quantum states of all atoms are identical and to Fermi-"gases" of fermionic isotopes of alkali metals, (e.g., K with atomic mass 40) in magnetic fields.[16,18] Finally, having maximum magneticity, no-pair clusters are also interesting for their potential applications in nanochemistry.

The present paper constitutes part of our ongoing program[3−6] to map the territory of these no-pair clusters in the periodic table. Coinage metals,[19−21] which possess valence configurations $nd^{10}(n+1)s^1$ that are analogous in a way to the monovalent alkali metals, seem particularly appealing candidates. Thus, in previous studies we compared the NPFM bonding in the clusters of sodium[6] and copper[22] to those of lithium. The sodium clusters ${}^{n+1}Na_n$ ($n = 2−12$) were found to be much more weakly bound than those of the lithium clusters ${}^{n+1}Li_n$, whereas the ${}^{n+1}Cu_n$ ($n = 2−14$) clusters exhibited stronger bonding, reaching 18−19 kcal/mol per atom. Therefore, the no-pair clusters of the coinage metals may be significantly stickier than the corresponding alkali cluster, and especially so the gold cluster where relativistic effects may contribute to this stickiness.[23] In view of the great surge of interest in gold clusters,[24] an investigation of NPFM bonding in ${}^{n+1}Au_n$ is timely and may be of broad interest. Interestingly, gold surfaces are known to induce sudden magnetization upon adsorption of layers of organic thiols due to formation of "bonded triplet pairs", as proposed by the Naaman et al.[25] The present paper investigates, therefore, NPFM bonding in no-pair coinage clusters of M = Cu, Ag, and Au, and then models the binding energy by a suitable VB model, as done for other no-pair clusters.[3,5,22]

Methods and Details of Calculations

A. Software, Methods, Basis Sets, and Benchmarking. *Software.* All density functional calculations presented here were performed with the Gaussian03 program package.[26] All coupled cluster using single, double, and perturbative triples excitation (CCSD(T)) calculations were carried out with the MOLPRO 2006.1 program package.[27] Ab initio valence bond (VB) calculations were performed with the Xiamen-01 ab initio Valence Bond program[28] using the St-RECP basis set (see below). The VB calculations were used to obtain the repulsive interactions in the purely covalent structure of the dimer with ns^1ns^1 electronic configuration.

Methods and Basis Sets. As has been shown recently,[29] relativistic effects are important for the correct description of properties of the coinage metal clusters and especially for the no-pair states. Therefore, to create a benchmark for the larger clusters, we calculated the ground state and no-pair triplet state of the dimers with the CCSD(T) method using the Douglas−Kroll−Hess (DKH) quasi-relativistic Hamiltonian[30] and the Hamiltonian, which incorporates the relativistic effects via the normalized elimination of small component (NESC) method.[31] Due to program limitation, the truncated (up to f functions) and fully uncontracted aug-cc-pCXZ basis sets of Peterson[32] were used for the NESC method, with $X = T, Q, 5$.[29] By contrast with the former NESC calculations,[29] in the present DKH calculations, we employed Peterson's standard relativistic aug-cc-pVXZ-DK basis.[32] For comparison, we also tested the non relativistic CCSD(T) method with all-electrons non relativistic Peterson correlation consistent basis sets (aug-cc-PVXZ-NR, where $X = T, Q, 5$).[32] Since the use of unrestricted coupled cluster (UCCSD(T)) proved to be too time-consuming, even for the dimer, we used density functional theory (DFT) methods for all the higher clusters.

The DKH−CCSD(T) and NESC−CCSD(T) results provided the benchmark for selecting the appropriate density functional/basis set combinations for calculating the larger clusters. To this end, we have examined the applicability of the different density functionals (PW91, B3P86, B3LYP, TPSS, and BMK) with Peterson-type pseudopotential basis sets (aug-cc-pVXZ-PP, where $X = T, Q, 5$)[33] as well as the B3P86 density functional in combination with the 1997 Stuttgart relativistic small effective core potential (St-RECP) with extended valence basis set (with (8s7p6d)/[6s5p3d] contractions);[34] the latter combination UB3P86/St-RECP was used already in our previous paper on ${}^{n+1}Cu_n$ clusters.[22] Specifically, UB3P86/St-RECP was shown[22] to have a very small basis set superposition error (BSSE) and was compatible with the results of the extended augmented double-ζ atomic natural orbital (ANO) basis set of Roos et al.[35] As seen below, the best combinations were found to be UB3P86/St-RECP and UB3LYP/aug-cc-pVTZ-PP. For the sake of consistency with our previous studies,[22] the present study uses both methods.

Description of the Benchmark Calculations. The dimers in the ground and no-pair triplet states were used to benchmark the DFT methods against available experimental data, where available, as well as relativistic CCSD(T) calculations with all electrons Peterson basis sets and with aug-cc-pVXZ-PP pseudopotential basis sets of Peterson.[33]

Table 1 shows the equilibrium bond lengths (R_e) and bond dissociation energies (D_e) values for Cu_2 in the ground ${}^1\Sigma_g^+$ state and in the no-pair ${}^3\Sigma_u^+$ state with relativistic (RL) and nonrelativistic (NR) CCSD(T) calculations. Inspection of the

Table 1. Results of the Relativistic (RL) and Nonrelativistic (NR) CCSD(T)/aug-cc-pVTZ Calculations of R_e (Å) and D_e (kcal/mol) for the Ground-Singlet and the No-Pair Triplet States of Cu_2

method	R	D_e
	Singlet $^1\Sigma_g^+$	
RL (DKH)	2.224	45.28
NR	2.258	42.51
exptl	2.220[a]	48.208[b]
	Triplet $^3\Sigma_u^+$	
RL (DKH)	2.703	1.328
RL (NESC)	2.711	1.270
NR	2.935	0.735
exptl[c]	2.48	3.46 ± 0.6

[a] From ref 17b. [b] From ref 17c. [c] From ref 17a.

Table 2. R_e (Å) and D_e (kcal/mol) Results for $^3Cu_2$ using Relativistic DKH−CCSD(T) with All-Electron aug-cc-PVXZ Basis Sets and CCSD(T) Calculations with Pseudopotential aug-cc-PVXZ-PP (X = T, Q, 5) Basis Sets

	VTZ[a]	VTZ-PP	VQZ	VQZ-PP	V5Z[a]	V5Z-PP
R_e	2.703	2.689	2.677	2.671	2.662	2.659
D_e	1.328	1.372	1.519	1.614	1.678	1.703

[a] With BSSE correction, R_e = 2.753 Å and D_e = 0.863 kcal/mol for VTZ and 2.670 Å and 1.537 kcal/mol for V5Z.

ground state shows that the difference between relativistic and nonrelativistic calculations is less than 0.04 Å for the equilibrium distance (R_e) and about 3 kcal/mol for the bond dissociation energy (D_e), which is less than 6%. The importance of relativistic effects is much more pronounced for the triplet state of Cu_2. For this state, relativity leads to shortening of R_e by more than 0.2 Å, and it increases the D_e by almost a factor of 2. For the triplet state of the copper dimer, the agreement between two relativistic NESC and DKH methods is quite reasonable (less than 0.01 Å and 0.05 kcal/mol). Furthermore, the match of the relativistic calculations to experimental data is seen to be reasonable.[36]

Table 2 shows the effect of using pseudopotential (PP)-CCSD(T) vs all-electron DKH−CCSD(T) relativistic calculations with matching basis set sizes on the properties of the no-pair state of Cu_2. It is seen that the pseudopotential calculations give reasonable results for the bond length and the bonding energy. The PP calculations always lead to a slightly shorter R_e (0.014 Å for aug-cc-pVTZ-PP and only 0.003 Å for aug-cc-pV5Z basis sets) and a marginally larger D_e value (the discrepancy is less than 0.1 kcal/mol). Therefore, the combination of CCSD(T) with PP basis sets gives good benchmark values that can test DFT methods.

Table 3. R_e (Å) and D_e (kcal/mol) Results for $^3Cu_2$ using CCSD(T) and Different Density Functionals with Pseudopotential Basis Sets aug-cc-pVXZ-PP (X = T, Q, 5) and B3P86/St-RECP

	PW91	B3P86	B3P86	B3LYP	TPSS	BMK	CCSD(T)
				R_e			
pVTZ-PP	2.464		2.532	2.590	2.452	2.606	2.689
pVQZ-PP	2.464		2.532	2.590	2.451		2.671
pV5Z-PP	2.462		2.530	2.586		2.617	2.659
St-RECP		2.607					
				D_e			
pVTZ-PP	12.28		4.295	2.440	11.396	8.036	1.372
pVQZ-PP	12.28		4.284	2.414	11.395		1.614
pV5Z-PP	12.36		4.301	2.452		10.82	1.703
St-RECP		3.047					

Table 4. R_e (Å) and D_e (kcal/mol) Results for $^3Au_2$ using CCSD(T) and Different Density Functionals Methods with Pseudopotential aug-cc-pVXZ-PP (X = T, Q, 5) Basis Sets and B3P86/St-RECP

	B3P86	B3P86	B3LYP	TPSS	BMK	CCSD(T)
			R_e			
pVTZ-PP	2.795		2.875	2.7423	2.920	2.890
pVQZ-PP	2.793		2.872	2.7395	2.917	2.872
pV5Z-PP	2.790		2.869			2.866
St-RECP		2.932				
			D_e			
pVTZ-PP	6.403		3.988	10.724	4.683	3.767
pVQZ-PP	6.421		4.035	10.852	4.812	4.214
pV5Z-PP	6.454		4.094			4.301
St-RECP		4.415				

Tables 3 and 4 compares R_e and D_e values in the no-pair triplet states of Cu_2 and Au_2 calculated with CCSD(T)/aug-cc-pVXZ-PP (X = T, Q, 5) and DFT/aug-cc-pVXZ-PP (X = T, Q, 5) methods. Alongside these, we show the less time-consuming B3P86/St-RECP results. All investigated functionals, except for B3LYP/aug-cc-pVXZ-PP (X = T, Q, 5) and B3P86/St-RECP, are seen to considerably overestimate the bond dissociation energies (D_e) in comparison with the benchmark CCSD(T) values and with the experimental datum[17a] of 3.46 ± 0.6 kcal/mol. It should be noted that the experimentally measured values for the triplet electronic state reported in Table 1 were obtained from laser spectroscopic measurements on a matrix isolated copper dimer.[17a] The uncertainty in such a measurement makes it difficult to estimate, e.g., the matrix effects upon the electronic transitions of Cu_2, which could be significant. For this reason, this experimental D_e value may contain significant uncertainty.[17a] Since the bond dissociation energy is the key factor in our study, we rely henceforth on the UB3LYP/aug-cc-pVTZ-PP and UB3P86/St-RECP for calculating the larger clusters.

B. Geometry Optimization, State Identification, and Bond Dissociation Energy Calculations. *Geometry Optimization.* Different structures with different state symmetries were tested for each cluster size in order to find the most stable clusters. All calculations discussed here are the result of a full geometry optimization followed by the usual test for genuine minima using frequency calculations.

Tests for State Identity. For every cluster we used the TDDFT method, as a stability check, to ascertain the lowest energy solution for the ground and the no-pair states. The TDDFT tests were carried out with NWCHEM 5.1[37] and Gaussian03 programs. In the event where the tested state did not have the lowest solution in the TDDFT calculation, a new guess function was examined, and the geometry was reoptimized until the TDDFT calculation converged to the lowest one.

Since there is no guarantee that the no-pair states will be the lowest state of a given multiplicity for a particular cluster, we routinely verified that the singly occupied orbitals in the state of choice were *only* the σ-types. These orbitals were examined in two ways: (i) Initially, using canonical

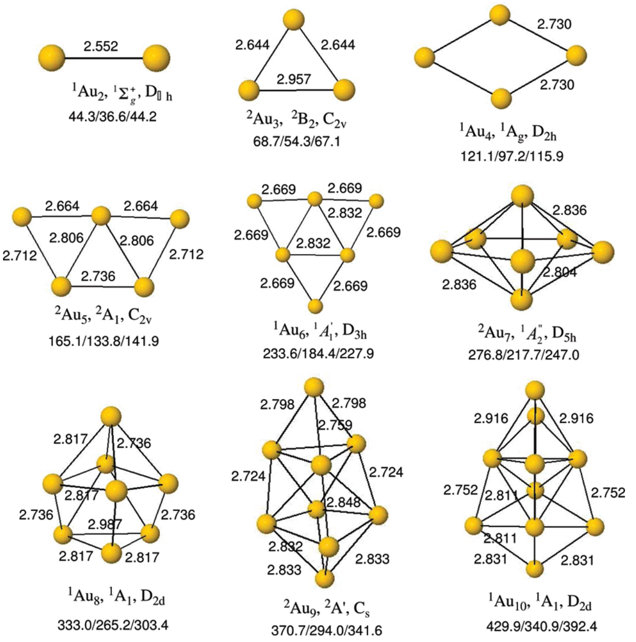


Figure 1. UB3P86/St-RECP optimized structures with their point groups and state assignments for the most stable coinage metal clusters in their ground electronic states. The bond length values are shown only for Au_n. Bond dissociation energies (D_e, in kcal/mol) are shown below the structures in the order Cu/Ag/Au.

Kohn−Sham (KS) orbitals, we made sure that no radial orbitals (perpendicular to the surface of the cluster and, hence, not of a σ M−M character) were singly occupied. (ii) Subsequently, these singly occupied KS orbitals were localized, and the resulting orbitals were ascertained to be largely (98%) confined to a single atom and to be dominated by the highest s-type AO of Cu, Ag, or Au atoms.

Bond Dissociation Energy Calculations. The dissociation energies and the dissociation energies per atom, D_e and D_e/n, were corrected for BSSE. BSSE on the D_e and D_e/n values were found using the counterpoise method (using the keyword counterpoise = n [n is a number of coinage atoms in the cluster] in Gaussian-03). In our previous study,[22] it was found that BSSE values are very small and do not change the behavior of D_e/n vs the cluster size n. Thus, the BSSE corrected and uncorrected of D_e/n vs n plots are virtually parallel to one another, different by almost a constant quantity (see Supporting Information, Figure S1). For this reason, in the present analyses, we used BSSE uncorrected values of D_e/n for all coinage metals. The study generated many results that are summarized in the Supporting Information.

Results

A. Ground State Structures of Coinage Metal Clusters. The geometry optimization in the ground and no-pair states revealed several "geometric isomers" for each studied cluster size. Figure 1 displays the UB3P86/St-RECP optimized structures of the most stable geometric isomers of the coinage metal clusters, M_n, where n varies from 2 to 10. The point group symmetries are common to all the M_n clusters for a given n. Therefore, to compact the information, we show in Figure 1 bond lengths only for the Au_n clusters, while the rest of the bond lengths and other geometric parameters are given in the Supporting Information, Tables S1−S3 and Figures S2 and S3.

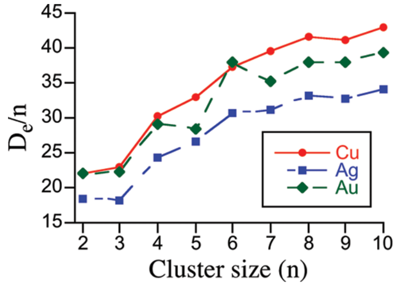


Figure 2. Dependence of the bond dissociation energies per atom (D_e/n, in kcal/mol) on the cluster size for the ground states of the coinage metal clusters, M_n. Copper data is in red, silver is in blue, and gold is in green.

It is seen from Figure 1 that for $n = 2-6$, the most stable isomers of the ground-state M_n clusters are two-dimensional (2D) structures. Starting with $n = 7$ and on to larger clusters, the most stable structures are seen to be three-dimensional.[38] These point-group symmetries and geometries are virtually the same as those published recently for the ground states of Ag_n,[39] using the spin-unrestricted Perdew and Wang (PW91) density functional method implemented in the Demon-KS3P5 program package with the all-electron orbital basis set contracted as (633321/53211*/521+) and in conjunction with the corresponding (5,5;5,5) auxiliary basis set for describing the s, p, and d orbitals. This match of the results for different functionals implies that the most stable structures obtained in our calculations are most likely independent of the density functionals and basis sets.

The numbers underneath the structures in Figure 1 correspond to the total D_e, in kcal/mol, in the order Cu/Ag/Au. It is seen that the ground state of the coinage metal clusters, even for the dimer, has significant bonding energies. The D_e increases considerably by about 10 times, reaching values of 340−430 kcal/mol for $n = 10$.

If, however, we consider the bond dissociation energy per atom (D_e/n), which is one of the measures of cluster stability, we find that the D_e/n quantity does not change as drastically as the total D_e. As can be seen in Figure 2, the D_e/n increases by less than a factor of 2 from the dimer to the M_{10} cluster and reaches the value around 40 kcal/mol. As shall be seen later, this is very different from the behavior we found for the no-pair states of the coinage metal clusters.

One of the important geometric features of the cluster is an average bond length between the first-neighbor bonded atoms. Since all the M_n clusters show similar trends, we have shown in Figure 3 the variation of this distance only for Au_n clusters. It is seen that the average bond length of the ground-state clusters falls into three distinct areas: the dimer, which has the shortest bond length, the planar clusters with $n = 3-6$, possessing intermediate bond lengths, and the three-dimensional (3D) clusters with $n > 6$, which exhibit the longest bond lengths. Thus, the average distance depends on the dimensionality of the clusters, and the changes within each group are smaller than between the groups. Comparison

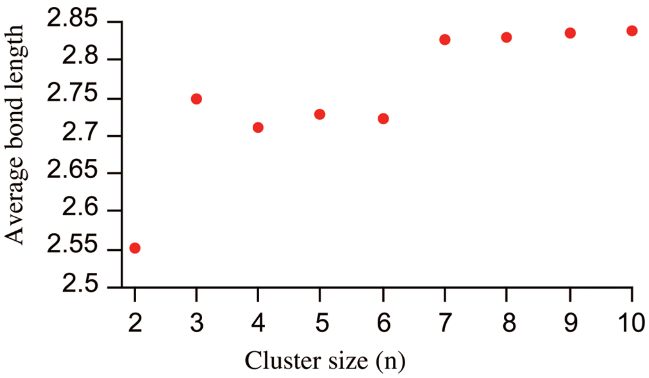


Figure 3. Dependence of the average first-neighbor bond lengths (in Å) of ground state of gold clusters plotted against the cluster size.

of Figures 2 and 3 shows that D_e/n increases as the average bond lengths between the first-neighbor-bonded atoms increases. The reason for this seemingly counterintuitive behavior is because as one moves from the dimer to the large clusters, the number of the bonds per atom increases; thus, the average number of the bonds per atom is 0.5 for the dimer, 1−1.5 for the 2D clusters, and 2−2.4 for the 3D-clusters, which belong to the 3D group. As such, while each particular bond in the large clusters becomes weaker, as in the smaller ones, the total bond dissociation energy increases considerably due to the much larger number of the bonds. Later this behavior too will be contrasted with the trend exhibited by the no-pair state of the coinage metal clusters.

B. Structures of Coinage Metal Clusters in the No-Pair State. In contrast to the ground state, where 3D sets in only for clusters with $n \geq 7$, the most stable no-pair state structures have 3D geometries starting already with $^5M_4$, which is tetrahedral.[40] These clusters are depicted in Figure 4, which displays the UB3P86/St-RECP geometries; the UB3LYP/aug-cc-pVTZ-PP geometries follow the same trends and are relegated to the Supporting Information (Figures S6 and S7 and Tables S7 and S8). The bond lengths are indicated in Figure 4 only for $^{n+1}Au_n$, and the data for the other metals are given in the Supporting Information (Tables S4−S6 and Figures S4 and S5).

As we found for the ground state, the most stable geometries of the no-pair clusters in Figure 4 are very similar for all coinage atoms. Only two exceptions exist for the clusters with $n = 5$ and 6, where the most stable isomer depends on the metal, as seen in Figure 5. Thus, in the case of $^6M_5$, one finds trigonal-bipyramidal or square-pyramidal structures depending on M, but the difference in the total energy of the two isomers for a given metal is small, e.g., for $^6Au_5$ the total energy difference is only 2.8 kcal/mol. The same applies to $^7M_6$: the total energy difference between the alternative structures for a given metal is small. The unusual C_{2v} structure of $^7Cu_6$ cluster was created by adding the sixth copper atom to one of the faces of the trigonal-bipyramidal structure, while keeping other bonds, as in the D_{3h} structure of $^6Cu_5$ cluster.

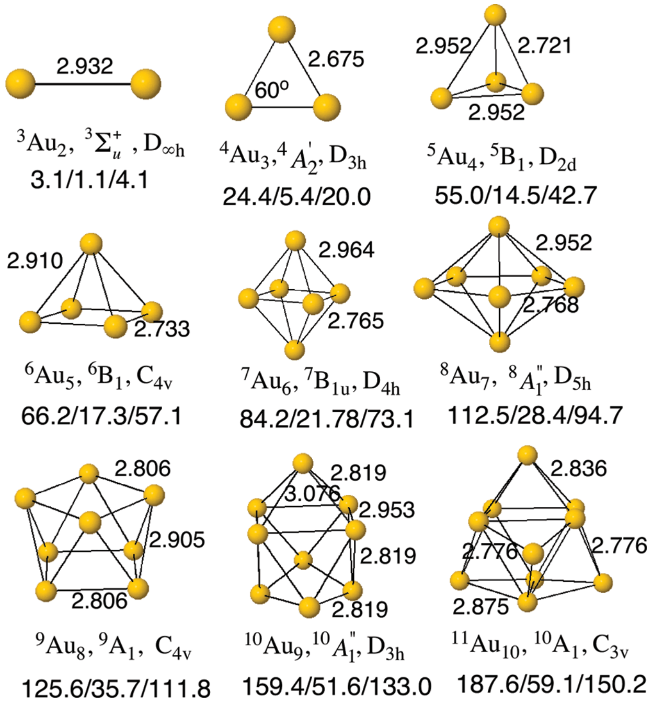

2.932
^{3}Au$_2$, $^3\Sigma_u^+$, D$_{\infty h}$
3.1/1.1/4.1
2.675
60°
^{4}Au$_3$,$^4A_2'$, D$_{3h}$
24.4/5.4/20.0
2.952
2.721
2.952
^{5}Au$_4$, ^{5}B$_1$, D$_{2d}$
55.0/14.5/42.7
2.910
2.733
^{6}Au$_5$, ^{6}B$_1$, C$_{4v}$
66.2/17.3/57.1
2.964
2.765
^{7}Au$_6$, ^{7}B$_{1u}$, D$_{4h}$
84.2/21.78/73.1
2.952
2.768
^{8}Au$_7$, $^8A_1''$, D$_{5h}$
112.5/28.4/94.7
2.806
2.905
2.806
^{9}Au$_8$, ^{9}A$_1$, C$_{4v}$
125.6/35.7/111.8
2.819
3.076
2.953
2.819
2.819
^{10}Au$_9$, $^{10}A_1''$, D$_{3h}$
159.4/51.6/133.0
2.836
2.776
2.776
2.875
^{11}Au$_{10}$, ^{10}A$_1$, C$_{3v}$
187.6/59.1/150.2

Figure 4. UB3P86/St-RECP optimized structures, their point groups, and state assignments for the most stable coinage metal clusters in the no-pair high-spin states, $^{n+1}M_n$. The bond length values are shown only for $^{n+1}Au_n$. Bond dissociation energies (D_e, in kcal/mol) are shown below the structures in the order Cu/Ag/Au.

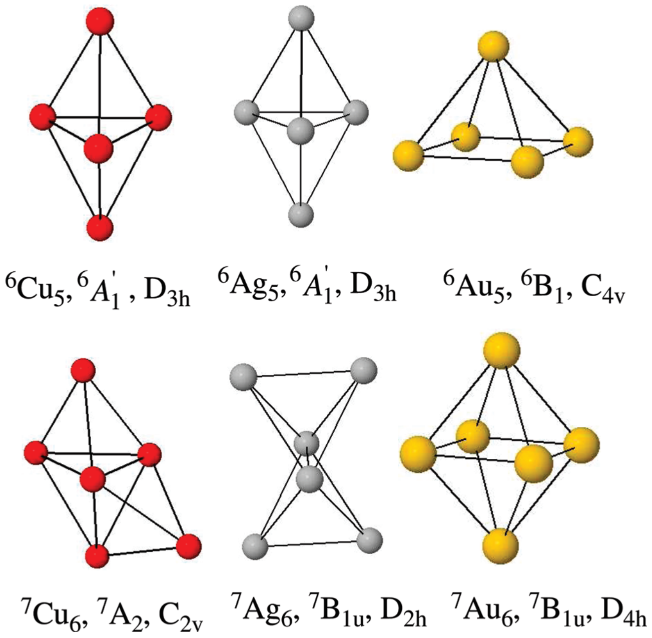

^{6}Cu$_5$, $^6A_1'$, D$_{3h}$
^{6}Ag$_5$, $^6A_1'$, D$_{3h}$
^{6}Au$_5$, ^{6}B$_1$, C$_{4v}$
^{7}Cu$_6$, ^{7}A$_2$, C$_{2v}$
^{7}Ag$_6$, ^{7}B$_{1u}$, D$_{2h}$
^{7}Au$_6$, ^{7}B$_{1u}$, D$_{4h}$

Figure 5. UB3P86/St-RECP optimized structures for $^6M_5$ and $^6M_5$ clusters, their point groups, and state assignments.

At present, we do not have an explanation for these structural variations for $n = 5$ and 6, and perhaps such an explanation is not warranted, since the relative energies of the isomers are very close. For all other clusters, the structures are metal-independent and are close to being highly symmetrical species. For example, the D_{2d} symmetry of the $^5M_4$ clusters (Figure 4) is actually very close to a pure tetrahedral T_d symmetry. Indeed, except for the clusters with $^8M_7$, generally the point-group symmetry of the no-pair clusters is always higher than the symmetry of the corresponding clusters in their ground state. Moreover, in previous alkali metal clusters, we noted the same phenomenon,[3,22]

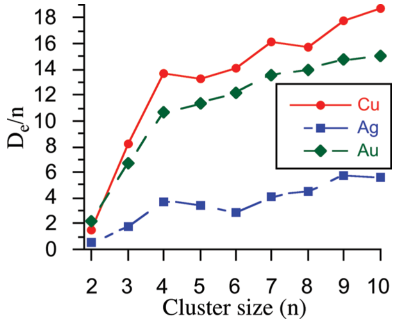


Figure 6. Dependence of the bond dissociation energies per atom, D_e/n (in kcal/mol), on cluster size for the no-pair states of the coinage metal clusters.

and hence we are seeing a topological behavior of NPFM bonding, which seems to transcend the identity of the metal. In fact, as we argued before,[3,22] this behavior of the no-pair state is predictable based on VB theory, which shows that the high-spin state clusters attempt to create structures that maximize the coordination number for each atom in the cluster and minimize the repulsive interaction in the fundamental all-s[1] configuration. Generally, the minimized repulsive interaction requires identical bond lengths, and a maximal coordination number means also that the structure has the most symmetrical 3D packing.

It is also interesting to point out the difference between some structures calculated previously with the UB3P86/LANL2DZ method[22] and those obtained in present paper at the B3P86/St-RECP level. There are two main differences. First, the $^6Cu_5$ cluster was previously assigned a C_{4v} point-group symmetry, and now it has D_{3h} symmetry for the most stable structure. At the B3P86/St-RECP level, the difference in the total energy between two structures is 3.3 kcal/mol in favor of D_{3h} symmetry. Second, the $^9Cu_8$ cluster, had a C_{2v} symmetric structure in our previous work,[22] and here it has C_{4v} symmetry, but the difference in total energies between C_{2v} and C_{4v} point group symmetries is negligible. Clearly, the no-pair clusters may exhibit also some fluxionality in their structures.

The total D_e values noted in Figure 4, underneath the structures, exhibit a behavior very different than the trend in ground-state clusters. Thus, D_e starts from a very small value for dimers and increases steeply for the larger no-pair clusters, reaching values that are 50-fold larger than the D_e for the dimers. For example, the D_e for the no-pair dimers of copper and gold is only 3−4 kcal/mol, and it reaches values around 170 kcal/mol for the $^{11}M_{10}$ (M = Au, Cu). As will be discussed, this nonadditive behavior can also be predicted using our VB model of bonding.[3,22]

Let us consider now the D_e/n quantity, which is a measure of the cluster stability. As can be seen from Figure 6, D_e/n increases dramatically by about 10−15 times in contrast with the ground state, where it increased by less than a factor of 2. Thus, the value starts from less than 1 kcal/mol for the dimers and reaches 18 and 15 kcal/mol for copper and gold atoms, respectively, which are remarkably high binding energies for clusters with no electron pairing. The no-pair silver clusters are rather weakly bonded, and the D_e/n value converges to less than 6 kcal/mol at $^{11}Ag_{10}$. In all the series, the steepest increase of D_e/n occurs in the transition from the dimer to the trimer. Another significant increase of D_e/n occurs also between the trimer and the tetramer. According to our VB model, which will be discussed later, this behavior is associated with the very significant increase in the total coordination number of the cluster; from 2 for $^3M_2$ to 6 for $^4M_3$ and then to 12 for $^5M_4$. Further changes in the total coordination number become less considerable for the larger clusters (18 for $^6M_5$, 24 for $^7M_6$, and 30 for $^8M_7$).

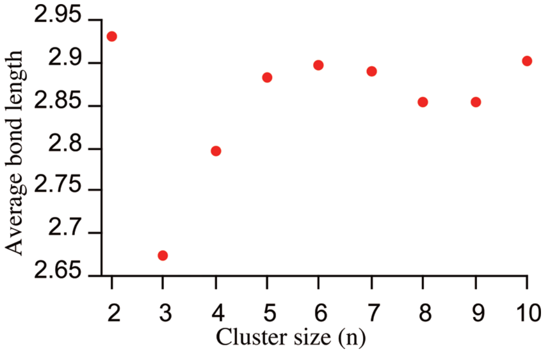


Figure 7. Dependence of the average bond length between the first-neighbor-bonded atoms (in Å) for gold clusters in the high-spin state plotted against the cluster size.

Table 5. D_e/n Values (in kcal/mol) Calculated for No-Pair States, $^{n+1}M_n$, of the Copper and Gold Clusters Using UB3LYP/aug-cc-pVTZ-PP and UB3P86/St-RECP Methods

	UB3LYP/ aug-cc-pVTZ	UB3P86/ St-RECP		UB3LYP/ aug-cc-pVTZ	UB3P86/ St-RECP
Cu_2	1.22	1.52	Au_2	1.99	2.21
Cu_3	9.07	8.14	Au_3	9.33	6.67
Cu_4	12.94	13.73	Au_4	11.58	10.68
Cu_5	11.39	13.23	Au_5	11.73	11.41
Cu_6	12.29	14.03	Au_6	11.42	12.18
Cu_7	13.80	16.07	Au_7	12.54	13.52
Cu_8	13.49	15.59	Au_8	12.90	13.98
Cu_9	14.85	17.71	Au_9	13.25	14.78
Cu_{10}	15.68	18.76	Au_{10}	13.46	15.02

The dependence of the average bond length between first-neighbor-bonded atoms on the size of the cluster in the no-pair states of the gold is presented in the Figure 7. The trend is again similar for all other coinage metals, and for this reason, we present here only the results for the gold clusters. As we pointed out above, this dependence in the no-pair state is very different from the trend we found for the ground state. The longest bond distance is observed in the dimer, and it decreases significantly, by about 0.3 Å, for the trimer, which has the shortest bond length among all cluster sizes. For the tetramer, the average bond length increases by about 0.15 Å, and starting from *n* = 5, the average Au−Au distance changes very little in the range of 2.85−2.90 Å.

Discussion

Bonding in the ground-state clusters is relatively easy to understand since it derives from electron pairing and from delocalization of the electrons.[22,41] The intriguing finding is the significant bonding in the states where all the spins are up. To aid the discussion, we display in Table 5 the computed bond dissociation energy per atom (D_e/n) for the no-pair

Table 6. B3P86/St-RECP Calculated and VB Model Estimated BDE/n (kcal/mol) for the No-Pair States of Coinage Metal Clusters

	BDE/*n*					
	Cu		Ag		Au	
n	DFT	model[a]	DFT	model[b]	DFT	model[c]
2	1.52	1.72	0.54	0.84/0.15	2.21	2.16
3	8.14	8.75	1.80	2.34/2.67	6.67	7.09
4	13.73	12.12	3.63	3.38/3.56	10.68	10.11
5	13.23	14.15	3.45	4.01/4.09	11.41	11.65
6	14.03	15.51	3.64	4.07/4.10	12.18	13.14
7	16.07	16.47	4.05	4.73/4.70	13.52	14.01
8	15.70	15.34	4.46	4.41/4.37	13.98	13.05
9	17.71	16.93	5.73	4.88/4.81	14.78	14.45
10	18.76	18.21	5.91	5.26/5.16	15.02	15.56
11		18.58		5.38/5.25		15.89
12		19.50		5.65/5.51		16.70
13		19.43		5.63/5.48		16.65
14		19.37		5.62/5.46		16.59

[a] $\delta\varepsilon_{\text{rep}}$ = 16.42 and $\delta\varepsilon_{\text{mix}}$ = 1.3245 kcal/mol for Cu. [b] The values for Ag correspond to the two alternative fits: the two-parameter fit ($\delta\varepsilon_{\text{rep}}$ = 0.8029 and $\delta\varepsilon_{\text{mix}}$ = 0.1653 kcal/mol) and the one-parameter fit ($\delta\varepsilon_{\text{rep}}$(VB) = 8.602 and $\delta\varepsilon_{\text{mix}}$ = 0.5931 kcal/mol). [c] $\delta\varepsilon_{\text{rep}}$ = 6.0256 and $\delta\varepsilon_{\text{mix}}$ = 0.6902 kcal/mol for Au.

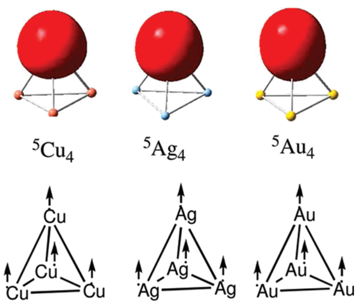


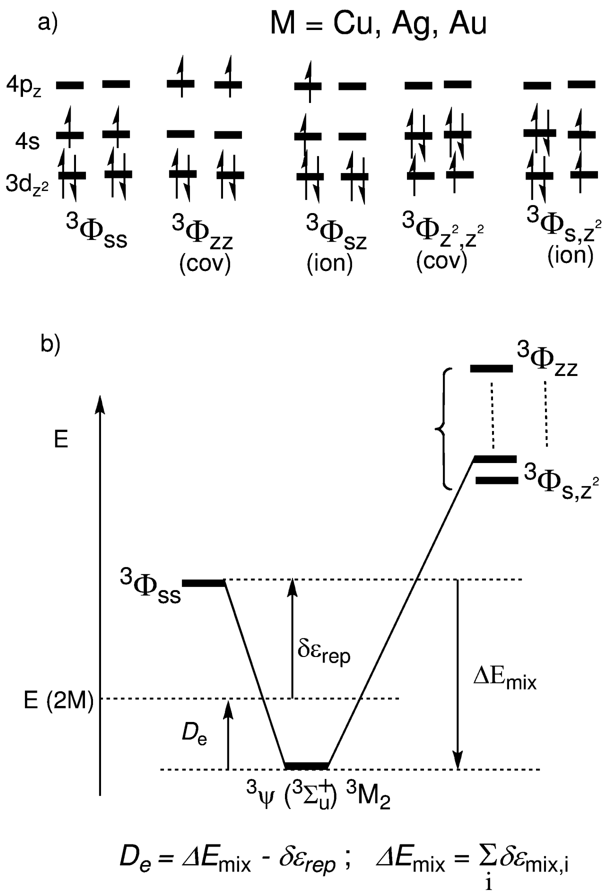
Figure 8. Localized orbitals for $^5M_4$; M = Cu, Ag, Cu. For each cluster, we show only one orbital. Below these orbitals are schematic representations of the electronic structures of these clusters, using dots to represent the electrons and arrows to represent the spin.

states of copper and gold clusters at the UB3LYP/aug-cc-pVTZ-PP and UB3P86/St-RECP levels of theory. The values for silver were obtained only at the UB3P86/St-RECP level and are shown in Table 6.

Inspection of Table 5 shows that, except for the $^4Au_3$ case, the agreement between both methods is quite reasonable, with UB3P86/St-RECP giving a generally larger D_e/n. The converged D_e/n values at $^{11}M_{10}$, for M = Cu and Au, are very impressively large, and they are certainly not weak van der Waals interactions; they are more in the realm of chemical bonds. These values become all the more impressive when one looks at the localized orbitals of the no-pair clusters. Figure 8 shows one of these orbitals for the $^5M_4$ clusters (M = Cu, Ag, Au) which, according to the natural localized molecular orbital (NLMO) analyses,[41] are 98.2% localized with small tails on other atoms. Thus, it is apparent that the electronic structure is largely localized with one electron per site, all the electrons having parallel spins, and that the local orbitals have dominant *ns* (*n* = 4−6 for M = Cu, Ag, Au) characters with some outward hybridization.

So where does the NPFM bonding come from? Why does it get so strong as the cluster increases and then converges very quickly at about *n* = 10? Can we account for the jumps in the D_e/n quantity? What happens as we change M from Cu to Ag and then to Au? Is it possible to find a rationale for the symmetric clusters and for the high coordination numbers that typify these clusters? This will be done by using VB theory and by modeling of the D_e/n quantity.

Scheme 2. (a) Some of the VB Configurations That Contribute to NPFM Bonding in the No-Pair Dimers[a] and (b) The Corresponding VB-mixing diagram[b]



[a] $^3M_2(^3\Sigma_u^+)$ for M = Cu, Ag, and Au. [b] The equation in part (b) is the bond dissociation energy (D_e) expression.

A. Valence Bond Analyses of the NPFM Bonding in the No-Pair States of the Coinage Metal Clusters. As was argued previously,[3,6,22] NPFM bonding originates due to the *ionic-covalent fluctuations of the triplet pairs*. The various types of VB structures that contribute to the wave function of the $^3M_2$ coinage metal dimers as well as the corresponding VB mixing diagram that leads to NPFM bonding are shown in Scheme 2.

Thus, as shown in Scheme 2a, the fundamental configuration is the covalent $^3\Phi_{s,s}$ with the two valence electrons in the (*n*+1)s AOs (4s, 5s, 6s) of the two coinage metal atoms. There are higher-lying VB structures, which involve singly occupied nd_{z^2} (*n* = 3−5) and (*n+1*)p$_z$ AOs. Some of these are ionic triplet configurations, like $^3\Phi_{s,z^2}$, which involves electron transfer from the nd_{z^2} AO of one metal to the (*n*+1)s AO of the second or from $^3\Phi_{s,z}$, which involves an electron transfer from the (*n*+1)s AO of one atom to the (*n*+1)p$_z$ AO of the second. In addition, there are excited covalent configurations, where the two valence electrons occupy the $(n+1)\mathrm{p}_z$ AOs of the two atoms, as in ${}^3\Phi_{z,z}$, or the $n\mathrm{d}_{z^2}$ AOs of the two atoms, as in ${}^3\Phi_{z^2,z^2}$. By itself, the fundamental ${}^3\Phi_{\mathrm{s,s}}$ configuration is purely repulsive, and the repulsive term $\delta\varepsilon_{\mathrm{rep}}$, in Scheme 2b, arises from the two triplet electrons as well as from the $\mathrm{d}^{10}-\mathrm{d}^{10}$ closed shell Pauli repulsions. The NPFM bonding will arise only from the mixing of the excited ionic and covalent configurations, each of which contributes a $\delta\varepsilon_{\mathrm{mix,i}}$ element as shown in Scheme 2b. Thus, at a given M−M distance, the net NPFM bonding will be a balance between the repulsive interactions in the fundamental structure ($\delta\varepsilon_{\mathrm{rep}}$) and the sum of the mixing interaction terms due to all the excited configurations ($\Delta E_{\mathrm{mix}} = \sum_i \delta\varepsilon_{\mathrm{mix,i}}$). As we move to larger and larger clusters, there will always be one fundamental configuration with a singly occupied ns orbital for each atom, ${}^{n+1}\Phi_{\mathrm{s1,s2,\ \ldots,\ s}n}$. However, now the number of excited ionic and covalent configurations increases in a nonlinear manner, since each atom can have ionic and covalent triplet configurations with each neighboring atom, thus dramatically augmenting the stabilization energy.

As we showed previously,[3,6] the bond dissociation energy (D_{e}) due to NPFM bonding can be expressed in a simple analytical form. Thus, assuming that the elementary repulsion term ($\delta\varepsilon_{\mathrm{rep}}$) is the same for all pairs of bonded atoms and that it involves only the close neighbor atoms, this allows us to use the repulsion term extracted from a VB calculation of the respective dimer molecule and to evaluate the total repulsion by multiplying this pair repulsion by the number of close neighbor M···M pairs in the cluster. In addition, assuming for simplicity that the various excited configurations contribute each an identical close-neighbor mixing term $\delta\varepsilon_{\mathrm{mix,i}}$, which is the same as in the corresponding dimer ${}^3\mathrm{M}_2$, allows us to evaluate the total mixing term for any cluster size. This is done by simply counting the number of ionic and covalent excited configurations a given atom has with its close neighbors and by multiplying the resulting number of configurations by the elementary mixing term. We further truncate the number of excited configurations to the lowest excitations involving electron shifts from the nd AOs to singly occupied $(n+1)$s AOs and from $(n+1)$s to $(n+1)$p.

These simplifications allow us to model the NPFM-binding energy based on eq 1:

$$D_{\mathrm{e}} = \left[\frac{(N_{\mathrm{AO}}^2 + 9)C_{\mathrm{tot}}}{2} + N_{\mathrm{AO}}\right]\delta\varepsilon_{\mathrm{mix}} - \frac{C_{\mathrm{tot}}\delta\varepsilon_{\mathrm{rep}}}{2} \qquad (1)$$

Here N_{AO} is the number of singly occupied and virtual AOs (per atom) that participate in the populating of the n-electrons in ${}^{n+1}\mathrm{M}_n$, while C_{tot} is the total coordination number that sums all the close neighbors of all atoms in the cluster. Only s- and p-AO's are counted for N_{AO}. Details of the deriving equaition 1 can be found in the Supporting Information, VB Model equations section.[22]

Due to limitations of the VB software,[28] we are able to calculate only the energy of the fundamental configuration, ${}^3\Phi_{\mathrm{s,s}}$, and evaluate thereby the pair repulsion term in the ${}^3\mathrm{M}_2$ species. The mixing terms are then obtained by least-squares fitting of eq 1 to the computed D_{e}/n quantities. To test the fit quality, we used a second method whereby we fit eq 1 by least-squares fitting of both the repulsion and the mixing terms. We have checked both approaches and obtained results that are quite similar and are all given in the Supporting Information (Figures S8−S11).

Using UB3P86/St-RECP D_{e}/n data, the best-fitted pair-repulsion terms are 16.42 for Cu, 6.0256 for Au, and 0.8029 kcal/mol for Ag atoms. The values for ${}^{n+1}\mathrm{Cu}_n$ and ${}^{n+1}\mathrm{Au}_n$ are much larger than those obtained for ${}^{n+1}\mathrm{Li}_n$, which makes physical sense since the coinage metals have in addition $\mathrm{d}^{10}-\mathrm{d}^{10}$ repulsive terms, which the Li cluster does not have. Indeed, the VB calculated repulsion terms are 14.28 for Cu (at $R_{\mathrm{e}} = 2.60$ Å) and 11.76 kcal/mol for Au (at $R_{\mathrm{e}} = 2.93$ Å), which, while not identical to the fitted values, are large, in the right order, and much larger than the corresponding elementary repulsion calculated by VB for Li (1.504 kcal/mol[3]). The smaller repulsion of Au vs Cu may well reflect the relativistic shrinkage of the 6s orbitals of Au, which lower the $6\mathrm{s}^1-6\mathrm{s}^1$ repulsion. The fitted $5\mathrm{s}^1-5\mathrm{s}^1$ repulsive term of Ag is much too small, and this may reflect the poorer quality of the fit. Indeed, the VB calculation for the $5\mathrm{s}^1-5\mathrm{s}^1$ fundamental structure of ${}^3\mathrm{Ag}_2$ gives significant values, which depends on the equilibrium distance taken for the dimer; 8.602 (at $R_{\mathrm{e}} = 3.12$ Å) and 5.902 kcal/mol (at $R_{\mathrm{e}} = 3.35$ Å, which is obtained with CCSD(T)/St-RECP calculations).

The best-fitted elementary mixing terms are 1.3245, 0.6902, and 0.1653 kcal/mol for Cu, Au, and Ag atoms. The relative ordering of these values is in line with the calculated d−s orbital energy gaps for the atoms (see Supporting Information, Tables S9 and S10). According to the VB mixing model (Scheme 2), larger gaps will result in small mixing terms and vice versa for smaller gaps. The gaps can in turn be understood based on various effects that have been discussed by Pyykkö and Desclaux.[42] Thus, Cu is affected by the "$3\mathrm{d}^{10}$ contraction" and hence the 3d−4s gap should be smaller than the 4d−5s gap in Ag, while in Au, the 5d−6s gap shrinks relative to Ag by the relativistic shrinkage.[42] Further support of these considerations is provided by inspecting the σ and σ^* orbitals of the M_2 dimers, which shows that Cu has the largest d contribution, while Ag has the smallest. All the dimers have as well $(n+1)$p contributions to the σ and σ^* orbitals, but these contributions are quite similar to the three atoms. Thus, the order of the mixing terms ($\delta\varepsilon_{\mathrm{mix}}$), obtained from the two-parameter fit, is physically reasonable. Again, the $\delta\varepsilon_{\mathrm{mix}}$ obtained for Ag from the two-parameter fit procedure may seem very small. Using the VB computed $\delta\varepsilon_{\mathrm{rep}}$ gave values of $\delta\varepsilon_{\mathrm{mix}} = 0.5931$ kcal/mol, which is still smaller than the corresponding values for Cu and Au.

Inserting these fitted values into the eq 1 enables us to calculate D_{e}/n values for the larger clusters. All the D_{e}/n values are collected in Table 6, along with the UB3P86 computed ones. Figure 9 shows plots of the VB modeled and UB3P86 computed D_{e}/n values for the ${}^{n+1}\mathrm{Cu}_n$, ${}^{n+1}\mathrm{Au}_n$, and ${}^{n+1}\mathrm{Ag}_n$ clusters vs the cluster size n.

It is apparent from Table 6 that the quality of the fit is very good for Cu and Au, having R^2 values of 0.97 and 0.98 for Cu and Au, respectively, and of a lesser quality for Ag atoms with $R^2 = 0.89$. However, taking the series together, it is clear that the VB model describes well the entire pattern

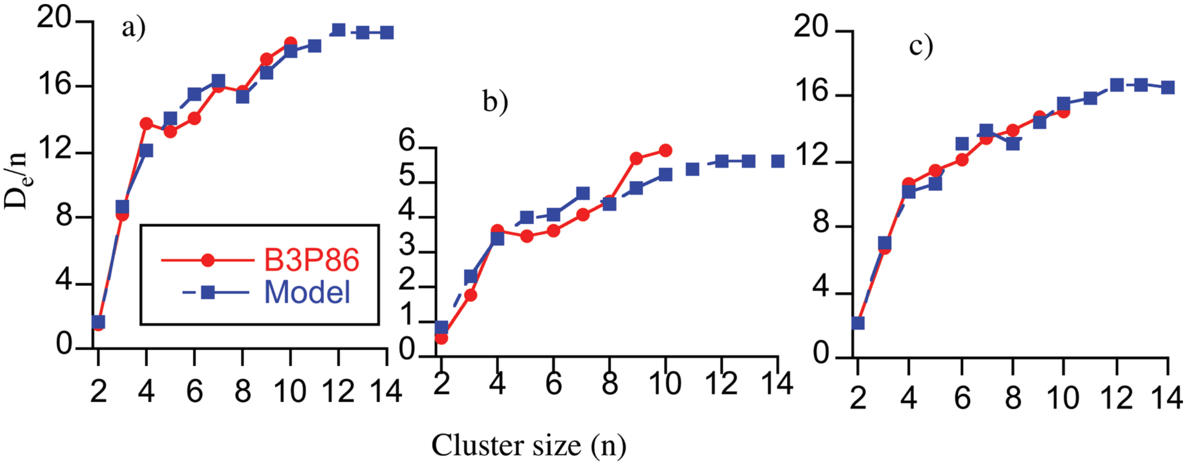


Figure 9. Fit between UB3P86/RECP calculated (red) for (a) copper, (b) silver, and (c) gold clusters and VB model estimated D_e/n (blue) (in kcal/mol) as function of cluster size.

of NPFM bonding in these three coinage metals. Furthermore, as shown in the Supporting Information (Figures S8−S11), the quality of the fits is retained with different combinations of the parameters. Furthermore, inspection of Figure 9 shows that the VB-modeled D_e/n curve fits nicely the UB3P86 calculated one. The VB-modeled curve reproduces the steep rise of the D_e/n observed when moving from $^3M_2$ to $^{n+1}M_n$ clusters, and it converges at approximately 19.5, 16.5, and 5 kcal/mol for the copper (Figure 9a), the gold (Figure 9c), and the silver (Figure 9b), respectively.

The VB model can be used also to account for the seemingly odd behavior of the Cu vs Au clusters. Thus, as can be seen in the Tables 5 and 6, $^3Au_2$ is more strongly bound than $^3Cu_2$. By contrast, as the cluster grows, the trend is reversed, and for the $^{n+1}M_n$ clusters with $n > 10$, the Cu clusters are more strongly bonded with a converged D_e/n value of 19.4 kcal/mol relative to 16.6 kcal/mol for the Au clusters. The VB model in Scheme 2b and eq 1 nicely explains this reversal. Thus, eq 1 shows that the total D_e is a balance between the mixing and repulsion terms with a larger multiplier for the $\delta\varepsilon_{mix}$ term (in eq 1) compared with that of the $\delta\varepsilon_{rep}$ term. This larger multiplier signifies the fact that the number of excited VB configurations, which can mix into the fundamental VB structure, increases much faster than the number of pair repulsions as the cluster grows. Thus, the dimer $^3Cu_2$, with the larger repulsive term, is more weakly bound compared with $^3Au_2$. However, as the cluster grows, the number of contributing VB configurations increases and the total mixing term starts to dominate, and, since the elementary mixing term for Cu is significantly larger than the corresponding one for Au, the D_e and D_e/n values for the $^{n+1}Cu_n$ clusters become larger than for those of the $^{n+1}Au_n$ clusters.

B. NPFM Bonding of Resonating Bound Triplet Pairs. The above discussion shows that the VB modeling inherent in eq 1 captures qualitatively and semiquantitatively the essence of the NPFM bonding in the no-pair clusters of the coinage metals. NPFM bonding arises primarily from bound triplet electron pairs that spread over all the close neighbors of a given atom in the clusters.

The bound triplet pair owes its stabilization to the resonance energy provided by the mixing of the local ionic configurations, $^3M(\uparrow\uparrow)^-M^+$ and $M^+\,^3M(\uparrow\uparrow)^-$, and the various excited covalent configurations (involving p_z and d_{z^2} AOs) into the fundamental covalent structure $^3(M\uparrow\uparrow M)$ with the s^1s^1 electronic configuration. As was demonstrated for $^{n+1}Li_n$ clusters,[5] the mixing of the excited covalent structures into $^3(M\uparrow\uparrow M)$ generates a covalent structure with hybrid orbitals that keep the triplet electrons further apart compared with the fundamental s^1s^1 structure and thereby lowers the triplet repulsion. This is augmented by the mixing of the ionic structures, which buttress the bonding by covalent−ionic resonance energy.[43] Thus, if we consider each diatomic triplet pair and its ionic plus covalent fluctuations as a local NPFM bond, we can regard the electronic structure of a given $^{n+1}M_n$ cluster as *a resonance hybrid of all the local NPFM bonds that each atom forms with all of it close neighbors*.

In the case of alkali metals, the local FM bond involves only two electrons in s and p orbitals, while in the coinage metal clusters, there are also filled 3d orbitals that contribute components of three electron bonding due to the participation of these orbitals in the ionic fluctuations (Scheme 2). Thus, the no-pair coinage metal clusters are more strongly bonded than the corresponding alkali metal clusters.[4−6] Moreover, both $^{n+1}Cu_n$ and $^{n+1}Au_n$ possess stronger binding energies than the corresponding $^{n+1}Li_n$ clusters.

Concluding Remarks: Bonded Triplet Pairs

The paper discusses no-pair ferromagnetic (NPFM) bonding in the maximum-spin states of coinage metal clusters as a result of *bonded triplet pairs*. It is shown that the bonding energy per atom, D_e/n, grows rapidly with the cluster size, exhibits a strongly nonadditive behavior, and converges to values as large as 16−19 kcal/mol for gold and copper; values which are of the order of normal spin-paired bonds in metals.

The valence bond analysis of the problem shows that a weak stabilization of the triplet pair in the dimer can become a remarkably strong force that binds together monovalent atoms without a single electron pair. This is achieved because the steeply growing number of VB structures exerts on the triplet pair a cumulative effect of stabilization that is maximized when the cluster is compact with an optimal coordination number of the atoms. Thus, *the nonadditive behavior of the binding energy is scaled by the number of VB structures available for mixing with the fundamental repulsive structure*, $^{n+1}\Phi_{s(1), ..., s(i), ..., s(n)}$.

A more complete mini-periodic table of NPFM bonding will have to include the heavy alkali metals (K, Cs, Fr) and the group III metalloids, like B, Al, and so on. In view of the importance of the ionic structures, the heteroatomic clusters may be even more strongly bonded. Some future work thus lies ahead.

Acknowledgment. This research is supported by an ISF (Israel Science Foundation) grant (53/09) to S.S. Mark Danovich is acknowledged for help with the least-squares fitting described in the paper.

Supporting Information Available: Tables of Cartesian coordinates, figures with optimized structures, point groups and state assignments of all calculated clusters are available for B3P86/St-RECP and B3LYP/aug-cc-pVTZ-PP levels. Plots of the D_e/n vs cluster size (n) for all coinage metal clusters calculated using UB3P86/St-RECP method and eq 1 are also available. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) McAdon, M. H.; Goddard, W. A., III. *J. Phys. Chem.* **1988**, *92*, 1352.

(2) Glukhovtsev, M. N.; Schleyer, P. v. R. *Isr. J. Chem* **1993**, *33*, 455.

(3) Danovich, D.; Wu, W.; Shaik, S. *J. Am. Chem. Soc.* **1999**, *121*, 3165–3174.

(4) de Visser, S. P.; Alpert, Y.; Danovich, D.; Shaik, S. *J. Phys. Chem. A* **2000**, *104*, 11223.

(5) de Visser, S. P.; Danovich, D.; Wu, W.; Shaik, S. *J. Phys. Chem. A* **2002**, *106*, 4961.

(6) de Visser, S. P.; Danovich, D.; Shaik, S. *Phys. Chem. Chem. Phys.* **2003**, *5*, 158.

(7) Higgins, J.; Hollebeck, T.; Reho, J.; Ho, T.-S.; Lehmann, K. K.; Rabitz, H.; Scoles., G.; Gutowski, M. *J. Chem. Phys.* **2000**, *112*, 5751.

(8) Brühl, F. R.; Miron, R. A.; Ernst, W. E. *J. Chem. Phys.* **2001**, *115*, 10275.

(9) Fioretti, A.; Comparat, D.; Crubellier, A.; Dulieu, O.; Masnou-Seeuws, F.; Pillet, P. *Phys. Rev. Lett.* **1998**, *80*, 4402.

(10) For a few of the earlier calculations of these bound dimmers, see: (a) Kutzelnigg, W.; Staemler, V.; Gélus, M. *Chem. Phys. Lett.* **1972**, *13*, 496. (b) Olson, M. L.; Konowalow, D. D. *Chem. Phys.* **1977**, *21*, 393. (c) Konowalow, D. D.; Olson, M. L. *Chem. Phys.* **1984**, *84*, 462.

(11) It should be noted that the mixing of these structures is not equivalent to a simple 2s−2p$_z$ hybridization effect, as evidenced by the fact that Hartree−Fock (HF) wave function for $^3Li_2$ ($^3\Sigma_u^+$), is unbound and repulsive throughout the internuclear distance of 3Å, despite the significant hybridization of the 2s and 2p$_z$ orbitals, more so than in the post HF wave functions.

(12) Higgins, J.; Callegari, C.; Reho, J.; Stienkemeier, F.; Ernst, W. E.; Lehmann, K. K.; Gutowski, M.; Scoles, G. *Science* **1996**, *273*, 629.

(13) Higgins, J.; Ernst, W. E.; Callegari, C.; Reho, J.; Lehmann, K. K.; Scoles, G. *Phys. Rev. Lett.* **1996**, *77*, 4532.

(14) Higgins, J.; Callegari, C.; Reho, J.; Stienkemeier, F.; Ernst, W. E.; Gutowski, M.; Scoles, G. *J. Phys. Chem. A* **1998**, *102*, 4952.

(15) Reho, J.; Higgins, J.; Nooijen, M.; Lehmann, K. K.; Scoles, G.; Gutowski, M. *J. Chem. Phys.* **2001**, *115*, 10265.

(16) (a) Cvitas, M. T.; Soldan, P.; Houston, J. M. *Phys. Rev. Lett.* **2005**, *94*, 033201. (b) Quemener, G.; Honvault, P.; Launay, J.-M.; Soldan, P.; Potter, D. E.; Houston, J. M. *Phys. Rev. A: At., Mol., Opt. Phys.* **2005**, *71*, 032722.

(17) (a) Bondybey, V. E. *J. Chem. Phys.* **1982**, *77*, 3771. (b) Huber, K. P.; Herzberg, G. *Constants of Diatomic Molecules,*; Van Nostrand Reinhold: New York, 1979. (c) Rohlfing, E. A.; Valentini, J. J. *J. Chem. Phys.* **1986**, *84*, 6560.

(18) (a) Cvitas, M. T.; Soldan, P.; Houston, J. M.; Honvault, P.; Launay, J.-M. *Phys. Rev. Lett.* **2005**, *94*, 033201. (b) Cvitas, M. T.; Soldan, P.; Houston, J. M.; Honvault, P.; Launay, J.-M. *Phys. Rev. Lett.* **2005**, *94*, 200402.

(19) McAdon, M. H.; Goddard, W. A., III. *J. Chem. Phys.* **1988**, *88*, 277.

(20) Morse, M. D. *Chem. Rev.* **1986**, *86*, 1049.

(21) Lombardi, J. R.; David, B. *Chem. Rev.* **2002**, *102*, 2431.

(22) de Visser, S. P.; Kumar, D.; Danovich, M.; Nevo, N.; Danovich, D.; Sharma, P. K.; Wu, W.; Shaik, S. *J. Phys. Chem. A* **2006**, *110*, 8510.

(23) See for example, (a) Pyykkö, P. *Angew. Chem., Int. Ed.* **2002**, *41*, 3573. (b) Pyykkö, P. *Chem. Rev.* **1997**, *97*, 597.

(24) See for example: (a) Schwerdtfeger, P. *Angew. Chem., Int. Ed.* **2003**, *42*, 1892. (b) Hakkinen, H. *Chem. Soc. Rev.* **2008**, *37*, 1847. (c) Bao, K.; Goedecker, S.; Koga, K.; Lancon, F.; Neelov, A. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2009**, *79*, 041405(R)

(25) (a) Carmeli, I.; Leitus, G.; Naaman, R.; Reich, S.; Vager, Z. *J. Chem. Phys.* **2003**, *118*, 10372. (b) L'vov, V. S.; Naaman, R.; Tiberkevich, V.; Vager, Z. *J. Chem. Phys. Lett.* **2003**, *381*, 650. (c) Naaman, R.; Vager, Z. *Phys. Chem. Chem. Phys.* **2006**, *8*, 2217.

(26) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(27) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schutz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D.;

Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. *MOLPRO*, version 2006 1; University College Cardiff Consultants Limited: Wales, U.K., 2006; see http://www.molpro.net.

(28) (a) Song, L.; Mo, Y.; Zhang, Q.; Wu, W. *XMVB: An ab initio Non-orthogonal Valence Bond Program*; Xiamen University: Xiamen, China, 2003. (b) Song, L.; Mo, Y.; Zhang, Q.; Wu, W. *J. Comput. Chem.* **2005**, *26*, 514.

(29) Danovich, D.; Filatov, M. *J. Phys. Chem. A* **2008**, *112*, 12995.

(30) Reiher, M. *Theor. Chem. Acc.* **2006**, *116*, 241.

(31) Dyall, K. G. *J. Chem. Phys.* **1997**, *106*, 9618.

(32) Balabanov, N. B.; Peterson, K. A. *J. Chem. Phys.* **2005**, *123*, 064107.

(33) Peterson, K. A.; Puzzarini, C. *Theor. Chem. Acc.* **2005**, *114*, 283.

(34) Dolg, M.; Stoll, H.; Preuss, H.; Pitzer, R. M. *J. Phys. Chem.* **1993**, *97*, 5852.

(35) Pou-Amerigo, R.; Merchan, M.; Nebot-Gil, I.; Widmark, P. O.; Roos, B. *Theor. Chim. Acta* **1995**, *92*, 149.

(36) (a) For Zora calculations of Cu_2, Ag_2 and Au_2, see: van Wüllen, C. *J. Phys. Chem.* **1998**, *109*, 392. (b) For Douglas-Kroll calculations of Au_2 using MP4 and CCSD(T) levels: Hess, B. A.; Kaldor, U. *J. Chem. Phys.* **2000**, *112*, 1809.

(37) Bylaska, E. J.; de Jong, W. A.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; Wang, D.; Apra, 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.; Wu, Q.; Van Voorhis, T.; Auer, A. A.; Nooijen, M.; Brown, E.; Cisneros, G.; Fann, G. I.; Fruchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J. A.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Pollack, L.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers*, version 5.1; Pacific Northwest National Laboratory: Richland, Washington, 2007.

(38) For other computational studies, see: (a) DFT, MP2 and CCSD(T) calculations of the ground states of Au_6 and Au_8clusters: Olson, R. M.; Varganov, S.; Gordon, M. S.; Metiu, H.; Chretien, S.; Piecuch, P.; Kowalski, K.; Kucharski, S. A.; Musial, M. *J. Am. Chem. Soc.* **2005**, *127*, 1049. (b) Structural and electronic properties of silver clusters up to Ag_{21}: Zhao, J.; Luo, Y.; Wang, G. *Eur. Phys. J. D* **2001**, *14*, 309. (c) B3LYP and PW91PW91/LANL2DZ calculations of Cu_n, Ag_n, Au_n (n =2−6): Zhao, S.; Ren, Y.; Wang, J.; Yin, W. *J. Phys. Chem. A* **2009**, *113*, 1075. (d) DFT calculations of silver clusters up to n=12, using scalar relativistic model potential: Fournier, R. *J. Chem. Phys.* **2001**, *115*, 2165. (e) PW91 calculations of properties of silver clusters: Pereiro, M.; Baldomir, D. *Phys. Rev. A: At., Mol., Opt. Phys.* **2007**, *75*, 033202. (f) MP2 calculations of gold clusters up to n=6 in the ground state sho planar geometries: Bravo-Perez, G.; Garzon, I. L.; Novaro, O. *J. Mol. Struct (Theochem)* **1999**, *493*, 225. (g) DFT with plane wave basis sets of Au clusters find that the transition to 3D starts at An_{15}: Xiao, L.; Wang, L. *Chem. Phys. Lett.* **2004**, *392*, 425.

(39) Pereiro, M.; Baldomir, D.; Arias, J. E. *Phys. Rev. A: At., Mol., Opt. Phys.* **2007**, *75*, 063204.

(40) Full configuration-interaction study of the tetrahedral Li_4 cluster confirmed the nature of the no-pair bond by an independent method: Monari, A.; Pitarch-Ruiz, J.; Bendazzoli, G. L.; Evangelisti, S.; Sanches-Martin, J. *J. Chem. Theory and Computations* **2008**, *4*, 404.

(41) Glendening, E. D.; Badenhoop, J. K.; Reed, A. E.; Carpenter, J. E.; Bohmann, C. M.; Morales, C. M.; Weinhold, F. *NBO 5.0*, Theoretical Chemistry Institute: University of Wisconsin, Madison, WI, 2001.

(42) Pyykkö, P.; Desclaux, J.-P. *Acc. Chem. Res.* **1979**, *12*, 276.

(43) Shaik, S.; Hiberty, P. C. *A Chemist's Guide to Valence Bond Theory*, Wiley-Interscience: New York, 2007.

CT100088U

JCTC Journal of Chemical Theory and Computation

Dual Grid Methods for Finding the Reaction Path on Reduced Potential Energy Surfaces

Steven K. Burger and Paul W. Ayers*

Department of Chemistry & Chemical Biology, McMaster University,
1280 Main St. West, Hamilton, Ontario, Canada

Received January 7, 2010

Abstract: Two new algorithms are presented for determining the minimum energy reaction path (MEP) on the reduced potential energy surface (RPES) starting with only the reactant. These approaches are based on concepts from the fast marching method (FMM), which expands points outward as a wavefront on a multidimensional grid from the reactant until the product is reached. The MEP is then traced backward to the reactant. Since the number of possible grid points that must be considered grows exponentially with increasing dimensionality of the RPES, interpolation is important for maintaining manageable computational costs. In this work, we use Shepard interpolation, which we have modified to resolve problems in overfitting. In contrast to FMM, which accurately locates the MEP, the new algorithms focus on locating the single rate-limiting transition state and provide only a rough estimate of the MEP. They do this by mapping out the RPES on a coarse grid and then refining a least action path on a finer grid. This is done so that the majority of the interpolation is done on the finer grid, which minimizes the amount of extrapolation inherent in an outward searching algorithm. The first method scans the entire PES before iteratively locating the transition state (TS) for the MEP on the lower bound estimate of the fine PES. The second method explores the coarse grid in a similar manner to FMM and then iteratively locates the rate-limiting TS in the same manner as the first method. Both methods are shown to be capable of rapidly obtaining (in less than 30 constrained optimization cycles) an approximation to the MEP and the rate limiting TS for three example systems: the 4-well potential, the molecule *N*-hydroxymethyl-methylnitrosaminee (HMMN), and a cluster model of DNA-uracil glycosylase.

1. Introduction

For many chemical problems, we are interested in the kinetics of a reaction, which requires knowing the mechanism and the energy barrier. Common kinetically interesting examples are gas and solution phase molecular reactions, enzyme mechanisms,[1] and conformation changes of proteins.[2] When dealing with such systems, we would like to know all of the kinetically accessible minima and transition state (TS) structures. With this information, we can determine the reaction rate with a variety of methods such as transition state theory[3] or the reaction path Hamilton method.[3,4]

Also of interest is the minimum energy path (MEP) between the reactant and product. This is defined as the steepest descent path from the TS to each minimum and it represents the most probable path the system would take at 0 K. It can be obtained relatively easily if the TS structures are known, since it reduces the problem to an initial value problem,[5] solvable with an implicit Runge−Kutta method. However, the MEP is generally less interesting for computational chemists than the TS structures since any thermodynamic path[6] connecting the rate limiting TS to the end points will give the correct kinetics.

Finding TS structures is an optimization problem and the methods used are similar to the methods used for minimization. However, the problem is complicated by the fact that one eigenvector of the Hessian must have a negative eigenvalue. This greatly increases the difficulty of the problem,

* Corresponding author e-mail: ayers@mcmaster.ca.

10.1021/ct100012y © 2010 American Chemical Society
Published on Web 03/30/2010

since it implies that a good guess of the Hessian is available and computing the Hessian analytically is usually computational expensive. If a good initial guess is available though, then there are a number of algorithms[7−14] that can get the TS starting from a point relatively close to the solution.

When the location of the TS is difficult to approximate then, instead of using a single point to find the TS, a string of points can be used to approximate the MEP. The highest point on this path is the best guess of the TS. There are a large number of string methods available,[15−25] all of which take advantage of the fact that the MEP must be a minimum in all directions perpendicular to the path. This class of methods tends to be easy to parallelize and will give a good approximation to the entire path, from which approximate TS structures can be obtained and then refined with TS searching methods.[7,8,25−27] Two major shortcomings of these methods are that many points are needed to accurately represent the path and lower barrier paths may be missed if one starts with a poor initial guess of the path.

The difficulty in finding the MEP and TS is related to the dimensionality of the problem. Fortunately, for many chemical problems a few key coordinates[28] can be identified, such as the interatomic distance involved in bond breaking or forming. The energy can then be formulated in terms of a reduced set of coordinates by minimizing all other degrees of freedom, thus removing their contribution. An extreme case of this is the coordinate driving method, where the mechanism is determined by changing only one coordinate while minimizing all of the other degrees of freedom. However, one dimension is usually not enough to describe a reaction. As a result, discontinuities will often emerge, due to the fact that other important degrees of freedom are sensitive to very small changes in the driven coordinate.

If all of the important degrees of freedom are included in the reduced set of coordinates, then the reduced potential energy surface (RPES) will be smooth and interpolation methods can be used. Shepard interpolation[29,30] has been shown to work well for the fast-marching and string methods[31] and it is our choice for the methods outlined here. Shepard interpolation only requires values for the potential; however, the quality of the interpolation can be dramatically improved by using derivative terms as well. For chemical problems, the gradient can be computed at a similar cost to the potential, so it is usually included. Higher derivatives are usually too costly to compute directly, but they can be approximated by interpolated moving least-squares (IMLS).[32,33]

Given a good error estimate for the interpolated RPES, either a string method or the fast-marching method (FMM)[34−37] can be used to explore the surface.[31] Unlike string methods, FMM has the nice property that it exhaustively searches the RPES to find the MEP and does not require knowledge of the product state. FMM propagates a wavefront and effectively fills up the surface on a grid. Once the product is found, the MEP is obtained by integrating the steepest descent path of an action surface back from the product.

FMM is greatly improved with interpolation, but its main shortcoming is that each new point is obtained by extrapolation rather than interpolation, since the algorithm expands outward from the current set of determined points. To get around this issue, two new algorithms are developed which evaluate the points on two grids: (1) a coarse grid that allows a rough estimate of the MEP to be obtained, (2) a finer grid that is used to determine the rate limiting TS structure. While extrapolation is the only alternative to directly evaluating the potential on the coarser grid, on the finer grid, new points can be approximated with Shepard interpolation. With interpolation we show that these methods work well for three different systems of increasing complexity.

2. Theory and Computational Methods

Least Action Surface. The fast marching method (FMM),[34−37] like all reaction path methods, is based on finding the path of length q which minimizes the integral,

$$S(q) = \int_0^q f(q(\tau))d\tau \tag{1}$$

where τ is a parametrization of the arc length and $f(\tau)$ is a cost function which can be defined as follows:

$$f(q) = \left[\frac{E - V(q)}{E - V(q_{\min})}\right]^{l/2} \tag{2}$$

E is the classically highest allowed energy of the system, $V(q_{\min})$ is the lowest potential value and l determines the cost being considered. If $l = 0$ then solving eq 1 will minimize the distance between two points, while if $l = 1$ we obtain the least time path. In the limit $l \to \infty$, minimizing eq 1 results in the MEP.[35] If we differentiate eq 1 we obtain the Hamilton−Jacobi differential equation,

$$|\nabla S(q)| = f(q) \tag{3}$$

Equation 3 can be solved practically as a finite difference equation. For the 2D case, this has the form,

$$\max\left(\frac{S_{i,j} - S_{i-1,j}}{\Delta q_1}, \frac{S_{i,j} - S_{i+1,j}}{\Delta q_1}, 0\right)^2 + \max\left(\frac{S_{i,j} - S_{i,j-1}}{\Delta q_2}, \frac{S_{i,j} - S_{i,j+1}}{\Delta q_2}, 0\right)^2 = (f(q))^2 \tag{4}$$

The full details of solving eq 4 are given in ref 35. Once the action has been computed at all points $S_{i,j}$ on the 2D grid, then the MEP can be determined by integrating backward (backtracing) from the product on the least action surface. This can be done only at grid points or on an interpolated surface. For the methods in this work, we do backtracing on the grid points.

Shepard Interpolation. To reduce the number of energy and gradient evaluations, interpolation is used when the error in the interpolant is sufficiently small. How large an error is tolerable is set as a user-defined parameter. Shepard interpolation[29,30,38−42] has been used with FMM and string methods[31] and we use the same basic scheme here as well. Shepard interpolation uses a set of points $\mathbf{X}^{(i)}$ where the Taylor expansion at each point is,

$$T_n^{(i)}(\mathbf{X}) = V(\mathbf{X}^{(i)}) + (\mathbf{X} - \mathbf{X}^{(i)})\cdot\nabla V(\mathbf{X}^{(i)}) + \frac{1}{2}(\mathbf{X} - \mathbf{X}^{(i)})\cdot\nabla\nabla V(\mathbf{X}^{(i)})\cdot(\mathbf{X} - \mathbf{X}^{(i)}) + ... \quad (5)$$

such that n is the highest order of the expansion and $\mathbf{X}$ is the coordinate of the point of interest. For the methods in this work, we calculate the first two terms in the Taylor series ($V(\mathbf{X}^{(i)})$ and $\nabla V(\mathbf{X}^{(i)})$) and use weighted least-squares to determine the $p = (1)/(n!)\prod_{i=0}^{n-1}(d + i)$ components of the higher-order ($n > 1$) derivatives. Rewriting eq 5 as a weighted least-squares equation we get,

$$\min_{\mathbf{x}} ||\mathbf{W}(\mathbf{A}\mathbf{x} - \mathbf{b})|| \quad (6)$$

where $\mathbf{x} = \{\nabla^2 V(\mathbf{X}^{(i)}),\nabla^3 V(\mathbf{X}^{(i)}),...\}$ is a vector of length p containing the unknown higher order terms we are interested in; $\mathbf{b}$ is a linear combination of $V(\mathbf{X}^{(i)})$ and $\nabla V(\mathbf{X}^{(i)})$ of length $M(d + 1)$, where M is the number of neighboring points used and d is the dimensionality of the system; $\mathbf{A}$ is a matrix of the $(\mathbf{X} - \mathbf{X}^{(i)})$ terms in eq 5; and $\mathbf{W}$ is a diagonal weighting matrix. For the weight matrix, if we use the Bettens−Collins isotropic formula[39] or other similar forms,[31] then there can be problems with overfitting when too few points make significant contributions to the Shepard interpolant. Specifically, problems arise unless M is appreciably larger than $p/d - 1$. To get around this, we can change the weighting function to ensure that the weighting is more evenly distributed among the points by using the usual form for the diagonal terms,

$$w^{(i)}(\mathbf{X}) = \frac{v^{(i)}(\mathbf{X})}{\sum_{j=1}^{M} v^{(j)}(\mathbf{X})} \quad (7)$$

with,

$$v^{(i)}(\mathbf{X}) = e^{-1/2\left(\frac{||X - X^{(i)}||}{\sigma^{(i)}}\right)^2} \quad (8)$$

But instead of basing the trust radius $\sigma^{(i)}$ on the grid spacing, we sort the neighbors of point $\mathbf{X}^{(i)}$ based on distance and then chose the $k = p/d - 1$ element so that $\sigma^{(i)} = ||\mathbf{X} - \mathbf{X}^{(k)}||$. This ensures that enough points are within one standard deviation of the weighting function so that overfitting does not occur.

Determining the error in the coefficients requires the residual of the least-squares fit,

$$\sigma^2 = \frac{\mathbf{b}^T\mathbf{b} - (\mathbf{A}\mathbf{x})^T\mathbf{b}}{(d + 1)M - p} \quad (9)$$

from which the covariance matrix can be determined,

$$\mathbf{V} = (\mathbf{A}^T\mathbf{A})^{-1}\sigma^2 \quad (10)$$

The error from the higher order fitted terms in the Taylor series is estimated as follows:

$$\varepsilon_T(\mathbf{X}) = \sqrt{\sum_{i=1}^{M(d+1)}\sum_{j=1}^{p} V_{ij}A(\mathbf{X})_{ij}^2} \quad (11)$$

Of course the residual in the Taylor series is still not accounted for, but usually this term is significantly smaller than the error introduced by fitting the higher order terms.

Equation 5 can be fit to any order so long as $p \leq M(d + 1)$. However, overfitting will be a problem if $p/M(d + 1) \approx 1$. To determine if the next order is a good model for the surface, we can check to see if there is a “lack of fit”.[43] This is generally done by testing the general linear hypothesis, $\mathbf{b}_{p-q...p} = \mathbf{0}$. To test this hypothesis, the ratio,

$$\frac{\left(\frac{\sigma_{n-1}^2 - \sigma_n^2}{q}\right)}{\left(\frac{\sigma_n^2}{n - p}\right)} \quad (12)$$

is compared against the $F(q, n - p)$ distribution.[44] In eq 12 σ_n^2 is the standard deviation when derivatives are fit up to the order n. Unfortunately, we found that this method does not work well for the problems we considered. Instead, we used a more practical method, leave one out cross-validation (LOOCV). In this scheme, each of the M neighbor points used in the fitting is left out in sequence while the remaining M-1 points are used to fit $\mathbf{x} = \{\nabla^2 V(\mathbf{X}^{(i)}),\nabla^3 V(\mathbf{X}^{(i)}),...\}$. Each point left out is used to estimate one term of a weighted mean squared error.

$$\mathrm{MSE}(n) = \sum_{k=1}^{M} w^{(i)}(\mathbf{X}^{(k)})(V(\mathbf{X}^{(k)}) - T_n^{(i)}(\mathbf{X}^{(k)}))^2 \quad (13)$$

This is done for each order between 2 and 5. The derivatives are fit up to and including the order which gave the lowest value for eq 13.

Once the higher order terms have been fit, the potential at any point $\mathbf{X}$ on the surface is obtained by the sum,

$$\tilde{V}(\mathbf{X}) = \sum_{i=1}^{M} w^{(i)}(\mathbf{X})T^{(i)}(\mathbf{X}) \quad (14)$$

where $T^{(i)}(\mathbf{X})$ are given by eq 5, M is the number of neighbor points, and $w^{(i)}(\mathbf{X})$ is the weight function. The error for this sum can be estimated as follows:

$$\varepsilon_{\tilde{V}}(\mathbf{X}) = \sqrt{\sum_{i=1}^{M} w^{(i)}(\mathbf{X})(\tilde{V}(\mathbf{X}) - T_n^{(i)}(\mathbf{X}))^2} \quad (15)$$

where the same weight function from ref 31 is used in both eqs 14 and 15. This works well when one point does not dominate the sum, which we define as $w^{(j)}(\mathbf{X}) > 0.9$. If one term does dominate, then eq 15 will likely underestimate the error, and eq 11 will generally be a better estimate.

Dual Grid−Low Path Methods. Both algorithms are based on a dual grid approach. The methods differ mainly in their treatment of the coarse grid. In the low path method, one evaluates the energy and gradient at every point on the coarse grid. The boundary low-path method, by contrast, evaluates points on the coarse grid in a similar way to FMM, avoiding points that lie outside the boundary where the error in extrapolation is too large.

We denote the upper and lower bounds on the d coordinates under consideration as **ub** and **lb**. The reactant and product configurations are $\mathbf{R}_{\text{react}}$ and $\mathbf{R}_{\text{prod}}$, respectively. The coarse grid consists of the set of points, $X_i^{(k)} = R_{\text{react},i} + c_i \Delta R_{\text{large},i}$, for which $lb_i < X_i^{(k)} < ub_i$. Here $1 \le i \le d$ denotes the particular coordinate of interest, c_i is an integer, and $\Delta\mathbf{R}_{\text{large}}$ is the vector of grid spacings for the coarse grid. Similarly, points on the fine grid are defined by $X_i^{(k)} = R_{react,i} + c_i \Delta R_{small,i}$, where $\Delta\mathbf{R}_{\text{small}}$ is the vector of grid spacings for the small grid. A parameter, α, is used to construct a lower error bound on the true RPES. In keeping with our previous notation, values of the potential energy evaluated using computational chemistry software are given as $V(\mathbf{X}^{(k)})$ and interpolated values of the potential are $\tilde{V}(\mathbf{X}^{(k)})$.

At first, the algorithm constructs an interpolation of the full RPES on the coarse grid and then locates the best guess at the TS. Next we set $\alpha = 1$, so that the error is subtracted from the value of the interpolant; this provides an (approximate) lower bound to the true potential energy. The TS is then located on the lower-bound surface. If the TS is located at a grid point which has already been evaluated (i.e., where the error is zero), then we identify this conformation as the rate limiting TS. Otherwise, the energy and gradient are evaluated at this point, the surface is reinterpolated and the process repeats.

The transition-state estimate will be accurate, up to grid-spacing of the fine grid, as long as: (a) the coarse grid is fine enough that alternative pathways are not missed and (b) the error estimate of the interpolated potential is not underestimated. The full algorithm is as follows:

Algorithm 1: Low Path Method (LPM).

(a) Set $\Delta\mathbf{R}_{\text{large}}$, $\Delta\mathbf{R}_{\text{small}}$, **lb**, **ub**, $\mathbf{R}_{\text{react}}$, $\mathbf{R}_{\text{prod}}$, E_{barr}, and $\alpha = 0$.

(b) Coarsely scan the RPES evaluating $V(\mathbf{X}^{(k)})$ and $\nabla V(\mathbf{X}^{(k)})$ at the points $X_i^{(k)} = R_{\text{react},i} + c_i \Delta R_{\text{large},i}$ which satisfy $lb_i < X_i^{(k)} < ub_i$ for $i = 1...d$ and $c_i \in \mathbb{Z}$.

(c) Fit the higher-order derivatives of the potential to the evaluated points using eq 6.

(d) Interpolate all unevaluated points $X_i = R_{\text{react},i} + c_i \Delta R_{\text{small},i}$, $lb_i < X_i < ub_i$ on the fine grid to get $\{V(\mathbf{X}),\varepsilon(\mathbf{X})\}$ where $\varepsilon(\mathbf{X})$ is the error given by either eq 11 or eq 15. For evaluated points set $\varepsilon(\mathbf{X}) = 0$.

(e) Determine the action S at each point on the fine grid by solving eq 4, with $f(\mathbf{X}) = [(E - V(\mathbf{X}) + \alpha\varepsilon(\mathbf{X}))/(E - V_{\min}(\mathbf{X}))]^{1/2}$, where E is an upper-bound estimate of the potential at the highest TS.

(f) Backtrace from $\mathbf{R}_{\text{prod}}$ on the action surface to get the MEP.

(g) If the highest point X_{TS} on the MEP is evaluated, then

i. if α=1 THEN STOP; ELSE set α=1.

(h) Evaluate $\{V(X_{\text{TS}}),\nabla V(X_{\text{TS}})\}$. Set $\varepsilon(X_{\text{TS}}) = 0$. GOTO (c).

The algorithm first iterates until the rate limiting TS is found on the interpolated surface without consideration of the error. Then with $\alpha = 1$, a lower bounded RPES is used to find the TS within the accuracy of the grid. This can be skipped if the coarse grid size is sufficiently small, but otherwise alternate paths with lower energy TS structures may be missed.

The low-path method (LPM) works well when we are interested in the full RPES, $\mathbf{lb} < \mathbf{R} < \mathbf{ub}$. Often, however, there are large regions of conformation space where the potential energy is too high to be of interest. It would be more efficient not to explore those regions. FMM is particularly good at only exploring the low-energy regions, so we propose a second algorithm, called the boundary low-path method (BLPM) that (a) uses FMM to explore the RPES until the product state is located and then (b) uses the same methodology as LPM to find the TS on the fine grid.

The key new idea in the BLPM is the construction of a boundary set, B, on the fine grid that separates the region of the potential energy surface where the interpolant is sufficiently accurate from the region where the interpolation cannot be trusted. A point on the fine grid, $\mathbf{X}^{(i)}$, is a boundary point if it satisfies our error criterion ($\varepsilon(\mathbf{X}^{(i)}) < \varepsilon_{\max}$), but one of its neighbors on the fine grid does not.

Algorithm 2: Boundary Low Path Method (BLPM).

(a) Set $\Delta\mathbf{R}_{\text{large}}$, $\Delta\mathbf{R}_{\text{small}}$, **lb**, **ub**, $\mathbf{R}_{\text{react}}$, $\mathbf{R}_{\text{prod}}$, E_{barr}, $\varepsilon_{\max}$, and set $\alpha = 0$.

(b) Follow steps (c)−(e) in Algorithm 1 to interpolate the RPES and to get the action values.

(c) Construct the boundary set:

$$B = \{\mathbf{X}^i | \varepsilon(\mathbf{X}^i) < \varepsilon_{\max}, \exists \mathbf{X}^k : |X_j^i - X_j^k| \le \Delta R_{\text{small},j}, \quad j = 1...d, \varepsilon(\mathbf{X}^k) > \varepsilon_{\max}\}$$

(d) Take the element of B which has the lowest action $b_{\min} = \arg\min_b\{S(b), b \in B\}$.

(e) If $S(b_{\min}) > S(\mathbf{R}_{\text{prod}})$, then set α=1 and GOTO to Algorithm 1, starting at step (f).

(f) Evaluate the nearest neighbors of $b_{\min}$ on the course grid to obtain the set, $\{(V(\mathbf{X}^k),\nabla V(\mathbf{X}^k)) | |b_{\min,j} - X_j^k| \le \Delta R_{\text{large},j}, j = 1...d\}$. If all of the surrounding points on the larger grid are evaluated then only evaluate $\{V(b_{\min}),\nabla V(b_{\min})\}$.

(g) GOTO (b).

This method has fewer initial evaluations but has drawbacks. Since BLPM evaluates points outward from the reactant on the RPES in a similar fashion to FMM, the error in the approximated energies tends to be larger during this first step because more extrapolation is used. Also it may be more difficult to parallelize than LPM since the grid points that need to be evaluated are less predictable.

3. Results and Discussion

To compare BLPM and LPM, we examined the 4-well potential, a gas phase molecular dissociation and a cluster model of an enzyme. All systems were done in two-dimensions so they could be visualized. The code is broken up into a number of Fortran 90[45] programs, which communicate with external files. For the energy and gradient calculations, we used system calls to Gaussian03[46] and for the cost function we used $l = 15$.

The 4-Well Potential. Analytic systems can be good at demonstrating flaws in certain methods. The 4-well potential provides an example for how string methods can fail when starting with a linear interpolation as an initial guess of the path. The highest TS on the MEP is located at (−0.274223,

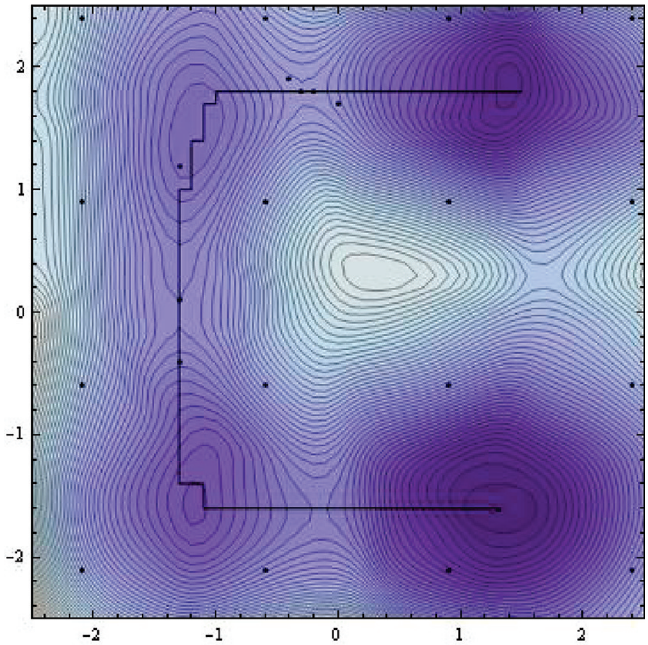


Figure 1. A contour plot of the four-well potential with Shepard interpolation using the fine grid spacing $\Delta\mathbf{R}_{\mathrm{small}} = (0.1,0.1)$. The evaluated points from LPM are shown as black dots. The backtrace path on the grid is the black curve. LPM converges after 24 evaluations to within the accuracy of the fine grid.

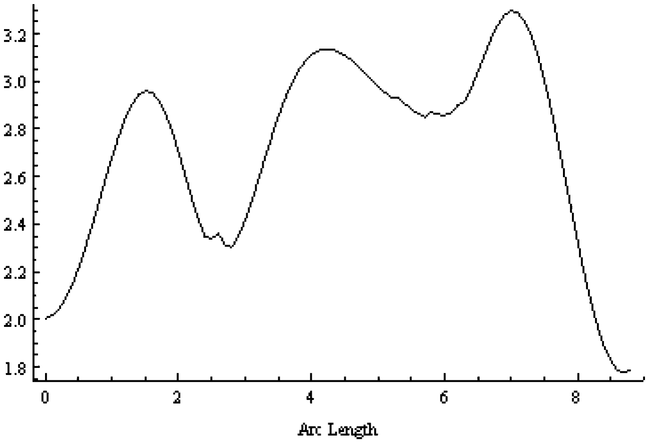


Figure 2. The energy plot of the MEP for the LPM. The intermediates are poorly resolved since they are furthest from the evaluated points.

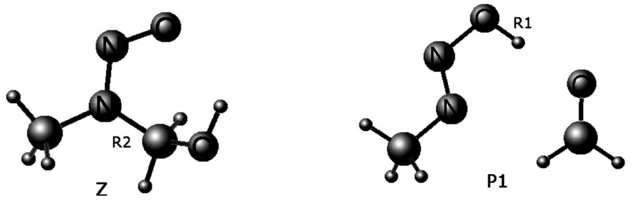


Figure 3. The g03 optimized end points for *N*-hydroxymethyl-methylnitrosamine (HMMN) using HF/3-21G. R1 is the hydroxyl O−H bond distance and R2 is the N−C distance. The Z and P1 labels correspond to the structures from ref 47.

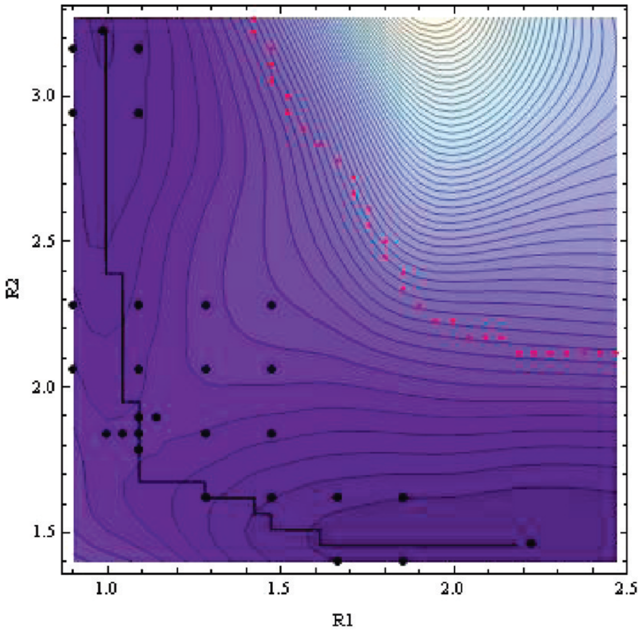


Figure 4. The HF/3-21G potential energy surface for HMMN demethylation after the BLPM has converged, where the distances are in Angstroms. The line of dots at the top of the plot is the boundary for an error tolerance of 0.01 au. The large black dots are points are optimized points on the surface and the curve is the approximate MEP. Most evaluations are clustered near the TS.

1.79308) and has a potential value of 3.2961. We set $\Delta\mathbf{R}_{\mathrm{large}} = (1.5,1.5)$, $\Delta\mathbf{R}_{\mathrm{small}} = (0.1,0.1)$, $\mathbf{lb} = (-2.5, -2.5)$, $\mathbf{ub} = (2.5,2.5)$, $E_{\mathrm{barr}} = 6$, and $\varepsilon_{\mathrm{shep}} = 0.2$ for the LPM. The coarse grid points are shifted by (0.4, 0.4) away from the lower bound. The converged results are shown in Figure 1 with the interpolated MEP shown in Figure 2. LPM converges on the grid to $\mathbf{R}_{\mathrm{TS}} = (-0.3,1.8)$ with $V(\mathbf{R}_{\mathrm{TS}}) = 3.2953$. The method requires 16 evaluations to calculate the potential on the initial grid, 1 evaluation for the starting point and 7 more evaluations on the finer grid for a total of 24. The points in the last step are largely focused on the regions with the highest barriers, with one point located near the second intermediate where the error is particularly large. Operating on the finer grid, it takes FMM 241 evaluations to resolve the TS to the same degree of accuracy, and it takes 75 with a grid spacing (0.5, 0.5).

***N*-Hydroxymethyl-Methylnitrosamine (HMMN).** This system, which is shown in Figure 3, is taken from ref 47. Rather than examine the entire reaction we simply looked at the demethylation step which results in the product methyldiazohydroxide. The compound is interesting since has been shown that it may methylate DNA bases in vivo.[48] Gaussian 03[46] was used to evaluate each point using the keywords OPT and MODRED in the heading with the bond variables R1 and R2, shown in Figure 3, kept frozen (http://www.chemistry.mcmaster.ca/ayers/projects.html). For this system, we use $\Delta\mathbf{R}_{\mathrm{large}} = (0.38,0.44)$, $\Delta\mathbf{R}_{\mathrm{small}} = (0.0475, 0.055)$, $\mathbf{lb} = (0.9,1.4)$, $\mathbf{ub} = (2.5,3.3)$, $\varepsilon_{\mathrm{shep}} = 0.01$ au, $\mathbf{R}_{\mathrm{react}} = (0.98,3.22)$ and $\mathbf{R}_{\mathrm{prod}} = (2.22,1.46)$, starting from the product “P1” rather than “Z”. The potential energy surface is shown in Figure 4. LPM requires 34 constrained geometry optimizations to converge to the TS while BLPM requires just 20.

When the coarser grid is made finer by setting $\Delta\mathbf{R}_{\mathrm{large}} = (0.19,0.22)$, then the LPM takes significantly longer using

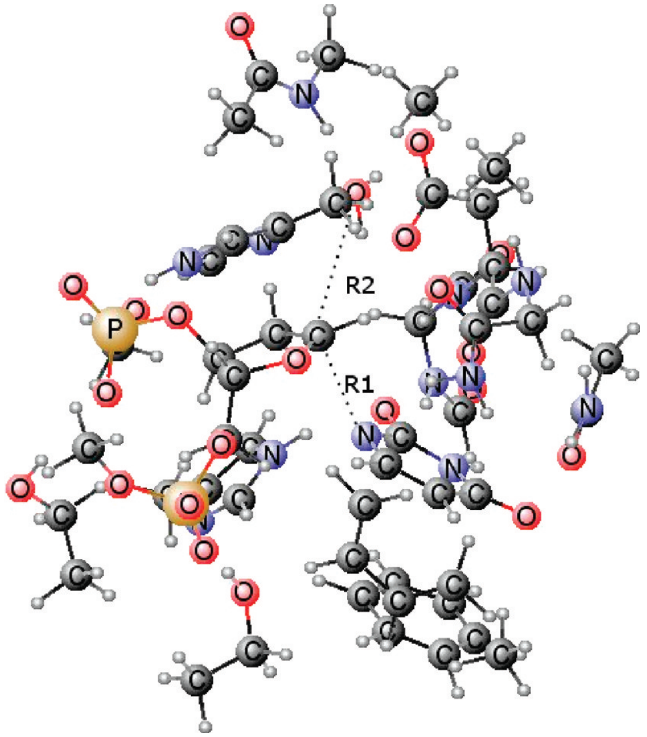


Figure 5. A Uracil-DNA glycosylase cluster model based on ref 50. The two coordinates used for the BLPM are shown as R1 and R2. This oxocarbenium cation/anion intermediate was optimized with AM1 keeping key unreactive atoms fixed.

85 constrained optimizations to converge compared to BLPM, which only takes 28. In both cases, the methods converge to within 0.2 kcal/mol of the true barrier.

Cluster Model of Uracil-DNA Glycosylase. To test this method on a larger system, we selected a cluster model of 177 atoms, shown in Figure 5, from Uracil-DNA Glycosylase based on the crystal structure 1EHM.[49] As in ref 50, a select number of atoms were frozen to keep the cluster together and AM1 was used to minimize the end point structures. For this system, we only tested BLPM with the parameters: $\Delta\mathbf{R}_{\text{large}} = (0.2,0.15)$, $\Delta\mathbf{R}_{\text{small}} = (0.05, 0.05)$, $\mathbf{lb} = (1.5,3.75)$, $\mathbf{ub} = (2.7,4.5)$, $\varepsilon_{\text{shep}} = 0.05$ au, $\mathbf{R}_{\text{react}} = (1.5,4.46)$, and $\mathbf{R}_{\text{prod}} = (2.52,3.84)$. The maximum allowed error was set to 0.01au. The exact AM1 TS, (2,12, 4.07), was located with Gaussian03 using the keyword opt(QST3) starting from the grid method's final structure at (2.1, 4.05). However, Gaussian03 had trouble converging for this structure, taking more than 500 steps before finishing. The interpolated RPES is shown in Figure 6 for BLPM, which converged in just 15 optimization cycles to (2.10, 4.05). Each constrained optimization cycle took about 30 iterations using the normal convergence criterion and 15 iterations using a loose criterion. The error was particularly large near the boundary at the top and bottom of the RPES, and the algorithm guessed at paths that would run along the boundaries. After the algorithm placed points near the boundary of the allowed region, it converged relatively quickly to the correct TS region.

4. Conclusions

Two new methods, the low-path method (LPM) and the boundary low-path method (BLPM), are proposed for finding the transition state (TS) structures on a reduced dimensional potential energy surface (RPES). Although it is more expensive to obtain points on the RPES than it is to evaluate points on the full-dimensional PES, the reduced dimensionality simplifies finding TS structures and allows the use of interpolation methods. Specifically for LPM and BLPM, Shepard interpolation was used. To prevent overfitting and to get a better error estimate, new methods were devised for determining the trust radius and the highest order of the interpolant.

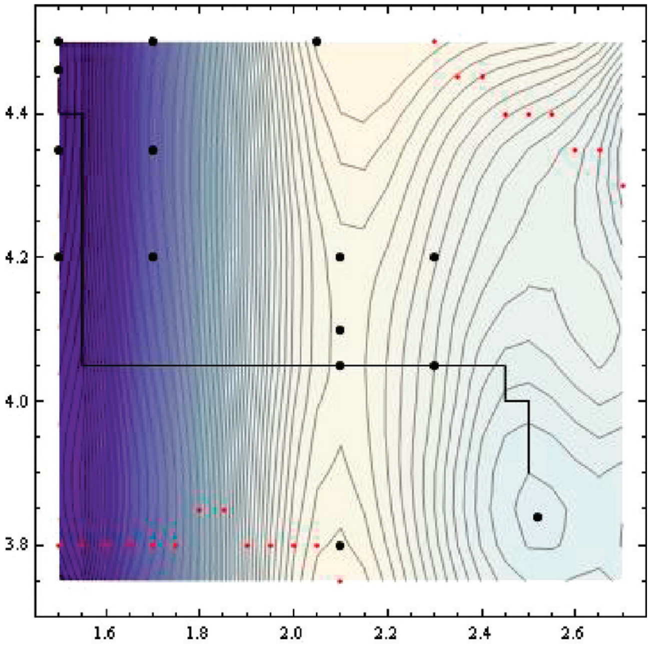


Figure 6. The RPES of the Uracil-DNA glycosylase cluster model shown in Figure 5. The *x*-axis is R1, the length of the glycosylic bond in Angstroms and the *y*-axis is R2, the distance between C1′ on the sugar ring and the oxygen on the water molecule, also in Angstroms. The boundary points are the string of points near the top and bottom of the plot. The dots are evaluated points and the line is the approximate MEP that is obtained by backtracing from the product.

The methods were shown to be able to rapidly locate the TS for an analytical function and two molecular systems. We attribute the success of these methods not only to the fact they work on the RPES, but also to the fact that, (a) they attempt to interpolate, rather than extrapolate, and (b) that they focus on providing an accurate description of the TS region, rather than the minimum energy path.

Acknowledgment. The authors thank the Chemistry Department of McMaster University, the Hamilton Foundation, NSERC, and the Canada Research Chairs for financial support.

References

(1) Kraut, D. A.; Carroll, K. S.; Herschlag, D. Challenges in enzyme mechanism and energetics. *Annu. Rev. Biochem.* **2003**, *72*, 517–571.

(2) Hammes, G. G. Multiple conformational changes in enzyme catalysis. *Biochemistry* **2002**, *41* (26), 8221.

(3) Morokuma, K.; Kato, S. *Potential Energy Surfaces and Dynamics Calculatoins*; Plenum: New York, 1981.

(4) Miller, W. H.; Handy, N. C.; Adams, J. E. Reaction Path Hamiltonian for Polyatomic Molecules. *J. Chem. Phys.* **1980**, *72* (1), 99.

(5) Gonzalez, C.; Schlegel, H. B. An improved algorithm for reaction path following. *J. Chem. Phys.* **1989**, *90* (4), 2154–2161.

(6) Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993**, *93* (7), 2395–2417.

(7) Bofill, J. M.; Anglada, J. M. Finding transition states using reduced potential-energy surfaces. *Theor. Chem. Acc.* **2001**, *105* (6), 463–472.

(8) Bofill, J. M. Updated Hessian matrix and the restricted step method for locating transition structures. *J. Comput. Chem.* **1994**, *15* (1), 1–11.

(9) Culot, P.; Dive, G.; Nguyen, V.; Ghuysen, J. A quasi-Newton algorithm for first-order saddle-point location. *Theor. Chim. Acta* **1992**, *82*, 189–205.

(10) Munro, L. J.; Wales, D. J. Defect migration in crystalline silicon. *Phys. Rev. B* **1999**, *59*, 3969–3980.

(11) Kumeda, Y.; Wales, D. J.; Munro, L. J. Transition states and rearrangement mechanisms from hybrid eigenvector-following and density functional theory.: Application to C10H10 and defect migration in crystalline silicon. *Chem. Phys. Lett.* **2001**, *341* (1), 185.

(12) Baker, J. *J. Comput. Chem.* **1986**, *7*, 385.

(13) Henkelman, G.; Jónsson, H. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Chem. Phys.* **1999**, *111*, 7010.

(14) Burger, S. K.; Ayers, P. W., Methods for finding transition states on reduced potential energy surfaces. J. Chem. Phys. 2010, (accepted).

(15) E, W. N.; Ren, W. Q.; Vanden-Eijnden, E. Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *J. Chem. Phys.* **2007**, *126* (16), 164103.

(16) E, W. N.; Ren, W. Q.; Vanden-Eijnden, E. String method for the study of rare events. *Phys. Rev. B* **2002**, *66* (5), 052301.

(17) Henkelman, G.; Jonsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **2000**, *113* (22), 9978–9985.

(18) Henkelman, G.; Uberuaga, B. P.; Jonsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **2000**, *113* (22), 9901–9904.

(19) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. String method in collective variables: Minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.* **2006**, *125* (2), 024106.

(20) Peters, B.; Heyden, A.; Bell, A. T. A growing string method for determining transition states: Comparison to the nudged elastic band and string methods. *J. Chem. Phys.* **2004**, *120* (17), 7877–7886.

(21) Quapp, W. A growing string method for the reaction pathway defined by a Newton trajectory. *J. Chem. Phys.* **2005**, *122* (17), 174106.

(22) Quapp, W. Reaction pathways and projection operators: Application to string methods. *J. Comput. Chem.* **2004**, *25* (10), 1277–1285.

(23) Burger, S. K.; Yang, W. T. Sequential quadratic programming method for determining the minimum energy path. *J. Chem. Phys.* **2007**, *127* (16), 164107.

(24) Burger, S. K.; Yang, W. T. Quadratic string method for determining the minimum-energy path based on multiobjective optimization. *J. Chem. Phys.* **2006**, *124* (5), 054109.

(25) Ayala, P. Y.; Schlegel, H. B. A combined method for determining reaction paths, minima, and transition state geometries. *J. Chem. Phys.* **1997**, *107* (2), 375–384.

(26) Jensen, F. Locating transition structures by mode following: A comparison of six methods on the Ar8 Lennard-Jones potential. *J. Chem. Phys.* **1995**, *102*, 6706.

(27) Peng, C.; Ayala, P.; Schlegel, H.; Frisch, M. Using redundant internal coordinates to optimize equilibrium geometries and transition states. *J. Comput. Chem.* **1996**, *17* (1), 49–56.

(28) Budzelaar, P. Geometry optimization using generalized, chemically meaningful constraints. *J. Comput. Chem.* **2007**, *28* (13), 2226–2236.

(29) Ischtwan, J.; Collins, M. A. Molecular-Potential Energy Surfaces by Interpolation. *J. Chem. Phys.* **1994**, *100* (11), 8080–8088.

(30) Collins, M. A. Molecular potential-energy surfaces for chemical reaction dynamics. *Theor. Chem. Acc.* **2002**, *108* (6), 313–324.

(31) Burger, S. K.; Liu, Y.; Sarkar, U.; Ayers, P. W. Moving least-squares enhanced Shepard interpolation for the fast marching and string methods. *J. Chem. Phys.* **2009**, *130*, 024103.

(32) Dawes, R.; Thompson, D. L.; Guo, Y.; Wagner, A. F.; Minkoff, M. Interpolating moving least-squares methods for fitting potential energy surfaces: Computing high-density potential energy surface data from low-density ab initio data points. *J. Chem. Phys.* **2007**, *126* (18), 084107.

(33) Kawano, A.; Guo, Y.; Thompson, D. L.; Wagner, A. F.; Minkoff, M. Improving the accuracy of interpolated potential energy surfaces by using an analytical zeroth-order potential function. *J. Chem. Phys.* **2004**, *120* (14), 6414–6422.

(34) Dey, B. K.; Ayers, P. W. Computing tunneling paths with the Hamilton-Jacobi equation and the fast marching method. *Mol. Phys.* **2007**, *105* (1), 71–83.

(35) Dey, B. K.; Ayers, P. W. A Hamilton-Jacobi type equation for computing minimum potential energy paths. *Mol. Phys.* **2006**, *104* (4), 541–558.

(36) Dey, B. K.; Bothwell, S.; Ayers, P. W. Fast marching method for calculating reactive trajectories for chemical reactions. *J. Math. Chem.* **2007**, *41* (1), 1–25.

(37) Dey, B. K.; Janicki, M. R.; Ayers, P. W. Hamilton-Jacobi equation for the least-action/least-time dynamical path based on fast marching method. *J. Chem. Phys.* **2004**, *121* (14), 6667–6679.

(38) Crittenden, D. L.; Jordan, M. J. T. Interpolated potential energy surfaces: How accurate do the second derivatives have to be. *J. Chem. Phys.* **2005**, *122* (4).

(39) Bettens, R. P. A.; Collins, M. A. Learning to interpolate molecular potential energy surfaces with confidence: A Bayesian approach. *J. Chem. Phys.* **1999**, *111* (3), 816–826.

(40) Jordan, M. J. T.; Thompson, K. C.; Collins, M. A. The Utility of Higher-Order Derivatives in Constructing Molecular-Potential Energy Surfaces by Interpolation. *J. Chem. Phys.* **1995**, *103* (22), 9669–9675.

(41) Schatz, G. C. The analytical representation of electronic potential-energy surfaces. *Rev. Mod. Phys.* **1989**, *61* (3), 669–688.

(42) Farwig, R. Rate of convergence of shepard global interpolation formula. *Math. Comput.* **1986**, *46* (174), 577–590.

(43) Draper, N. R.; Smith, H. *Applied Regression Analysis*. 2nd ed.; John Wiley & Sons, Inc: New York, 1980.

(44) Neter, J.; Kutner, M.; Wasserman, W.; Nachtsheim, C., *Applied Linear Statistical Models*; McGraw-Hill/Irwin: New York, 1996.

(45) Burger, S.; Liu, Y.; Ayers, P. Fast Marching Method Fortran 90 code v. 1.0, 2009.

(46) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R. ; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al.Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*, Gaussian, Inc.: Wallingford, CT, 2003.

(47) Lu, C. L.; Liu, Y. D.; Zhong, R. G. Theoretical Investigation of mono- and bi-function alkylating agents. *J. Mol. Struct. THEOCHEM* **2009**, *893*, 106–110.

(48) Anderson, L. M.; Souliotis, V. L.; Chhabra, S. K.; Moskal, T. J.; Hargaugh, S. D.; Kyrtopouls, S. A. *Int. J. Cancer* **1996**, (66), 130.

(49) Parikh, S.; Walcher, G.; Jones, G.; Slupphaug, G.; Krokan, H.; Blackburn, G.; Tainer, J. Uracil-DNA glycosylase-DNA substrate and product structures: Conformational strain promotes catalytic efficiency by coupled stereoelectronic effects. *Proc. Natl. Acad. Sci.* **2000**, *97* (10), 5083.

(50) Dinner, A. R.; Blackburn, G. M.; Karplus, M. W. Uracil-DNA glycosylase acts by substrate autocatalysis. *Nature* **2001**, *413*, 752.

CT100012Y

JCTC Journal of Chemical Theory and Computation

Electronic Continuum Model for Molecular Dynamics Simulations of Biological Molecules

I. V. Leontyev and A. A. Stuchebrukhov*

Department of Chemistry, University of California Davis, One Shields Avenue, Davis, California 95616

Received November 2, 2009

Abstract: Electronic polarizability is an important factor in molecular interactions. In the conventional force fields such as AMBER or CHARMM, however, there is inconsistency in how the effect of electronic dielectric screening of Coulombic interactions, inherent for the condensed phase media, is treated. Namely, the screening appears to be accounted for via effective charges only for neutral moieties, whereas the charged residues are treated as if they were in a vacuum. As a result, the electrostatic interactions between ionized groups are exaggerated in molecular simulations by a factor of about 2. The model discussed here, MDEC (Molecular Dynamics in Electronic Continuum) provides a theoretical framework for modification of the standard nonpolarizable force fields to make them consistent with the idea of uniform electronic screening of partial atomic charges. The present theory states that the charges of ionized groups and ions should be scaled, i.e., reduced by a factor of about 0.7. In several examples, including the interaction between Na^+ ions, which is of interest for ion-channel simulations, and the dynamics of an important salt bridge in cytochrome *c* oxidase, we compared the standard nonpolarizable MD simulations with MDEC simulations and demonstrated that the MDEC charge scaling procedure results in more accurate interactions. The inclusion of electronic screening for charged moieties is shown to result in significant changes in protein dynamics and can give rise to new qualitative results compared with the traditional nonpolarizable force fields simulations.

1. Introduction

At present, the majority of molecular dynamics simulations are performed by using nonpolarizable models such as AMBER,[1,2] CHARMM,[3] GROMOS,[4] and OPLS.[5] Presumably, the effects of electronic polarization and screening of electrostatic interactions are incorporated in the effective charges and other empirical parameters of the force fields; however, the extent to which this is so has never been entirely clear. [Throughout the paper the term "electronic screening" means a reduction of the electric field and electrostatic interactions due to an electronic relaxation of the environment. The origin of the effect is discussed elsewhere, e.g., in ref 6.] The importance of electronic polarizability is well recognized: for example, roughly half of the solvation free energy of ions is due to electronic polarization of the medium, and the interaction between charges is roughly half as weak, due to only electronic screening compared with that in a vacuum; therefore a significant effort is being undertaken to develop accurate fully polarizable force fields for biomolecules, see, e.g., refs 7–14.

Yet, in many cases, nonpolarizable models have been remarkably successful in modeling complex molecular systems.[15] For example, the properties of liquid water are described quite accurately without introducing electronic polarizability explicitly; likewise, the hydration free energies can be computed quite accurately using nonpolarizable simulations.[16] On the other hand, the simulation of polarization effects in low-polar solvents, e.g., ethers,[14] and especially in nonpolar solvents, e.g., alkanes,[13,17] meets serious problems. The nonpolarizable models can also significantly underestimate the magnitude of the dielectric response in a low-dielectric interior protein environment. For example, the dielectric constant of the inner part of cytochrome *c* was found to be only about 1.5,[18] which is lower than pure

* Corresponding author e-mail: stuchebr@chem.ucdavis.edu.

10.1021/ct9005807 © 2010 American Chemical Society
Published on Web 04/08/2010

electronic dielectric constant $\varepsilon_{el} \cong 2.0$.[19] Many other shortcomings of nonpolarizable MD simulations have been recently discussed in the literature, see ref 20 and references therein.

The polarizable models aim at resolving the problems mentioned above. Most of such models involve various kinds of coupled polarizable sites[7–14] and the computationally expensive procedure of achieving self-consistency of polarization of such sites at each molecular dynamics time step. Although, with the Extended-Lagrangian technique,[7,11–14] the computation cost of polarizable simulations can be significantly reduced, the implementation of such models is yet to be completed; at present, even the simplest classical Drude oscillator model[11–14] is still not readily available for application to many biological systems.

As fully polarizable force fields are being developed, there is also a growing interest in improving the existing empirical nonpolarizable models to capture more accurately the effects of electronic polarization and screening in MD simulations. Given a specially designed (but empirical in nature) procedure of how the partial charges are selected,[1–3] the charges of neutral residues do reflect, at least approximately, the effects of electronic screening—in a way, how, for example, TIP3P or similar fixed-charge models of water do. One issue of concern, however, is that the electrostatic interactions of ions are described in standard nonpolarizable force fields, such as CHARMM or AMBER, by their original integer charges (e.g., ± 1, for Na^+ and Cl^-), i.e., as if these ions were in a vacuum, completely disregarding the effect of electronic dielectric ($\varepsilon = \varepsilon_{el}$) screening inherent to the condensed phase medium. The interaction of such bare charges obviously is overestimated by a factor of about 2 (the screening factor ε_{el} is about 2 for most of organic media[13]). Thus, in simulation of ion channels ions (e.g., several K^+ ions in the same channel, just a few angstroms apart) interact very strongly, and therefore their interactions are important to describe correctly (see, e.g., ref 21) or, for interaction of the ions with water molecules or other partial atomic charges of the protein, for that matter. The same is true for *charged* residues in the protein, such as Arg^+ or Glu^-, partial charges of which carry their original net values ±1. The use of the bare charges in nonpolarizable simulations would be appropriate for a vacuum, but not for the condensed phase, where all charges are essentially immersed in the electronic continuum, which weakens their interactions by about a factor of 2—a typical electronic (or high-frequency) dielectric constant ε_{el} of any organic material.

Given the phenomenological nature of the force fields, one can argue that, in fact, partial charges should be considered only as formal parameters. However, they are often used, for example, in hybrid QM/MM calculations, where one needs to evaluate the electric field of the protein medium to which the QM system is exposed. The use of CHARMM or AMBER charges in such calculations has become standard and has been adopted in many studies.[22,23] Obviously, the electric potential of the charges should reflect the electronic screening of the medium.

One may also think that the atomic charges, dipoles, etc. are chosen in the force fields in such a way as to make the medium "over-polarized"[14] so that the effective nuclear relaxation/polarization alone would reflect both the effects of electronic and actual nuclear polarization. In this case, however, there is a question of relaxation time scale: on the time scale of nuclear motion, the electronic polarization and screening occurs almost instantaneously, reducing at once all electrostatic interactions by a factor of 2, whereas the effective polarization evolves on the time scale of the nuclear motions.

The question arises then as to whether it is possible to introduce an appropriate scaling of bare charges of ionized groups and ions to correctly reflect the electronic screening? Here, we argue that the bare charges of the ionic groups can and should be scaled, in particular when the electrostatic potentials of such groups are considered.

More generally, we discuss a principle of uniform charge-scaling based on which one could systematically build a nonpolarizable force field for simulations of condensed media. The principle is based on a simple idea of a uniform electronic continuum, with an effective dielectric constant $\varepsilon \sim 2$, and point charges moving in it. The resulting model, which combines a nonpolarizable (fixed-charge) force field for nuclear dynamics (MD) with a phenomenological electronic continuum (EC) is referred to as MDEC (Molecular Dynamics in Electronic Continuum). In this model, the effects of electronic screening are reduced to simple scaling of the partial charges. The model is similar but not equivalent to standard nonpolarizable force fields used in most MD simulations; we propose a simple scaling procedure that makes nonpolarizable force fields such as AMBER and CHARMM uniformly consistent with the idea of electronic screening, which naturally improves the quality of these force fields.

Several examples of MDEC calculations and the effects of electronic polarization will be discussed, including the interaction between Na^+ ions, which is of interest for ion-channel simulations, and the dynamics of an important salt bridge in cytochrome *c* oxidase.

2. MDEC Model

A detailed discussion of the MDEC model is given in previous publications.[24,25] Here, we restate main features of the model essential for subsequent calculations.

2.1. Screening Effect and Effective Charges. As frequently stated in the literature,[26] the partial atomic charges of nonpolarizable models, e.g., TIP3P[27] or SPC/E,[28] empirically incorporate the effect of electronic polarization in molecular interactions. There are different aspects of electronic polarization, however, that differently affect electrostatic interactions between individual molecules. Thus, the molecular dipole moment enhancement, usually considered[1–3,26] in the context of the electronic polarization, increases a strength of electrostatic interactions. On the other hand, the effect of electronic dielectric screening results in the reduction of the electrostatic interactions. Both factors are important for interaction of noncharged molecules; however, for interactions of ions, or ionized groups, where the direct Coulomb interaction dominates, the screening of the Coulomb interaction is of prime importance. Here, the screening effect will be described in terms of charge scaling.

Consider two ions, Q_1 and Q_2, in a solvent modeled by the dielectric of ε; the charges are located at the center of spheres with corresponding ionic radii, R_1 and R_2 (no dielectric inside the spheres), and separated by the distance r. An effective interaction between these ions is given by the potential of mean force (PMF):

$$\mathrm{PMF} \equiv U^{\mathrm{eff}}(r) = \Delta G(r) - \Delta G(\infty) \qquad (2.1)$$

where $\Delta G = \Delta G_{\mathrm{vac}} + \Delta G_{\mathrm{solv}}$ is the total free energy of the system (ion pair + solvent) composed of its vacuum (no solvent, ΔG_{vac}) and solvation (ΔG_{solv}) components; $\Delta G(\infty)$ is the sum of free energies of individual ions. The gradient of PMF $U^{\mathrm{eff}}(r)$ over r gives the effective force acting between ions Q_1 and Q_2 in the medium. Since we are interested in only electrostatic interactions, nonelectrostatic components of the free energies will be neglected. In the case when $r > R_1 + R_2$, the solvation free energy of the ion pair in the dielectric is accurately approximated by the well-known relation (15) from ref 29 and PMF is obtained as[29]

$$U^{\mathrm{eff}}(r) = \frac{Q_1 Q_2}{\varepsilon r} \qquad (2.2)$$

which shows that Coulomb interactions are reduced (screened) by a factor $1/\varepsilon$ due to relaxation of the polarizable environment. A more detailed treatment of the origin of the screening effect is given, e.g., in ref 6 and also in ref 25.

The magnitude of the screening factor ε depends on which part of the medium relaxation is considered explicitly (as moving charges q_i) and which part is described phenomenologically as a polarizable dielectric.[30] Since in nonpolarizable microscopic models the atomic motions are described explicitly, the screening factor should include only the electronic component of the medium polarization, $\varepsilon = \varepsilon_{\mathrm{el}}$. The static (i.e., time-independent) dielectric approximation in this case is quite accurate, because on the time scale of nuclear motion the electronic polarization occurs almost instantaneously, reducing at once all interatomic electrostatic interactions by a factor of $\varepsilon_{\mathrm{el}}$. The phenomenological parameter $\varepsilon_{\mathrm{el}}$ is known from the experiment as a high-frequency dielectric permittivity ($\varepsilon_{\mathrm{el}} = n^2$, where n is a refraction index of the medium) and typically is about 2. The resulting model, which combines a nonpolarizable (fixed-charge) force field for nuclear dynamics (MD) and a phenomenological electronic continuum (EC) for the electronic polarization is referred to as MDEC.[25]

The MDEC model[25] considers charges q_i moving in an electronic polarizable continuum of known dielectric constant $\varepsilon_{\mathrm{el}}$. In the uniform dielectric, all electrostatic interactions are scaled by a factor $1/\varepsilon_{\mathrm{el}}$. Since interactions are quadratic in charges, the effect of electronic dielectric screening can be taken into account implicitly by using scaled partial charges, $q_i^{\mathrm{eff}} = q_i/\sqrt{\varepsilon_{\mathrm{el}}}$; in this case, the Coulomb interaction between sites i and j automatically has the correct form $q_i^{\mathrm{eff}} q_j^{\mathrm{eff}}/r_{ij} = q_i q_j/\varepsilon_{\mathrm{el}} r_{ij}$ without explicitly introducing factor $1/\varepsilon_{\mathrm{el}}$. The unscaled original charges q_i are difficult to specify a priori in general (they are not the same as partial charges of a condensed medium molecule in a vacuum, see ref 25), unless one deals with ions or ionized groups in a protein, whose unscaled net charges are known. But charges q_i^{eff} can be found empirically by fitting experimental data[27,28] or scaled ab initio interaction energies.[3]

2.2. Solvation Free Energy. In the MDEC model, when the solvation free energy of a group is considered, the electronic polarization free energy is treated explicitly. The free energy consists of the nuclear part ΔG_{nuc} evaluated by MD and the pure electronic polarization energy part ΔG_{el} evaluated by using the polarizable continuum model[31] (i.e., by solving the Poisson equation with corresponding boundary conditions, with dielectric constant $\varepsilon = 1$ inside the solute region and $\varepsilon = \varepsilon_{\mathrm{el}}$ outside):

$$\Delta G = \Delta G_{\mathrm{nuc}} + \Delta G_{\mathrm{el}} \qquad (2.3)$$

Such an approach to electrostatic solvation free energy calculations, eq 2.3, was shown to work well both in high- and low-dielectric media[24,32] and will be further elaborated in this paper.

When the interaction of a solute with solvent molecules is considered in an MDEC simulation (in evaluating the ΔG_{nuc} part), the solute partial charges (found in an appropriate quantum-mechanical calculation, in a vacuum, or in a dielectric environment) should be scaled by $1/\sqrt{\varepsilon_{\mathrm{el}}}$, like all other charges when the forces between atoms are considered. If no scaling of solute charges was employed in the MD simulation, which is typical for standard MD simulations, see e.g. refs 16 and 33, the free energies obtained from MD, ΔG_{MD}, should be corrected directly afterward. Since in the linear response approximation the solvation free energy is quadratic in charges of the solute, ΔG_{MD} should be corrected by a factor $1/ \varepsilon_{\mathrm{el}}$, giving $\Delta G_{\mathrm{nuc}} = \Delta G_{\mathrm{MD}}/\varepsilon_{\mathrm{el}}$. The total MDEC polarization free energy of the medium then is

$$\Delta G = \frac{1}{\varepsilon_{el}} \Delta G_{\mathrm{MD}} + \Delta G_{\mathrm{el}} \qquad (2.4)$$

where ΔG_{MD} as stated above is the electrostatic solvation free energy obtained in nonpolarizable MD using unscaled solute charges (standard approach) and ΔG_{el} is the pure electronic part of the free energy. A more detailed description of the free energy simulation technique accounting for the electronic polarization can be found in refs 24 and 25.

2.3. Dielectric Constant of the Medium. The dielectric constant of the medium is often employed in the continuum electrostatic calculations, e.g., for solvation free energy evaluation. In microscopic calculations, on the other hand, the solvation free energy is obtained directly from MD simulations. The question arises often as to what is the effective dielectric constant of the medium, $\varepsilon_{\mathrm{MD}}$, that corresponds to a specific microscopic model of the system. The free energy relationships discussed in the previous section allow one to make a connection between the total (static) dielectric constant, ε_0, which includes both nuclear and electronic polarization effects, and the dielectric constant of nonpolarizable MD simulations, $\varepsilon_{\mathrm{MD}}$, which does not explicitly describe pure electronic polarization of the medium.

Suppose we consider a spherical ion or a pair of spherical ions; in this case, according to ref 29, the solvation energies will be proportional to their corresponding Born factors: $\Delta G \sim (1 - 1/\varepsilon_0)$, $\Delta G_{\mathrm{el}} \sim (1 - 1/\varepsilon_{\mathrm{el}})$ and $\Delta G_{\mathrm{MD}} \sim (1 - 1/\varepsilon_{\mathrm{MD}})$. Using eq 2.4 for the relationship between these free energies, we find

$$\varepsilon_0 = \varepsilon_{\mathrm{MD}} \cdot \varepsilon_{\mathrm{el}} \tag{2.5}$$

That is, the total dielectric constant of the medium ε_0 is not equivalent to that reproduced by the (nonpolarizable) MD simulation, $\varepsilon_{\mathrm{MD}}$; instead, the relationship between the two is given by the above formula. This indeed has been directly demonstrated[25] for several systems.

The above relation can be also obtained using a well-known expression[34] for the static dielectric constant:

$$\varepsilon_0 = \varepsilon_{\mathrm{el}} + \frac{4\pi}{3Vk_{\mathrm{B}}T}\langle M^2\rangle \tag{2.6}$$

Here, $\langle M^2\rangle$ is the mean square fluctuation of the total dipole of the dielectric sample V; k_{B} and T are the Boltzmann constant and temperature, respectively. According to the MDEC scaling procedure, the actual dipole moment μ of particles in the bulk is related to the effective moment μ^{eff} of these particles in the nonpolarizable model as $\mu = \sqrt{\varepsilon_{\mathrm{el}}}\mu^{\mathrm{eff}}$; therefore, $\langle M^2\rangle = \varepsilon_{\mathrm{el}} \langle M^2_{\mathrm{MD}}\rangle$, where $\langle M^2_{\mathrm{MD}}\rangle$ is the mean square fluctuation of the dipole moment observed in a nonpolarizable MD. Thus, eq 2.5 is obtained from eq 2.6 by noticing that $\varepsilon_{\mathrm{MD}}$ is defined via fluctuation $\langle M^2_{\mathrm{MD}}\rangle$ with $\varepsilon_{\mathrm{el}} = 1$ in eq 2.6.

Although the simple relation between dielectric constants eq 2.5 was derived using arguments strictly valid only for spherical ions, and for the bulk solvent modeled with periodic boundary conditions,[34] eq 2.5 is in fact more general and provides a good estimate of the static dielectric constant ε_0 in a wide range of different solutes.[18,35]

3. Applications of MDEC Model

3.1. Water Models. Many nonpolarizable force fields are essentially MDEC models. For example, TIP3P[27] or SPC/E[28] and similar models of water involve empirical charges that can be considered as scaled charges. TIP3P is particularly interesting in this regard as it is often used in biological simulations, and it serves as a reference for phenomenological parameter assignments of CHARMM.[3]

It is known that the dipole moment of a water molecule in a vacuum is 1.85D; in the liquid state, however, the four hydrogen bonds to which each water molecule is exposed on average strongly polarize the molecule, and its dipole moment falls somewhere in the range of 2.9D to 3.2D.[36-38] [It is recognized that in ab initio simulations of bulk water the water dipole cannot be defined unambiguously and depends on the partitioning scheme used;[39] as such, its actual value remains a matter of debate. Here, we rely upon calculations and the partitioning scheme of refs 37 and 38.] The significant increase of the dipole from $\mu_0 = 1.85$D to a value $\mu \approx 3$D, or even larger, is also supported by the Kirkwood–Onsager model,[40] see the Appendix, which estimates the enhanced polarization of a molecule due to the reaction field of the polarized environment. Yet, the dipole moment of the TIP3P water model is only 2.35D. The specific value of the TIP3P dipole moment can be understood as a scaled dipole, so that the dipole–dipole interactions are screened by the electronic continuum by a factor $1/\varepsilon_{\mathrm{el}}$. Indeed, if each dipole (or all partial charges) is scaled by a factor $1/\sqrt{\varepsilon_{\mathrm{el}}}$, one could consider interaction of the effective dipoles, $\mu^{\mathrm{eff}} = \mu/\sqrt{\varepsilon_{\mathrm{el}}} \simeq 2.35$D (for water $\varepsilon_{\mathrm{el}} = 1.78$), as if they were in a vacuum. This appears to be exactly what the fixed-charge water models do. Thus, charges of the TIP3P water model should be understood as scaled charges that reflect the effect of electronic screening.

The scaled nature of charges of the TIP3P water model is important to bear in mind when the interaction of such water models with a solute is considered. For example, if the charge of say a Na^+ ion is assigned to be +1 in a simulation, then it is obviously inconsistent with the charges of the water model, as the latter are scaled by a factor of $1/\sqrt{\varepsilon_{\mathrm{el}}}$, while the charge of the ion is not. Clearly the strength of the interaction is overestimated in this case by a missing factor of $1/\sqrt{\varepsilon_{\mathrm{el}}}$, i.e., about 0.7 (for proteins $\varepsilon_{\mathrm{el}} \sim 2$). The problem would not arise if the charge of the ion were appropriately scaled. (The reason why a seemingly incorrect charge gives reasonable aqueous solvation free energy is explained next.)

3.2. Conventional Force Fields. The conventional nonpolarizable force fields of AMBER,[1,2] CHARMM,[3] GROMOS,[4] or OPLS[5] are built on different principles than those discussed in this paper; yet the atomic partial charges of noncharged groups can be understood approximately as "scaled MDEC charges", because these empirical parameters were chosen in such a way as to reflect the condensed matter nature of the interaction. For example, in CHARMM, TIP3P water (an effective MDEC model) was used as a reference in the empirical procedure[3] of setting partial charges. In contrast, the charges of ionized groups *do not* reflect the effects of electronic screening.

In free energy simulations with nonpolarizable force fields (and unscaled charges), the pure electronic contribution to the electrostatic free energy is often completely ignored, as e.g. in refs 16 and 33. Yet, in many cases, such simulations pretty accurately reproduce experimental solvation energies; this may appear surprising, given the fact that about half of the total solvation free energy (for charged solutes typically 25–50 kcal/mol) comes from electronic polarization of the medium. In fact, the neglect of large electronic polarization free energy is almost completely compensated by the use of "incorrect" bare solute charges in such simulations. This fortuitous compensation of errors, however, occurs only in the high-dielectric media, as can be seen from the following argument.

Consider for example the Born solvation energy of Na^+ ion, $Q = +1$, in water; in simulations, one would have approximately

$$\Delta G = \frac{Q^2}{2R}\left(1 - \frac{1}{\varepsilon_{\mathrm{MD}}}\right) \tag{3.1}$$

where $\varepsilon_{\mathrm{MD}}$ is the dielectric constant of water that corresponds to a specific MD model employed in the calculation. No matter which model of water is used, $\varepsilon_{\mathrm{MD}}$ is much larger than unity; hence, the overall estimate of the solvation free energy is $Q^2/2R$, which is independent of properties of the solvent and can match pretty well the experimental value, provided the ionic radius R is chosen correctly. The interaction between two charges is taken to be then Q^2/r, completely disregarding the electronic screening of the interaction.

The MDEC model suggests instead that in MD simulations the charge Q should be scaled, *and* the electronic solvation free energy $\Delta G_{el} = Q^2/2R(1 - 1/\varepsilon_{el})$ is added explicitly. In this case, the nuclear part of the free energy calculated in MD will be $[(Q/\sqrt{\varepsilon_{el}})^2/2R](1 - 1/\varepsilon_{MD})$ and the total free energy, eq 2.3, is given by

$$\Delta G = \frac{(Q/\sqrt{\varepsilon_{el}})^2}{2R}\left(1 - \frac{1}{\varepsilon_{MD}}\right) + \frac{Q^2}{2R}\left(1 - \frac{1}{\varepsilon_{el}}\right) \quad (3.2)$$

Since $\varepsilon_{MD} = \varepsilon_0/\varepsilon_{el}$, the above expression correctly reproduces the expected result $Q^2/2R(1 - 1/\varepsilon_0)$. Notice that the charge is not scaled when the solvation is calculated in an electronic continuum. Notice also that it is only when $\varepsilon_{MD} \gg 1$ that the two expressions 3.1 and 3.2 approximately give the same result. Yet, for the interaction energy of two charges, the MDEC gives the correct expression $Q^2/r\varepsilon_{el}$, while the standard approach gives Q^2/r.

Given that the unscaled relation (eq 3.1) is only formally correct when $\varepsilon_{MD} = \varepsilon_0/\varepsilon_{el} \gg 1$, it is not surprising that the pure nonpolarizable approach works well in aqueous solutions ($\varepsilon_0/\varepsilon_{el} \sim 40$), as e.g. in refs 16 and 24; however, the approach fails (i.e., significantly underestimates the polarization effects) in low dielectric media ($\varepsilon_0/\varepsilon_{el} \sim 1$) as in refs 13, 14, and 41–43.

In contrast, an MDEC simulation of polarization effects is correct for both high- and low-dielectric media, as shown in the modeling of hydration free energies of ions,[24] dielectric constants of neat alcohols, alkanes,[25] and protein interiors of cytochrome *c* and cytochrome *c* oxidase[44] as well as nonequilibrium reorganization energies in water, dichloroethane, tetrahydrofuran, and supercritical carbon dioxide solvents.[32]

We next show that despite a complex nature of electronic polarization in a real system, the effect in practice can be described reasonably well by a simple charge scaling procedure; this opens a way to modify the standard force fields so as to improve the description of their charged groups by effectively incorporating the electronic screening of charges.

3.3. Ab Initio Interactions Modeled by Charge Scaling. Here, we consider the interaction of several charged species in ab initio calculations. The ab initio treatment captures the effects of electronic polarization of charged species themselves, while the effects of the polarization of the environment, and corresponding screening, are described here phenomenologically, by a continuum with dielectric $\varepsilon_{el} = 2.0$.

The interaction energies were calculated using a quantum-mechanical procedure identical to that of the CHARMM parametrization protocol,[3] with amino acids substituted by their corresponding model compounds (glutamate by propionate and arginine by *n*-propyl guanidinium). The isolated model compounds were optimized at the HF/6-31G(d) level.

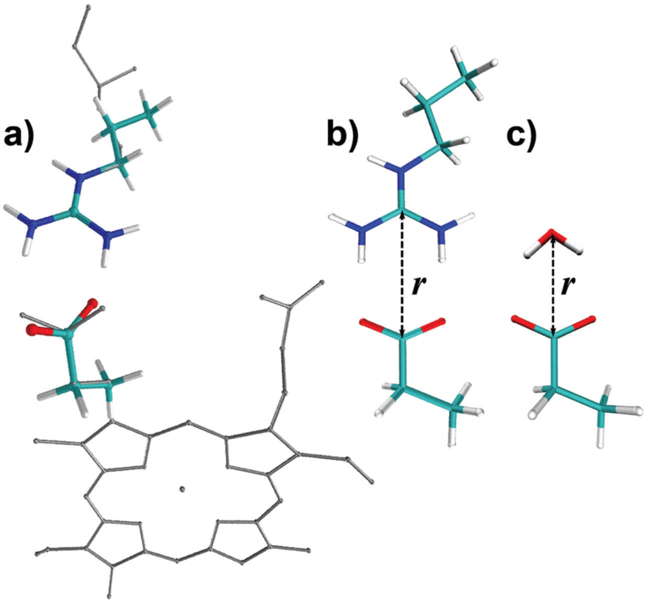


Figure 1. Mutual orientation of model compounds. (a) The initial configuration of the Glu^- and Arg^+ model compounds corresponds to the position of the salt bridge Arg438A−PropD of heme a_3 in cytochrome *c* oxidase (see section 3.5). (b) The optimized configuration of Glu^- and Arg^+ model compounds. (c) The optimized configuration of the Glu^- model compound and a water molecule. The interaction distance *r* is shown for optimized structures.

The optimized structures were then used to construct compound "supermolecules" consisting of a model compound and a single water molecule (or another compound). The supermolecule structures were optimized at the HF/6-31G(d) level by varying the interaction distance *r* and mutual angles, to find the optimum position with fixed monomer geometries. The mutual orientations of the model compounds in the supermolecules are shown for the interacting pairs Arg^+-Glu^- and Glu^--H_2O in Figure 1. The mutual angles were then fixed while the interaction distance *r* was varied. For each structure with a given *r*, the interaction energy was calculated as the difference between the total supermolecule energy and the sum of the individual monomer energies. The gas-phase interaction energies were calculated with model compounds in a vacuum, while the bulk-phase interactions were obtained with the model compounds immersed in the dielectric of $\varepsilon = 2$. The quantum-mechanical calculation in the dielectric utilized the PCM[31] technique and self-consistent reaction-field procedure implemented in Gaussian 03.[45] The PCM cavities were built using the Gaussian 03 default setup[45] for atomic radii without the generation of smoothing spheres (keyword NOADDSPH). In this united atom model, the radii of CH3, CH2, C, NH, NH2, OH2 and O are taken as 2.525 Å, 2.325 Å, 1.925 Å, 1.93 Å, 2.03 Å, 1.95 Å, and 1.75 Å, respectively. The calculations were performed with a surface grid element of 0.1 Å^2 average size.

In Figure 2a−c, the ab initio interactions between ions Na^+-Na^+ and Arg^+-Glu^- and between the Glu^- ion and water are compared with those modeled by the original and scaled CHARMM[3] force fields. (For amino acids, their corresponding model compounds are used.) In all cases, as expected, there is a significant screening effect of the

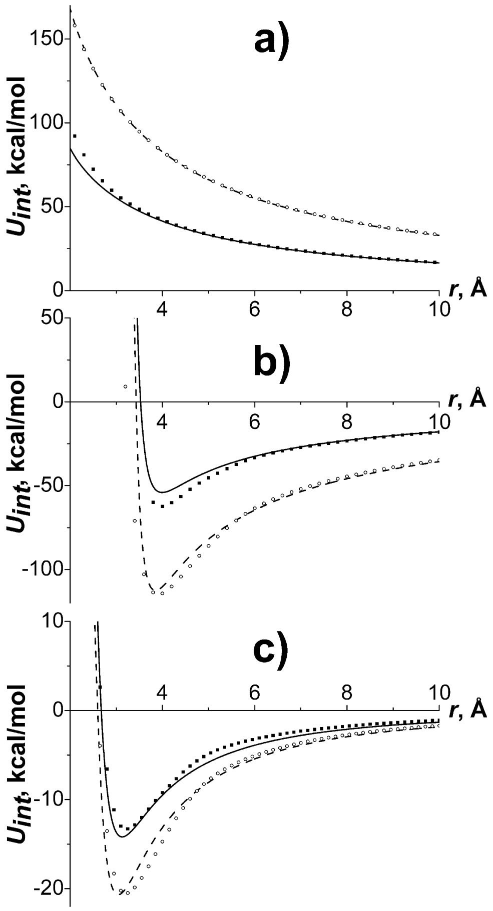


Figure 2. Interaction energies between (a) Na^+-Na^+ ions, (b) Arg^+ and Glu^- amino acids, and (c) Glu^- amino acid and water. Open circles stand for the energies obtained in the gas-phase HF/6-31G(d) calculation; filled squares are for the same interactions but calculated in the dielectric of $\varepsilon = 2.0$. The dashed lines represent interaction energies obtained by the standard CHARMM force field and using the TIP3P water model in (c). The solid lines represent the interaction energies obtained by the CHARMM force field with scaled charges ($\varepsilon_{el} = 2.0$) and the TIP3P water model in (c).

dielectric environment on the interaction energy; the effect as seen, however, can be pretty accurately reproduced by a simple scaling of charges. Notice that the charges are scaled by a factor $1/\sqrt{\varepsilon_{el}}$; in correspondence with what has been said about the TIP3P water model (section 3.1), in the example of the $Glu^- - H_2O$ pair, only charges of Glu^- were scaled. Notice how accurately the scaled CHARMM force field reproduces the results of ab initio calculations. Similar results are expected for other standard force fields (such as AMBER,[1,2] GROMOS,[4] etc.) where charged groups are also treated as having their original vacuum net charges.

Thus, the simple charge scaling procedure in standard nonpolarizable force fields can account for the effects of electronic screening not only in the interactions between ions but also between ions and water. Although, as seen in Figure 2a, some additional adjustments of the nonelectrostatic parameters might be useful to improve the interactions at shorter distances.

3.4. Forces between Ions in a Polarizable Environment. To test how well the phenomenological electronic continuum of the MDEC model describes a molecular solvent, with its structure and corresponding inhomogeneity of electronic polarization, here we examine a model of two ions A^- and A^+ dissolved in benzene; the low-dielectric environment is chosen to model the interior of a protein, or a lipid membrane. The solvent now is described by the polarizable Drude oscillator model,[11] whereas ions are treated by a standard nonpolarizable force field (Coulomb and Lennard-Jones interactions; the LJ parameters for ions correspond to those of Cl^- ion.) For such a system, we calculate the electrostatic part of the potential of mean force as defined by eq 2.1 and compare the results with those of scaled and unscaled CHARMM calculation, using the concepts of MDEC theory, see Figure 3. The PMF gradient over r gives the average electrostatic force acting between charged particles A^- and A^+ in the bulk. The solvation free energy $\Delta G(r)$ of ions was evaluated by three alternative techniques: by polarizable MD, by the standard MD technique using a nonpolarizable CHARMM force field, and by MD using eq 2.3 and the CHARMM force field with scaled ion charges, according to the MDEC model. Further details of the simulations are given in the Appendix.

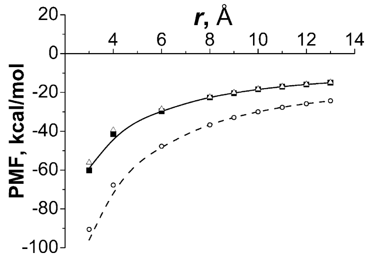


Figure 3. PMF for an ion pair A^+ and A^- in benzene. The squares, circles, and triangles stand for the MD results obtained with polarizable MD, nonpolarizable CHARMM, and CHARMM with scaled charges of the ions, respectively. Continuous curves are the least-squares fitting of the simulation points by the Coulomb function $-1/\varepsilon r$ (with the Ewald correction, see the Appendix). For polarizable simulations (solid line), the effective dielectric constant $\varepsilon = 1.88(\varepsilon_0)$ for nonpolarizable simulations (dashed line), $\varepsilon = 1.16(\varepsilon_{MD})$. The triangles correspond to *nonpolarizable* CHARMM simulations with *scaled* charges by a factor $1/\sqrt{(\varepsilon_0/\varepsilon_{MD})}$ according to the MDEC model, eq 2.3.

As seen in Figure 3, when the space between ionic spheres is larger than a size of solvent molecules, the effects of the solvent microscopic structure become unimportant, and the average interaction, both in polarizable and nonpolarizable models of benzene, can be approximated by a simple Coulomb law with an effective dielectric constant (obviously the LJ interactions are not important in this region). In the case of a polarizable Drude oscillator model for solvent benzene, the average interaction between ions is reproduced with an effective dielectric constant $\varepsilon_0 = 1.88$. [We notice that the experimental value of ε_0 for benzene is actually 2.3;[46] the underestimated MD value of ε_0 is a consequence of the reduced polarizability parameter employed in the benzene model,[11] which is ~20% lower than experimental benzene polarizability.]

According to MDEC theory, eq 2.5, the total dielectric constant of the medium ε_0 is a product of the electronic dielectric ε_{el} (due to Drude polarization of benzene mol-

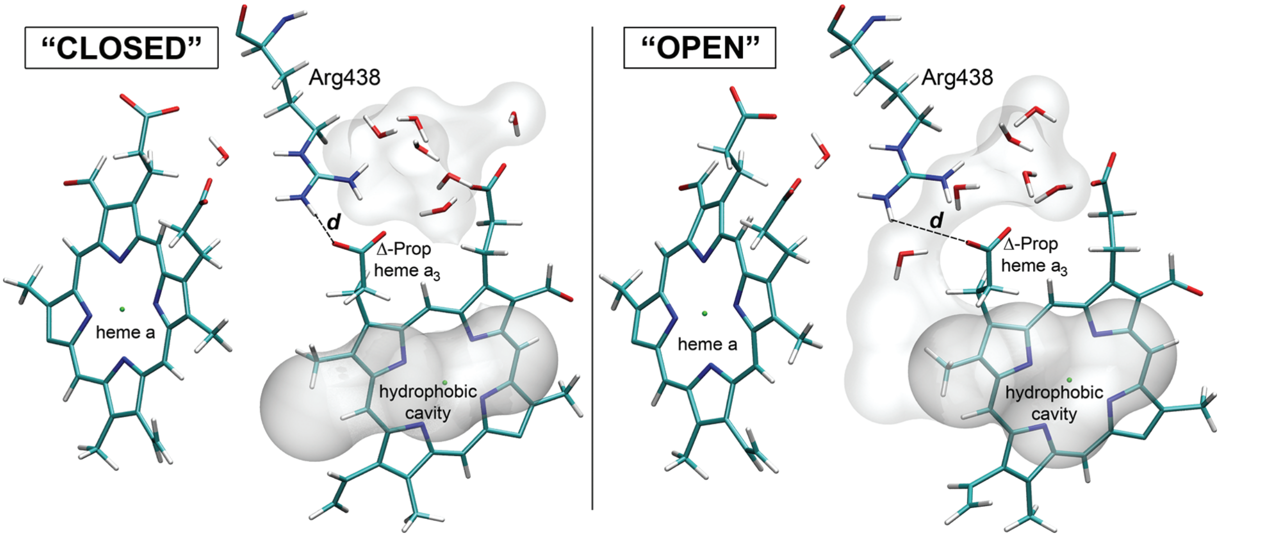


Figure 4. Salt bridge between Δ-propionate of heme a_3 and Arg438A in bovine cytochrome *c* oxidase. Configurations, when the gate is "OPEN" and "CLOSED", are shown.

ecules) and that of nuclei, $\varepsilon_{\mathrm{MD}}$. The latter was obtained in a separate simulation using the nonpolarizable CHARMM model of benzene, see Figure 3 (dashed line); the corresponding value is $\varepsilon_{\mathrm{MD}} = 1.16$. According to eq 2.5 then, the corresponding electronic dielectric constant of the polarizable model of benzene is $\varepsilon_{\mathrm{el}} = \varepsilon_0/\varepsilon_{\mathrm{MD}} = 1.62$. As seen in Figure 3 (solid line), in perfect agreement with MDEC theory, the results of *polarizable* benzene simulations are reproduced by scaling charges of ions (by a factor $1/\sqrt{1.62}$) and running *nonpolarizable* CHARMM simulations. Again, we see that all effects of electronic polarization can be incorporated by scaling charges of ions with a factor $1/\sqrt{\varepsilon_{\mathrm{el}}}$.

The significant deviation of the results of standard nonpolarizable MD from those of polarizable and MDEC techniques shown in Figure 3, in fact, can be rationalized without the PMF simulations. Since the scaling of Coulomb interactions for each microscopic model is given by the corresponding dielectric constant (as defined in section 2.3), the PMF profiles are approximated by the corresponding Coulomb functions $-1/\varepsilon_0 r$, $-1/\varepsilon_{MD} r$ and $-1/((\varepsilon_{\mathrm{el}}\varepsilon_{\mathrm{MD}})r)$ for the polarizable, nonpolarizable CHARMM, and MDEC techniques, respectively. Due to the relation (eq 2.5), PMF functions for polarizable MD and MDEC should be the same, while deviation from the CHARMM technique is estimated as $(1 - 1/\varepsilon_{\mathrm{el}})(1/\varepsilon_{\mathrm{MD}}r)$. Thus, for the low-dielectric media where $\varepsilon_{\mathrm{el}} = 2$ and $\varepsilon_{\mathrm{MD}} \sim 1$, the deviation is $\sim 1/2r$, which is significant even for larger separation distances (~16 kcal/mol for $r = 10$ Å). In the high-dielectric media ($\varepsilon_{\mathrm{MD}} \gg 1$), however, the difference will be much smaller. For instance, in water ($\varepsilon_{\mathrm{el}} = 1.8$, $\varepsilon_{\mathrm{MD}} \sim 100$ for the TIP3P model[47]), the deviation will be just $\sim 1/225r$, which is ~0.5 kcal/mol even for the shortest separation $r \leq 4$ Å (contact ion pair: $r \leq 2R_{\mathrm{vdW}}$). Since the difference 0.5 kcal/mol is on the order of statistical uncertainty of MD, the missing electronic screening effect is not noticeable in the standard nonpolarizable simulations of water solutions.

3.5. Dynamics of Salt Bridges in Proteins. To demonstrate the significance of accounting for the electronic polarization in protein dynamics simulations, we modeled fluctuations of an important salt bridge (Arg438A−PropD of heme a_3) in cytochrome *c* oxidase (C*c*O), see Figure 4. This salt bridge (SB) controls water penetration to the hydrophobic cavity in the catalytic center of C*c*O.[48,49] The strength of the electrostatic interaction of the salt bridge determines the rate of its opening/closing and, as a result, the probability of water transfer to/from the catalytic cavity.

The distance *d* between O2D of Δ-propionate and 2HH2 of Arg438 (as shown in Figure 4) has been chosen to characterize the fluctuations of the salt bridge gate during an MD run. The AMBER[1,2] force field was used. The simulation setup is similar to that of ref 44. Details of the MD simulations are given in the Appendix. The distribution functions for distance *d* obtained with scaled and original unscaled charges are shown in Figure 5a. Here, no water in the cavity was included in the simulation.

It is seen that the SB dynamics become qualitatively different once electrostatic interactions between the charged Arg438$^+$ and the COO$^-$ group of Δ-propionate are reduced by a factor of $1/\varepsilon_{\mathrm{el}}$ (in the simulations, $\varepsilon_{\mathrm{el}} = 2.0$). In contrast to the standard MD simulations,[49] the fluctuations observed in the scaled model are significantly larger, so that the internal water can now easily pass through the opened SB gate and enter the catalytic cavity. In fact, during a 5 ns MD run with scaled charges, several such water transitions were observed.

In Figure 5b, the distribution functions of *d* are shown from simulations that included water in (and around) the catalytic cavity of the enzyme. As we already pointed out, the electronic screening affects not only charge−charge interactions but interaction with water as well. Here, the TIP3P model is taken without modification; the charge scaling affects only the salt bridge groups. As seen in Figure 5b when the effects of electronic screening are included, even more dramatic changes are observed.

Thus, standard (unscaled charges) MD simulations with and without water in the cavity lead to the conclusion that the salt bridge is formed 100% of the time; here stability of the salt bridge is quantified by the criterion $d < 3$ Å, while the bridge is observed only 98% or even 63% of the time in

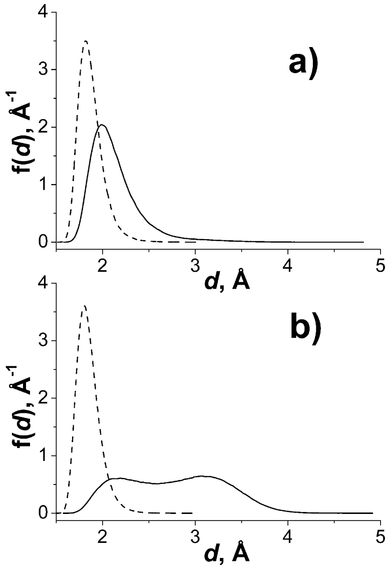


Figure 5. Distribution functions of the distance *d* between O2D (Δ-propionate of heme a_3) and 2HH2 (Arg438A) of C*c*O salt bridge: (a) no water in the catalytic cavity; (b) 4 water molecules added to the cavity. Dashed lines represent distributions obtained in the standard MD, while solid lines stand for the distributions obtained in the MD with scaled charges of the ionized groups.

simulations with scaled charges without (see Figure 5a) and with water (see Figure 5b), respectively.

It is clear that the account for electronic screening of charged groups can give rise to *qualitatively different* results in simulations of proteins. As we have shown, this can be achieved in a computationally effective way by simple charge scaling of ionized groups in the protein.

Unfortunately, there are no direct experimental data on the dynamics of the salt bridge discussed here to verify our proposal of electronic screening. However, as we argued in this paper, such a scaling is obvious from a theoretical point of view. An indirect comparison with an experiment, and support of charge scaling, is provided by some other computational studies, such as Zhu et al.,[50] where a heuristic approximation for the charge scaling of ionized side chains (variable dielectric constant ≥ 2) somewhat similar to ours was employed, which resulted in significant improvement in both side chain and loop prediction for protein conformations.

4. Conclusions

There is inconsistency in how the effect of electronic screening of Coulombic interactions, inherent for the condensed phase media, is treated in the conventional force fields such as AMBER[1,2] or CHARMM.[3] Namely, the screening appears to be accounted for via effective charges only for neutral moieties, whereas the charged residues are treated as if they were in a vacuum. As a result, the electrostatic interactions between ionized groups are exaggerated in molecular simulations by a factor of about 2.

The discussed MDEC (Molecular Dynamics in Electronic Continuum) model provides a theoretical framework within which the charge screening of both ionized and neutral residues can be achieved on the same footing. In a few examples, we compared the standard nonpolarizable MD simulations with MDEC simulations and demonstrated how the charges of ionized groups can be rescaled to correspond to the MDEC model. The present theory states that the charges of ionized groups of the protein, as well as charges of ions, in simulations with conventional nonpolarizable force fields such as CHARMM, AMBER, etc. should be scaled, i.e., reduced by a factor of about 0.7, to reflect the electronic screening of the condensed medium appropriate for biological dynamics simulations. If the charge-scaling procedure is employed, in the solvation free energy calculations, the electronic polarization energy should be treated explicitly, i.e., explicitly added to the nuclear part of the free energy, eq 2.3. *Ab initio* calculations of interaction energies (section 3.3) and MD simulations of the potential of mean force (section 3.4) indicate that the MDEC charge scaling procedure results in more accurate interactions not only between ions but also between ions and nonpolarizable water models, such as TIP3P, often used in biological simulations. Given the above examples and earlier reports,[24,25,32] we conclude that the MDEC model provides a way to more accurate modeling of condensed-phase molecular systems.

The ignored electronic screening between ions in standard MD simulations may be unnoticeable in molecular simulations of high-dielectric media such as water solutions; however, it has a dramatic effect (section 3.5) in the dynamics of charged systems such as salt bridges in a low-dielectric protein interior.

Acknowledgment. We would like to acknowledge many stimulating discussions of the subject of the paper with Professor Toby Allen of UC Davis and Dr. Igor Vorobyov. This work has been supported in part by the NSF grant PHY 0646273 and NIH GM054052.

Appendix

Kirkwood–Onsager Model of Water in the Bulk. Consider a polarizable point dipole in the middle of a sphere cut in the medium of dielectric constant ε_0; the radius of the sphere is a, the permanent dipole value (dipole of water in vacuum) is μ_0, and the dipole polarizability (i.e., electronic polarizability of water) is α. The increase of the dipole due to the reaction field of the polarized medium is[40]

$$\mu = \frac{\mu_0}{1 - \dfrac{\alpha}{a^3}\dfrac{2(\varepsilon_0 - 1)}{2\varepsilon_0 + 1}} \tag{A1}$$

Taking $\varepsilon_0 = 78$ for the dielectric constants of water, $\mu_0 = 1.85$D for a water dipole in a vacuum and $\alpha = 1.47$ Å^3 for the polarizability of water, for reasonable values of radius a in the range of 1.4–1.6 Å, one obtains a range of possible values of μ: 2.9–3.9D [Here the radius 1.4 Å corresponds to one-half of the most probable OO distance in the radial distribution function of bulk water, whereas 1.6 Å corresponds to one-half of the average intermolecular distance in the bulk]. Although this is a crude model, it nevertheless strongly indicates that the actual dipole of water in the bulk is much larger than the empirical value of 2.35D of the TIP3P model.

Potential of Mean Force Calculations. Consider two generic ions, A^- and A^+, with ionic radii, R^- and R^+, separated by a distance r, dissolved in polarizable benzene. In the continuum solvent approximation (in the dielectric region $r > R^- + R^+ + 2R_{\text{solvent}}$), the estimate for PMF $U^{\text{eff}}(r)$ of the ions is given by eq 2.2. If the finite system is simulated with periodic boundary conditions, then the expression (eq 2.2) should be corrected by the Ewald term:

$$U^{\text{eff}}(r) = -\frac{1}{\varepsilon r} + \frac{1}{\varepsilon}\text{Ewald}(r) \tag{A2}$$

where Ewald(r) is the interaction of A^- and A^+ ions with all periodic images (excluding the solvent). The effective value of ε is determined by fitting the continuous expression A2 to the MD simulated PMF data.

PMF was simulated employing eq 2.1 and different microscopic models of the benzene solvent: the polarizable Drude oscillator model[11] and the nonpolarizable CHARMM[3] and MDEC models. The ions were treated by a standard nonpolarizable force field (Coulomb and Lennard-Jones interactions; the LJ parameters for ions correspond to those of the Cl^- ion[3]). The solvation free energies of the ionic pair were evaluated by the standard technique of thermodynamic integration within the linear response approximation. The MDEC free energies were calculated employing eq 2.3 and the CHARMM force field (with scaled charges of ions) for MD simulation of the nuclear part of the free energy ΔG_{nuc}. The electronic part of the free energy ΔG_{el} was estimated by solving the continuum electrostatic problem for point charges +1,−1 in a dielectric of $\varepsilon = 1$ inside the solute region (defined by the radii R^- and R^+: $R^- = R^+ = R_{\text{vdw}}(\text{Cl}^-)$) and $\varepsilon = \varepsilon_{\text{el}}$ outside. The Poisson equation was solved by the PCM[31] technique implemented in our code,[51] which has been extensively tested especially on the two-site system. The value of ε_{el} corresponding to the polarizable model[11] was found to be $\varepsilon_0/\varepsilon_{\text{MD}} = 1.62$. The nuclear part of the free energy ΔG_{nuc} was obtained in MD simulation using the CHARMM force field with scaled (by factor $1/\sqrt{1.62}$) charges of the ions.

The MD system consisted of two ions A^- and A^+ and 350 benzene molecules within the MD box with a ~38 Å edge. A prior computation has shown that all potential energy terms calculated for the same configuration of the system by two programs CHARMM (version c32b1) and Gromacs[52] coincide with each other to a high precision. For practical reasons, all simulations were carried out by the Gromacs[52] MD package; however, results are expected to be identical to a CHARMM simulation. The electrostatic interactions were treated by the PME technique with a real space cutoff of 12 Å. The new Berendsen thermostat with a stochastic term and a Berendsen barostat were used with the coupling constant 0.1 ps, to keep the temperature at 298 K and the pressure at 1 atm. For each separation distance r, the system was equilibrated first during a 1 ns run, followed by a 5 ns data collection run. The MD time step was 1 fs in the nonpolarizable and 0.5 fs in polarizable simulations. The positions of the Drude particles were optimized with a frequency of 1 fs in the self-consistent procedure as implemented in Gromacs.[52]

To quantify the screening effect, the value of ε was determined by fitting the function "$- 1/\varepsilon r + \text{Ewald}(r)/\varepsilon +$ const" to the simulated PMF data. In the microscopic simulations, the const term appears due to some quadrupolar (nondielectric) and nonlinear contributions to the solvent polarization, but it does not essentially affect the force between the ions, $-\nabla U^{\text{eff}}(r)$, in the dielectric region ($r > R^- + R^+ + 2R_{\text{solvent}}$).

The comparison of PMF $U^{\text{eff}}(r)$ functions obtained with different force fields is shown in Figure 3. The shown PMF profiles are shifted by the corresponding values of the const, to reflect the correct boundary conditions at infinity, $U^{\text{eff}}(\infty) = 0$.

Fluctuations of Salt Bridges in Proteins. Cytochrome *c* oxidase is chosen as a probe protein due to its central role in energy metabolism in aerobic cells and our previous experience with this molecule. The enzyme is modeled by two subunits A and B taken from the fully reduced bovine heart cytochrome *c* oxidase structure (PDB code 1V55[53]). Additional water molecules were added between Glu-242 and the D channel, consistent with the *Rh. sphaeroides* structure.[54] The system was in the P_M state with Cu_A oxidized and heme *a* reduced. The partial charges of the redox centers, heme *a*, heme a_3, Cu_A, and Cu_B, are borrowed from ref 55, and the appropriate redox state was achieved by evenly distributing appropriate charge between metal ion and atoms directly coordinated to it. The structure was protonated according to the equilibrium protonation state of the residues determined previously.[48] The titratable residues with a proton occupancy larger than ~0.3 were treated as fully protonated, while all others as deprotonated, so as to avoid partial proton occupancies and the net charge on the molecule.

For MD simulations, Gromacs[52] with an AMBER force field ported by Eric J. Sorin[56] was used. The TIP3P model was used for water. It has been tested previously[44] that all potential energy terms calculated by the MD package AMBER7[57] and Gromacs for the same configuration of dehydrated C*c*O coincide with high precision.

The MD cell is formed by the protein immersed in a bath of water molecules and counterions K^+ and Cl^- with an effective salt concentration of 100 mM. The shortest distance from the protein to the edge of the MD box is 5 Å. Position restraints were applied to solvent exposed C_α atoms and membrane exposed heavy protein atoms. The electrostatic interactions were treated by the PME technique with a real space cutoff of 12 Å.

The new Berendsen thermostat with a stochastic term and coupling constant 0.3 ps was applied to keep the temperature at 298 K; also, the Berendsen pressure coupling with a reference pressure of 1 atm and a coupling constant of 0.5 ps was applied. All bonds to H atoms were constrained with the LINCS algorithm. The initial structure was energy-minimized and equilibrated in a 5 ns MD run before the sampling run. The sampling was collected in a 5 ns MD run with a 1 fs time step.

The interatomic distance *d* between O2D of Δ-propionate and 2HH2 of Arg438 (as shown in Figure 4) has been chosen to characterize the fluctuations of the salt bridge. The distribution functions of *d* are shown in Figure

5. Of particular interest for the C*c*O mechanism are configurations with large distances *d*; for such configurations, the water molecules from the internal cavity above the catalytic center can penetrate to the catalytic cavity of the enzyme, as shown in Figure 4.

References

(1) Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(2) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(3) Mackerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R.; Evanseck, J.; Field, M.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(4) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hunenberger, P. H.; Kruger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; Vdf Hochschulverlag AG an der ETH Zurich: Zurich, Switzerland, 1996.

(5) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(6) Frohlich, H. *Theory of Dielectrics*; Clarendon Press: Oxford, U.K., 1949.

(7) Rick, S. W.; Stuart, S. J.; Berne, B. J. Dynamical fluctuating charge force fields: Application to liquid water. *J. Chem. Phys.* **1994**, *101*, 6141–6156.

(8) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. Development of an Accurate and Robust Polarizable Molecular Mechanics Force Field from ab Initio Quantum Chemistry. *J. Phys. Chem. A* **2004**, *108*, 621–627.

(9) Patel, S.; Mackerell, A. D., Jr.; Brooks, C. L. CHARMM fluctuating charge force field for proteins: II Protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *J. Comput. Chem.* **2004**, *25*, 1504–1514.

(10) Stuart, S. J.; Berne, B. J. Effects of Polarizability on the Hydration of the Chloride Ion. *J. Phys. Chem.* **1996**, *100*, 11934–11943.

(11) Lopes, P. E. M.; Lamoureux, G.; Roux, B.; MacKerell, A. D. Polarizable empirical force field for aromatic compounds based on the classical drude oscillator. *J. Phys. Chem. B* **2007**, *111*, 2873–2885.

(12) Anisimov, V. M.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D., Jr. Polarizable Empirical Force Field for the Primary and Secondary Alcohol Series Based on the Classical Drude Model. *J. Chem. Theory Comput.* **2007**, *3*, 1927–1946.

(13) Vorobyov, I. V.; Anisimov, V. M.; MacKerell, A. D. Polarizable Empirical Force Field for Alkanes Based on the Classical Drude Oscillator Model. *J. Phys. Chem. B* **2005**, *109*, 18988–18999.

(14) Vorobyov, I.; Anisimov, V. M.; Greene, S.; Venable, R. M.; Moser, A.; Pastor, R. W.; MacKerell, A. D. Additive and Classical Drude Polarizable Force Fields for Linear and Cyclic Ethers. *J. Chem. Theory Comput.* **2007**, *3*, 1120–1133.

(15) Karplus, M. Molecular dynamics simulations of biomolecules. *Acc. Chem. Res.* **2002**, *35*, 321–323.

(16) Hummer, G.; Pratt, L. R.; Garcia, A. E. Free energy of ionic hydration. *J. Phys. Chem.* **1996**, *100*, 1206–1215.

(17) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

(18) Simonson, T.; Perahia, D. Internal and Interfacial Dielectric Properties of Cytochrome c from Molecular Dynamics in Aqueous Solution. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92*, 1082–1086.

(19) Simonson, T.; Perahia, D. Microscopic Dielectric Properties of Cytochrome c from Molecular Dynamics Simulations in Aqueous Solution. *J. Am. Chem. Soc.* **1995**, *117*, 7987–8000.

(20) Halgren, T. A.; Damm, W. Polarizable force fields. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236–242.

(21) Allen, T. W.; Andersen, O. S.; Roux, B. Molecular dynamics - potential of mean force calculations as a tool for understanding ion permeation and selectivity in narrow channels. *Biophys. Chem.* **2006**, *124*, 251–267.

(22) Gao, J. L.; Truhlar, D. G. Quantum mechanical methods for enzyme kinetics. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467–505.

(23) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.

(24) Vener, M. V.; Leontyev, I. V.; Basilevsky, M. V. Computations of solvation free energies for polyatomic ions in water in terms of a combined molecular-continuum approach. *J. Chem. Phys.* **2003**, *119*, 8038–8046.

(25) Leontyev, I. V.; Stuchebrukhov, A. A. Electronic continuum model for molecular dynamics simulations. *J. Chem. Phys.* **2009**, *130*, 085102.

(26) Guillot, B. A reappraisal of what we have learnt during three decades of computer simulations on water. *J. Mol. Liq.* **2002**, *101*, 219–260.

(27) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(28) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(29) Platzman, R.; Franck, J. The Role of the Hydration Configuration in Electronic Processes Involving Ions in Aqueous Solution. *Z. Phys.* **1954**, *138*, 411–431.

(30) Schutz, C. N.; Warshel, A. What are the dielectric constants of proteins and how to validate electrostatic models. *Proteins* **2001**, *44*, 400–417.

(31) Miertus, S.; Scrocco, E.; Tomasi, J. Electrostatic Interaction of a Solute with a Continuum - a Direct Utilization of Abinitio Molecular Potentials for the Prevision of Solvent Effects. *Chem. Phys.* **1981**, *55*, 117–129.

(32) Vener, M. V.; Tovmash, A. V.; Rostov, I. V.; Basilevsky, M. V. Molecular Simulations of Outersphere Reorganization Energies in Polar and Quadrupolar Solvents. The Case of Intramolecular Electron and Hole Transfer. *J. Phys. Chem. B* **2006**, *110*, 14950–14955.

(33) Geerke, D. P.; van Gunsteren, W. F. Force Field Evaluation for Biomolecular Simulation: Free Enthalpies of Solvation of Polar and Apolar Compounds in Various Solvents. *ChemPhysChem* **2006**, *7*, 671–678.

(34) Neumann, M.; Steinhauser, O. Computer simulation and the dielectric constant of polarizable polar systems. *Chem. Phys. Lett.* **1984**, *106*, 563–569.

(35) Simonson, T.; Brooks, C. L. Charge Screening and the Dielectric Constant of Proteins: Insights from Molecular Dynamics. *J. Am. Chem. Soc.* **1996**, *118*, 8452–8458.

(36) Badyal, Y. S.; Saboungi, M. L.; Price, D. L.; Shastri, S. D.; Haeffner, D. R.; Soper, A. K. Electron distribution in water. *J. Chem. Phys.* **2000**, *112*, 9206–9208.

(37) Silvestrelli, P. L.; Parrinello, M. Structural, electronic, and bonding properties of liquid water from first principles. *J. Chem. Phys.* **1999**, *111*, 3572–3580.

(38) Sharma, M.; Resta, R.; Car, R. Dipolar correlations and the dielectric permittivity of water. *Phys. Rev. Lett.* **2007**, *98*, 247401.

(39) Delle Site, L.; Lynden-Bell, R. M.; Alavi, A. What can classical simulators learn from ab initio simulations? *J. Mol. Liq.* **2002**, *98*−*9*, 79–86.

(40) Kirkwood, J. G. The dielectric polarization of polar liquids. *J. Chem. Phys.* **1939**, *7*, 911–919.

(41) Simonson, T.; Carlsson, J.; Case, D. Proton Binding to Proteins: pKa Calculations with Explicit and Implicit Solvent Models. *J. Am. Chem. Soc.* **2004**, *126*, 4167–4180.

(42) Archontis, G.; Simonson, T. Proton Binding to Proteins: A Free-Energy Component Analysis Using a Dielectric Continuum Model. *Biophys. J.* **2005**, *88*, 3888–3904.

(43) Vorobyov, I.; Li, L.; Allen, T. W. Assessing Atomistic and Coarse-Grained Force Fields for Protein? Lipid Interactions: the Formidable Challenge of an Ionizable Side Chain in a Membrane. *J. Phys. Chem. B* **2008**, *112*, 9588–9602.

(44) Leontyev, I. V.; Stuchebrukhov, A. A. Dielectric relaxation of cytochrome c oxidase: Comparison of the microscopic and continuum models. *J. Chem. Phys.* **2009**, *130*, 085103.

(45) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B. G. E.; Scuseria, M. A.; Robb, J. R.; Cheeseman, J. A.; Montgomery, J. T.; Vreven, K. N.; Kudin, J. C.; Burant, J. M.; Millam, S. S.; Iyengar, J.; Tomasi, V.; Barone, B.; Mennucci, M.; Cossi, G.; Scalmani, N.; Rega, G. A.; Petersson, H.; Nakatsuji, M.; Hada, M.; Ehara, K.; Toyota, R.; Fukuda, J.; Hasegawa, M.; Ishida, T.; Nakajima, Y.; Honda, O.; Kitao, H.; Nakai, M.; Klene, X.; Li, J. E.; Knox, H. P.; Hratchian, J. B.; Cross, C.; Adamo, J.; Jaramillo, R.; Gomperts, R. E.; Stratmann, O.; Yazyev, A. J.; Austin, R.; Cammi, C.; Pomelli, J. W.; Ochterski, P. Y.; Ayala, K.; Morokuma, G. A.; Voth, P.; Salvador, J. J.; Dannenberg, V. G.; Zakrzewski, S.; Dapprich, A. D.; Daniels, M. C.; Strain, O.; Farkas, D. K.; Malick, A. D.; Rabuck, K.; Raghavachari, J. B.; Foresman, J. V.; Ortiz, Q.; Cui, A. G.; Baboul, S.; Clifford, J.; Cioslowski, B. B.; Stefanov, G.; Liu, A.; Liashenko, P.; Piskorz, I.; Komaromi, R. L.; Martin, D. J.; Fox, T.; Keith, M. A.; Al-Laham, C. Y.; Peng, A.; Nanayakkara, M.; Challacombe, P. M. W.; Gill, B.; Johnson, W.; Chen, M. W.; Wong, C.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision B.04; Gaussian, Inc.: Pittsburgh, PA, 2003.

(46) Lide, D. R. *CRC Handbook of Chemistry and Physics*, 84th ed.; CRC Press/Taylor and Francis: Boca Raton, FL, 2003.

(47) Hochtl, P.; Boresch, S.; Bitomsky, W.; Steinhauser, O. Rationalization of the dielectric properties of common three-site water models in terms of their force field parameters. *J. Chem. Phys.* **1998**, *109*, 4927–4937.

(48) Popovic, D. M.; Stuchebrukhov, A. A. Electrostatic Study of the Proton Pumping Mechanism in Bovine Heart Cytochrome c Oxidase. *J. Am. Chem. Soc.* **2004**, *126*, 1858–1871.

(49) Wikstrom, M.; Ribacka, C.; Molin, M.; Laakkonen, L.; Verkhovsky, M.; Puustinen, A. Gating of proton and water transfer in the respiratory enzyme cytochrome c oxidase. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 10478–10481.

(50) Zhu, K.; Shirts, M. R.; Friesner, R. A. Improved Methods for Side Chain and Loop Predictions via the Protein Local Optimization Program: Variable Dielectric Model for Implicitly Improving the Treatment of Polarization Effects. *J. Chem. Theory Comput.* **2007**, *3*, 2108–2119.

(51) Vener, M. V.; Leontyev, I. V.; Dyakov, Y. A.; Basilevsky, M. V.; Newton, M. D. Application of the Linearized MD Approach for Computing Equilibrium Solvation Free Energies of Charged and Dipolar Solutes in Polar Solvents. *J. Phys. Chem. B* **2002**, *106*, 13078–13088.

(52) van der Spoel, D.; Lindahl, E.; Hess, B.; van Buuren, A. R.; Apol, E.; Meulenhoff, P. J.; Tieleman, D. P.; Sijbers, A. L. T. M.; Feenstra, K. A.; van Drunen, R.; Berendsen, H. J. C. Gromacs User Manual version 4.0, www.gromacs.org (accessed Mar 2010).

(53) Tsukihara, T.; Shimokata, K.; Katayama, Y.; Shimada, H.; Muramoto, K.; Aoyama, H.; Mochizuki, M.; Shinzawa-Itoh, K.; Yamashita, E.; Yao, M.; Ishimura, Y.; Yoshikawa, S. The low-spin heme of cytochrome c oxidase as the driving element of the proton-pumping process. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 15304–15309.

(54) Qin, L.; Hiser, C.; Mulichak, A.; Garavito, R. M.; Ferguson-Miller, S. Identification of conserved lipid/detergent-binding sites in a high-resolution structure of the membrane protein cytochrome c oxidase. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 16117–16122.

(55) Tashiro, M.; Stuchebrukhov, A. A. Thermodynamic Properties of Internal Water Molecules in the Hydrophobic Cavity around the Catalytic Center of Cytochrome c Oxidase. *J. Phys. Chem. B* **2005**, *109*, 1015–1022.

(56) Sorin, E. J.; Pande, V. S. Exploring the Helix-Coil Transition via All-Atom Equilibrium Ensemble Simulations. *Biophys. J.* **2005**, *88*, 2472–2493.

(57) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Wang, J.; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crow-ley, M.; Tsui, V.; Gohlke, H.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *AMBER 7 Users' Manual.*: University of California: San Francisco, 2002.

CT9005807

JCTC Journal of Chemical Theory and Computation

Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field

Devleena Shivakumar,*,† Joshua Williams,‡ Yujie Wu,† Wolfgang Damm,†
John Shelley,‡ and Woody Sherman†

*Schrödinger, Inc., 120 West 45th Street, 17th Floor, New York, New York 10036 and
101 SW Main Street, Suite 1300, Portland, Oregon 97204*

Received November 4, 2009

Abstract: The accurate prediction of protein−ligand binding free energies is a primary objective in computer-aided drug design. The solvation free energy of a small molecule provides a surrogate to the desolvation of the ligand in the thermodynamic process of protein−ligand binding. Here, we use explicit solvent molecular dynamics free energy perturbation to predict the absolute solvation free energies of a set of 239 small molecules, spanning diverse chemical functional groups commonly found in drugs and drug-like molecules. We also compare the performance of absolute solvation free energies obtained using the OPLS_2005 force field with two other commonly used small molecule force fields—general AMBER force field (GAFF) with AM1-BCC charges and CHARMm-MSI with CHelpG charges. Using the OPLS_2005 force field, we obtain high correlation with experimental solvation free energies ($R^2 = 0.94$) and low average unsigned errors for a majority of the functional groups compared to AM1-BCC/GAFF or CHelpG/CHARMm-MSI. However, OPLS_2005 has errors of over 1.3 kcal/mol for certain classes of polar compounds. We show that predictions on these compound classes can be improved by using a semiempirical charge assignment method with an implicit bond charge correction.

Introduction

The accurate determination of free energies (solvation and binding) has been a primary objective for computational methods since the inception of molecular modeling and computer-aided drug design.[1,2] The solvation free energy is a critical component of many important problems in the fields of chemistry, biology, and pharmaceutical sciences, such as protein folding, conformational transitions, protein−ligand binding, and transport of drugs across biological membranes.[3,4] Hence, a large number of studies in the literature have focused on the quantitative determination of solvation free energies. Furthermore, the prediction of solvation free energies provides an excellent opportunity for the testing of methods and force fields.

Although the concept and underlying theory of free energy calculations were laid out several decades ago,[5,6] free energy calculations have not been used routinely in drug discovery for three main reasons. First, free energy calculations are notoriously complicated to set up and take a lot of effort even for experts. Additionally, they are computationally expensive, often taking days of processor time even for a single calculation. Finally, the accuracy of the methodology applied to pharmaceutically relevant systems has not been systematically validated across a broad range of targets and ligands. The former two problems have been partly solved by the availability of automated free energy perturbation (FEP) and thermodynamic integration (TI) workflows within modern molecular dynamics (MD) and Monte Carlo (MC) software packages and by the continued growth (in number and speed) of accessible computational resources. The systematic validation across diverse and relevant systems remains a limitation and is one of the primary focuses of our group and others in the field.

* Corresponding author. Telephone: 646-366-9555. Fax: 646-366-9551. E-mail: devleena.shivakumar@schrodinger.com.

† Schrödinger, Inc., New York, New York.

‡ Schrödinger, Inc., Portland, Oregon.

10.1021/ct900587b © 2010 American Chemical Society
Published on Web 04/14/2010

The calculation of absolute solvation free energies is a more tractable problem than predicting binding free energies, since the solvent molecules equilibrates more quickly around a small organic solute than around the binding site of the protein, and there are fewer internal degrees of freedom to consider with a small molecule than with a protein−ligand complex. Since absolute solvation free energies have been experimentally determined for hundreds of small molecules, it allows for a direct comparison between the experimental and calculated values. Also, several groups have published calculated absolute solvation free energies, making it possible to compare the performance of various methodologies and force fields.

There are many physics-based methods available to calculate solvation free energies. Often a compromise has to be made between the speed and the accuracy of the model. At one end of the spectrum are methods that use continuum solvation models, such as generalized Born (GB)[7,8] and Poisson−Boltzmann (PB) models,[9−11] which are computationally efficient but neglect the role of explicit water molecules. On the other hand, explicit solvent models are considered to be more rigorous because they more closely represent the underlying physical system of interest. However, explicitly modeling solvent molecules adds to the computational cost and complexity of the simulations. The FEP or TI that are coupled to MD or MC computer simulations are commonly used to compute solvation free energies.[12,13] The first successful FEP calculations of relative solvation free energies were published in 1985 by Jorgensen and Ravimohan, who computed the difference between ethane and methanol using MC simulations and found excellent agreement with experiment.[14]

Recent studies have compared the performance of explicit solvent MD/FEP in predicting absolute solvation free energies for a few hundred small molecules using various force fields and solvent models.[15−18] A study comparing the accuracy of different implicit solvent models using MD/FEP calculations on a set of 504 neutral small molecules using GROMACS 3.3 MD simulation package[19] reported root mean square (rms) errors ranging from 2.01 to 2.43 kcal/mol, depending on the implicit solvent model, and correlation coefficients (R^2) from 0.69 to 0.77.[16] The same group reported an rms error of 1.26 kcal/mol, a correlation coefficient of 0.89, and a mean error of 0.68 for the same set of molecules using explicit solvent and the general AMBER force field (GAFF)[20] with partial charges from the Austin Model 1 using bond charge corrections (AM1-BCC). Another study by Shivakumar et al.[18] used a set of 239 neutral small molecules to compare implicit versus explicit solvent molecules using two popular small-molecule general atom force fields, GAFF[20] and CHARMm-MSI,[21] along with various charge models, such as semiempirical AM1-BCC[22] and quantum chemical charge methods, such as CHarges from ELectrostatic Potentials using a Grid-based method (CHelpG)[23] and Restrained Electrostatic Surface Potential fit (RESP).[24] The results from this study also showed that the best results were obtained with the AM1-BCC/GAFF combination, having a correlation coefficient of 0.87.

Here, we extend the work reported by Shivakumar et al.[18] by using the OPLS_2005 force field to compute absolute solvation free energies for the same set of 239 neutral small molecules and to compare the results to those previously obtained with the GAFF and CHARMm-MSI. We also discuss the use of an improved semiempirical charge assignment method based on the CM1A procedure of Cramer and Truhlar along with implicit bond charge correction terms, called CM1A-BCC, to improve the solvation free energy predictions for polar molecules.[25] Our ultimate goal is to extend the application of MD/FEP to calculate relative binding free energies of congeneric molecules to pharmaceutically relevant targets, which will be addressed in future work.

Theory

Molecular Dynamics/Free Energy Perturbation (MD/FEP). The MD/FEP simulations were carried out using the Desmond MD package, version 20108, as distributed by Schrödinger.[26−28] Desmond is a relatively new MD engine developed by D. E. Shaw Research and can be used to run a variety of molecular simulations, including minimization, standard MD, simulated annealing, and replica exchange (REMD) simulations, in addition to FEP. Desmond implements a number of novel algorithmic optimizations that are explained in detail elsewhere,[28−31] such as minimizing interprocessor communication, ensuring conservation of energy even with single precision calculations, and computing long-range electrostatic interactions in the Fourier space with a smooth particle mesh Ewald[32] implementation.

Total Free Energy. The FEP method allows one to compute the change in free energy of a chemical system as it evolves from state A to state B. The computation is essentially based on Zwanzig's equation:[6]

$$\Delta G_{A\rightarrow B} = -k_BT\ln\langle\exp(-(U_B - U_A)/k_BT)\rangle_A \qquad (1)$$

where k_B is the Boltzmann constant, T is the absolute temperature, U_A and U_B are the potential energies of the A and B states, respectively, and the $\langle\cdots\rangle_A$ denotes the ensemble average over configurations sampled from the A state. For absolute solvation free energy calculations, we can define A as the solute molecule in the gas phase and B as the solute molecule in the solution phase. Solvation is, thus, regarded as a transfer process where the solute molecule enters the solvent from the ideal gas. It is worth noting that the ΔG under this definition is called the *transfer* free energy—the free energy of transferring the solute from ideal gas to aqueous solution of the same concentration.

A thermodynamic cycle for transferring a solute from vacuum to solvent phase is shown in Figure 1. The solvation free energy can be calculated along the vertical line (2a and b in Figure 1), which consists of annihilating or creating the solute molecule both in vaccuo and in solvent. This scheme is also known as double annihilation since the solute is annihilated (or created) from both the vacuum and solution phases (routes 2a and b, respectively, in Figure 1). Since the total free energy change in the thermodynamic cycle must be zero, the following can be deduced

$$\Delta G_{\text{solvation}} = \Delta G^{0}_{\text{annihilation}} - \Delta G^{1}_{\text{annihilation}} \tag{2}$$

Instead of directly simulating the process in 1a from Figure 1, a common approach is to employ a thermodynamic cycle and compute ΔG for the solvation process by eq 2. An alternative approach is used in Desmond that takes advantage of the fact that we recover the above-defined $[\text{solute}]_{\text{vaccuo}}$ state from $[\text{solute}]_{\text{solvent}}$ by turning off the interactions between the solute and the other molecules. An advantage of this approach is that only one FEP simulation is needed instead of two annihilations. We define the potential of the system as a function of two order parameters, λ_{vdW} and λ_{coul}, to scale the van der Waals (vdW) and electrostatic potentials, respectively, between the solute (M*)* and the solvent (solv). The total potential ($U(\lambda_{\text{vdW}}, \lambda_{\text{coul}})$) of the system is described by the eq 3:

$$U(\lambda_{\text{vdW}}, \lambda_{\text{coul}}) = U^{\text{vdW}}_{\text{M,solv}}(\lambda_{\text{vdW}}) + U^{\text{coul}}_{\text{M,solv}}(\lambda_{\text{coul}}) + U^{\text{bonded}}_{\text{M}} + U^{\text{nonbonded}}_{\text{M}} + U^{\text{bonded}}_{\text{solv}} + U^{\text{nonbonded}}_{\text{solv}} \tag{3}$$

where $U^{\text{vdW}}_{\text{M, solv}}(\lambda_{\text{vdW}})$ and $U^{\text{coul}}_{\text{M, solv}}(\lambda_{\text{coul}})$ are the vdW and electrostatic potentials between the solute and the solvent, respectively; $U^{\text{bonded}}_{\text{M}}$ and $U^{\text{nonbonded}}_{\text{M}}$ are the bonded and the nonbonded potentials of the solute, respectively; and $U^{\text{bonded}}_{\text{solv}}$ and $U^{\text{nonbonded}}_{\text{solv}}$ are the bonded and the nonbonded potentials of the solvent, respectively. This description recovers the A state for ($\lambda_{\text{vdW}}, \lambda_{\text{coul}}) = 0$ and the B state for ($\lambda_{\text{vdW}}, \lambda_{\text{coul}}) = 1$. For values between 0 and 1, it gives a hybrid system that has scaled similarities to both A and B.

To avoid numerical problems commonly associated with $\lambda_{\text{vdW}} = 0$ for the Lennard-Jones (LJ) 12-6 potential, we use the following soft-core potential:[33]

$$U^{\text{vdW}}_{\text{M,solv}}(r, \lambda_{\text{vdW}}) = 4\lambda_{\text{vdW}}\varepsilon\left[\left(\frac{\sigma^6}{\alpha(1-\lambda_{\text{vdW}})^2\sigma^6 + r^6}\right)^2 - \frac{\sigma^6}{\alpha(1-\lambda_{\text{vdW}})^2\sigma^6 + r^6}\right] \tag{4}$$

α is a positive constant that controls the magnitude of the soft-core term: $\alpha(1 - \lambda_{\text{vdW}})^2\sigma^6$. We used a value of 0.5 for α. Note that this soft-core potential is a function of λ_{vdW} and reduces to the standard LJ potential for $\lambda_{\text{vdW}} = 1$. As λ_{vdW} approaches 0, the soft-core term eliminates the problematic singularity in the LJ potential.

$$\begin{tikzcd}[column sep=huge, row sep=huge]
{[\text{Solute}]_{\text{vaccuo}}} \arrow[r, "\Delta G_{\text{solvation}}", "1a"'] \arrow[d, "\Delta G^{0}_{\text{annihilation}}", "2a"'] & {[\text{Solute}]_{\text{solvent}}} \arrow[d, "\Delta G^{1}_{\text{annihilation}}", "2b"'] \\
{[\text{Nothing}]_{\text{vaccuo}}} \arrow[r, "\Delta G{=}0", "1b"'] & {[\text{Nothing}]_{\text{solvent}}}
\end{tikzcd}$$

Figure 1. Thermodynamic cycle for calculating absolute solvation free energies.

A correction term that accounts for the missing long-range dispersion energy due to the cutoff of the vdW potentials[34] is also added to the absolute free energy obtained using MD/FEP in Desmond. Assuming a homogeneous solvent beyond the cutoff radius, this term can be analytically obtained as follows:

$$E_{\text{lrc}} = -\frac{16}{3}\frac{\pi\rho}{r^3_{\text{cut}}}\sum_i \varepsilon_{i\text{w}}\sigma^6_{i\text{w}} \tag{5}$$

where ρ is the water number density of the system, $\varepsilon_{i\text{w}}$ and $\sigma_{i\text{w}}$ are LJ parameters for the i-th solute atom and water oxygen atom (water hydrogen atoms do not have LJ interactions for the water models used in this study), the summation is over the solute atoms, and r_{cut} is the cutoff radius for the LJ potential. Adding this correction yields results that are nearly independent of the LJ cutoff distance for reasonable cutoff values at a negligible additional calculation cost.

Bennett Acceptance Ratio Method. We used the Bennett acceptance ratio (BAR) method to estimate the free energy difference from the MD simulations.[35] For each window (i), we sampled the potential energy difference in both the forward [$W^{\text{f}}_i = U_{i+1}(x_i) - U_i(x_i)$] and the reverse [$W^{\text{r}}_i = U_{i+1}(x_i) - U_i(x_i)$] directions. The free energy difference ΔG_i between window i and i+1 is computed by solving the nonlinear equation:

$$\sum_{j=1}^{L_i}\frac{1}{1 + \frac{L_i}{L_{i+1}}\exp[(W^{\text{f}}_{i,j} - \Delta G_i)/k_{\text{B}}T]} - \sum_{k=1}^{L_{i+1}}\frac{1}{1 + \frac{L_{i+1}}{L_i}\exp[(\Delta G_i - W^{\text{r}}_{i+1,k})/k_{\text{B}}T]} = 0 \tag{6}$$

where L_i is the number of data in the forward or reverse energy ensemble. The statistical uncertainty of ΔG_i for window i was estimated using the subsampling bootstrap method,[36] and the total uncertainty is reported as the rms error, RMSE (square root of the sum of the squares of the errors), across all windows.

The OPLS_2005 Force Field. The OPLS_2005 force field, as implemented in the Schrödinger suite 2008, follows the functional form of the OPLS-AA family of force fields with additional stretch, bend, and torsional parameters for better coverage of ligand functional groups.[37] The nonbonded parameters are taken from the OPLS-AA force field, which were developed to reproduce heats of vaporizations and densities of pure liquids, are retained as published.[38–45] The torsional parameters specific to peptides (not relevant in this work) originate from Jensen et al. and Kaminski et al.[39,40] Stretch and bend parameters for 112 compounds are retained as published, and all other stretch and bend parameters are adjusted to reproduce the structures obtained with quantum mechanics at the B3LYP/6-31G* method and basis set. With the exception of the protein specific parameters, all torsional parameters were fit to reproduce the conformational energetics obtained from a B3LYP/6-31G* geometry optimization followed by an LMP2/cc-pVTZ(-f) single point energy calculation. The training set used to derive torsional parameters consists of $\sim$660 compounds that are a combination of molecules from papers[41,45] and a small number of additional compounds.

We also examine the effect of using a semiempirical charge assignment method, called CM1A-BCC, to explore whether predictions can be improved for classes of molecules that present the greatest challenge for the OPLS_2005 force field. The CM1A-BCC charges are obtained from a combination of Cramer−Truhlar's CM1A charge model[25] and fit bond charge correction terms (BCC)[46] for each atom. The CM1A charge model has been shown to produce accurate solvation free energies,[47] specifically when used in combination with other force field terms from OPLS-AA.[48]

Methods

The calculations presented in this work were performed using the OPLS_2005 all-atom force field with explicit solvent and were run with the default parameters in the Maestro v8.5 interface to Desmond, version 20108.[26] Comparisons to AM1-BCC/GAFF and CHelpG/CHARMm-MSI were made based on previous calculations from Shivakumar et al.[18] The starting 239 neutral small molecules were obtained from Shivakumar et al. and were solvated in an orthorhombic water box using a 10 Å buffer with no ions. All the simulations were run with the SPC water model; however, a subset of simulations was also run with the TIP3P and TIP4P water models for comparison. The solvated structures were minimized for 10 steps with the steepest-descent method followed by a maximum of 1990 steps with the limited memory Broyden−Fletcher−Goldfarb−Shanno (LBFGS) method.[49] The minimized system was further relaxed with a 24 ps molecular dynamics simulation at 300 K temperature and 1 atm pressure (corresponding to the "quick relaxation" protocol in the Schrödinger interface to Desmond). Finally, production simulations were run for 600 ps for each λ window using the same reference temperature and pressure, which were maintained by chained Nose−Hoover thermostats[50] and by a Martyna−Tobias−Klein (MTK) barostat.[51] The last 450 ps of each window were used for analysis. All simulations used a 10 Å cutoff radius for both vdW and electrostatic interactions along with smooth particle mesh Ewald[32] to calculate long-range electrostatic interactions. The convergence of the final result was checked on a subset of compounds by plotting the variation in RMSE as a function of simulation time (Supporting Information, Figure S1).

Table 1. λ Schedule[a]

window index	λ_{vdW}	λ_{coul}
1	0.0	0.0
2	0.106974	0.0
3	0.1745536	0.0
4	0.2252634	0.0
5	0.2816288	0.0
6	0.366175	0.0
7	0.5014272	0.0
8	0.7099106	0.0
9	1.0	0.25
10	1.0	0.5
11	1.0	0.75
12	1.0	1.0

[a] The λ_{vdW} and λ_{coul} refer to the scaling the vdW and coulombic interactions, respectively, at each window.

Table 2. Functional Group Compound Classification and Number of Compounds in Each Class

type	number	type	number
alkanes	8	esters	15
alkenes	10	ethers	11
alkynes	5	halogen, bromo	10
alcohols	17	halogen, chloroalkanes	11
aldehydes	6	halogen, chloroalkenes	5
aliphatic amines	16	halogen, chloroarenes	3
amides	5	halogen, flouro	6
arenes	14	halogen, iodo	8
aromatic amines	14	thiols	4
bifunctional amine	3	ketones	12
bifunctional groups	5	multiple halogens	15
branched alkanes	7	nitriles	5
carboxylic acids	5	nitro	7
cycloalkanes	5	sulfides	5
disulfides	2	total	239

λ Schedule. Since free energy is a state function, it is independent of the path taken for the transformation going from state A to state B. However, in practice, the choice of λ schedule determines the precision of the calculated free energy change, as well as the stability of the simulations. The λ schedule used for calculating the absolute solvation free energy in this work has 12 windows (see Table 1) and was devised such that the vdW terms are scaled to full strength before the charging of the atoms starts. This prevents atoms of opposite charge from approaching each other closely enough to fall into the attractive singularity in the electrostatic potential. The statistical mechanics of the separation of vdW and electrostatic terms has been discussed in detail in litature by Deng and Roux.[52] The λ_{vdW} values given in the table were empirically determined by minimizing the variances in free energy predictions for a small set of training molecules (unpublished results). The linear five-step schedule (from 0 to 1) used for λ_{coul} performed well on the small set of training molecules, although the results were fairly insensitive to the details of this schedule (see Results and Discussions Section).

Test Set Selection. We used the same set of 239 small neutral organic molecules as described in Shivakumar et al.[18,47,53] This allowed for a direct comparison of the results obtained here with other force fields and simulation programs. The test set spans diverse chemical functional groups commonly encountered in drug design (Table 2). These include saturated and unsaturated hydrocarbons, strained rings, conjugated systems, aromatic and heterocyclic rings, and many polar functional groups. All calculations were performed with molecules in their neutralized states, including carboxylic acids and amines. The two-dimensional (2D)

Table 3. Comparison of SPC, TIP3P, and TIP4P Water Models[a]

entry ID	title	exp	SPC	TIP3P	TIP4P
13-Compound Subset					
1	1-3-dioxolane	−4.1	−2.50 ± 0.21	−2.83 ± 0.19	−2.51 ± 0.16
2	4-methyl-1H-imidazole	−10.25	−8.50 ± 0.21	−8.30 ± 0.19	−8.24 ± 0.21
3	biphenyl	−2.64	−1.40 ± 0.24	−1.52 ± 0.21	−0.95 ± 0.24
4	butanal	−3.18	−1.80 ± 0.18	−1.80 ± 0.18	−1.48 ± 0.17
5	cyclopropane	0.75	2.18 ± 0.17	2.04 ± 0.15	2.50 ± 0.11
6	diethyldisulfide	−1.54	−1.28 ± 0.18	−1.29 ± 0.25	−1.00 ± 0.18
7	morpholine	−7.17	−5.54 ± 0.25	−5.22 ± 0.22	−5.12 ± 0.22
8	nitrobenzene	−4.12	−2.61 ± 0.19	−2.72 ± 0.19	−2.03 ± 0.19
9	piperazine	−7.4	−7.63 ± 0.21	−7.41 ± 0.22	−8.35 ± 0.26
10	priopionic	−6.47	−5.31 ± 0.19	−5.57 ± 0.19	−4.84 ± 0.21
11	propane-2-methoxy-2-methyl	−0.79	−0.44 ± 0.21	−0.43 ± 0.15	−0.39 ± 0.19
12	propiononitrile	−3.85	−3.39 ± 0.17	−3.33 ± 0.16	−3.18 ± 0.19
13	trimethylamine	−3.42	−2.35 ± 0.17	−1.74 ± 0.17	−1.50 ± 0.16
	average unsigned error (AUE)		1.08	1.08	1.46
	R^2		0.96	0.96	0.92
	slope		0.96	0.94	0.99
	intercept		0.89	0.82	1.25
Amino Acid Side Chain Analogues					
1	acetamide (Asn)	−9.71	−8.47 ± 0.20	−8.51 ± 0.23	−8.30 ± 0.22
2	phenol (Tyr)	−3.74	−4.64 ± 0.21	−5.40 ± 0.22	−4.07 ± 0.20
3	acetic acid (AspH)	−4.88	−5.44 ± 0.17	−5.53 ± 0.16	−4.80 ± 0.17
4	4-methyl-1H-imidazole (His)	−10.25	−8.50 ± 0.24	−8.44 ± 0.2	−8.25 ± 0.24
5	toluene (Phe)	−0.89	−0.74 ± 0.17	−0.79 ± 0.21	−0.24 ± 0.17
	average unsigned error (AUE)		0.92	1.08	0.89
	R^2		0.95	0.91	0.96
	slope		0.78	1.21	1.17
	intercept		0.96	1.06	0.13

[a] Energies are reported in kcal/mol. The experimental (exp) numbers are shown for comparison.

Table 4. Absolute Solvation Free Energies using 5- and 11-λ Coupling Parameter for Electrostatics[a]

entry ID	title	exp	$\lambda^{5}_{\text{coul}}$	$\lambda^{11}_{\text{coul}}$
1	1-3-dioxolane	−4.10	−2.50 ± 0.21	−2.58 ± 0.17
2	4-methyl-1H-imidazole	−10.25	−8.50 ± 0.21	−8.72 ± 0.19
3	biphenyl	−2.64	−1.40 ± 0.24	−1.48 ± 0.25
4	butanal	−3.18	−1.80 ± 0.18	−1.83 ± 0.18
5	cyclopropane	0.75	2.18 ± 0.17	2.00 ± 0.11
6	diethyldisulfide	−1.54	−1.28 ± 0.18	−1.23 ± 0.19
7	morpholine	−7.17	−5.54 ± 0.25	−5.28 ± 0.20
8	nitrobenzene	−4.12	−2.61 ± 0.19	−2.49 ± 0.14
9	piperazine	−7.40	−7.63 ± 0.21	−7.97 ± 0.19
10	priopionic	−6.47	−5.31 ± 0.19	−5.19 ± 0.17
11	propane-2-methoxy-2-methyl	−0.79	−0.44 ± 0.21	−0.43 ± 0.20
12	propiononitrile	−3.85	−3.39 ± 0.17	−3.25 ± 0.15
13	trimethylamine	−3.42	−2.35 ± 0.17	−1.67 ± 0.18
	average unsigned error (AUE)		1.08	1.17
	R^2		0.96	0.95
	slope		0.96	0.97
	intercept		0.89	0.96

[a] The 5-λ schedule corresponds to the default coupling scheme. All the energies are reported in kcal/mol.

chemical structures for all molecules within each of the groups are available from the Supporting Information (Figure S2).

Subset of Molecules for Solvent Model and Charge Assignment Comparison. A subset of 13 molecules was initially investigated to study effects of the solvent model and the charge assignment method. This subset was chosen by applying hierarchical clustering to the Tanimoto similarity matrix[54] derived from linear 2D molecular fingerprints, as implemented in Canvas.[26] From the initial 239 compounds in Table 2, one compound (the closest member to the cluster centroid) from each of 13 clusters was chosen (Table 3). Three different solvent models were compared here—SPC,[55] TIP3P,[56] and TIP4P.[56] The SPC and TIP3P are commonly used three-site water models that differ in potential terms and geometry. TIP4P is a commonly used 4-site water model that places negative charge on a dummy atom near the oxygen along the bisector of the H−O−H angle.

Results and Discussions

Solvent Models. The absolute solvation free energies using the SPC, TIP3P, and TIP4P water models for the 13 molecule subset (see Methods Section) are shown in Table 3. For completeness, the absolute solvation free energies were also calculated for a set of five molecules that represents neutral amino acid side chain analogues Glu, Tyr, Asp, His, and Phe (Table 3). The statistical uncertainty of the computed

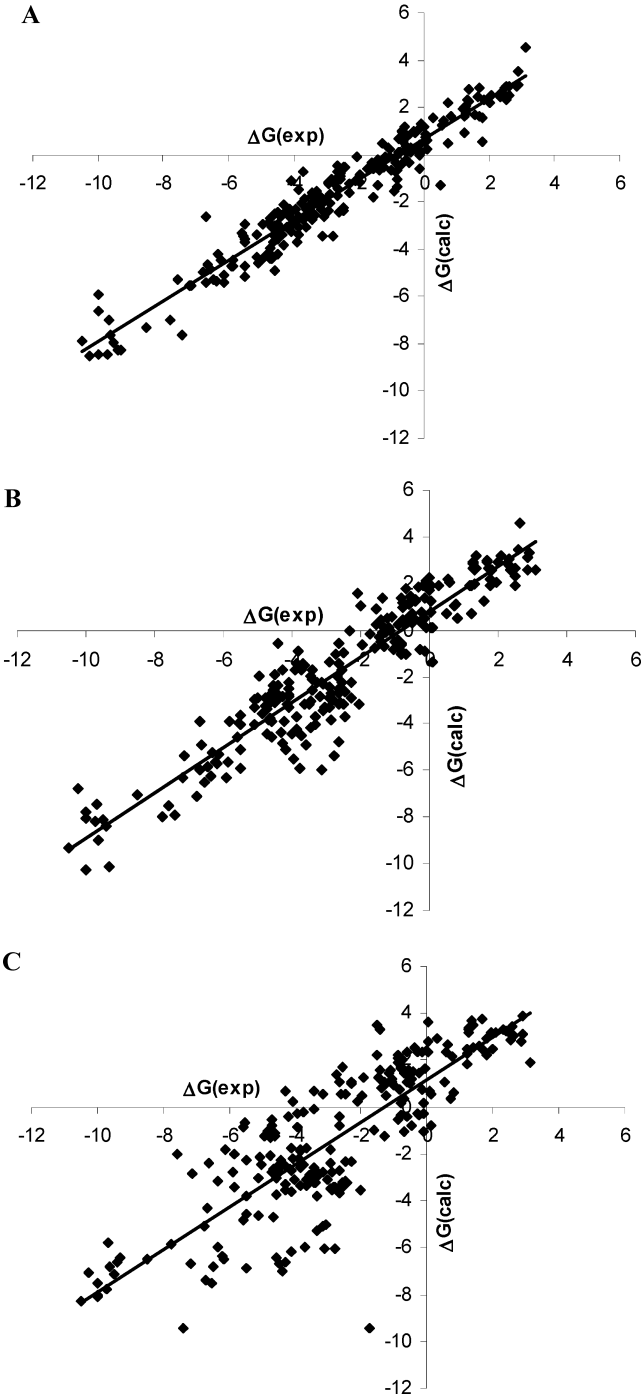


Figure 2. Plot of predicted versus experimental solvation free energy (in kcal/mol) using different force fields. (A) OPLS_2005, (B) AM1-BCC charges with GAFF vdW and bonded parameters (AM1-BCC/GAFF), and (C) CHelpG charges with CHARMm-MSI vdW and bonded parameters (CHelpG/CHARMm-MSI).

Table 5. Absolute Solvation Free Energies using a Modified λ Coupling Parameter for vdW[a]

entry ID	title	exp	$\lambda^{9}_{\mathrm{vdW}}$	$\Lambda^{13}_{\mathrm{vdW}}$
1	anthracene	−4.23	−3.06 ± 0.23	−2.87 ± 0.25
2	biphenyl	−2.64	−1.40 ± 0.24	−1.52 ± 0.25
3	nitrobenzene	−4.12	−2.61 ± 0.19	−2.30 ± 0.16

[a] The 9-λ schedule corresponds to the default coupling scheme. All the energies are reported in kcal/mol.

free energy is estimated using the subsampling bootstrap method as described in the Theory Section. OPLS_2005 parameters and charges were used for all molecules. The calculations on this subset of molecules were performed in order to find a good solvent model for further analysis on the full data set and were not intended as a complete study.

The correlation coefficient (R^2) between the experimental and the predicted solvation free energies using SPC, TIP3P, and TIP4P water models is 0.96, 0.96, and 0.92, respectively, for the small molecules reported in Table 3. In general, the solvent models perform comparably across this subset of molecules, and the same compounds present a challenge for all of the water models. These results suggest that there is not an extreme sensitivity to the choice of water model, at least for the current subset, representing small molecules. For amino acid side chain analogues (Table 3), the R^2 between the experimental and the predicted solvation free energy using SPC, TIP3P, and TIP4P water models is 0.95, 0.91, and 0.96, respectively. For the amino acid analogues, the TIP4P solvent model shows the lowest average unsigned error (AUE). This is in accordance with the previously published work where it has been reported that the hydration free energies of various small molecules in explicit solvent could be water-model dependent.[15] However, we chose to use the SPC model for the remaining calculations in this work because it has the lowest AUE, the best R^2, a good slope, and the smallest maximum error for the 13 small molecule subset and has similarly good results for the amino acid side chain analogues. Furthermore, simulations run faster with a three-point water model, like SPC, as compared to a four-point model, like TIP4P.

Effect of Additional λ Windows. The alchemical transformations presented in this work use a 5-λ schedule with linear scaling ($\lambda^{5}_{\mathrm{coul}} = 0.0, 0.25, 0.5, 0.75$, and 1) for the electrostatic component of the solvation free energy. To our knowledge, there has been no direct study reported in the literature to optimize the electrostatic λ schedule. Work by Deng and Roux[52] uses an 11-λ schedule ($\lambda^{11}_{\mathrm{coul}} = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$,and 1) to linearly scale the electrostatics, whereas Mobley et al[15] uses a 5-λ schedule. A reduced number of λ windows is desirable to speed the calculations but has the potential to produce inaccurate results.

The results for the 13-compound subset using the 5- and 11-λ schedules are shown in Table 4. The solvation free energies are equivalent within the errors of the computation, suggesting that the reduced $\lambda^{5}_{\mathrm{coul}}$ schedule is sufficient for the calculation of absolute solvation free energies of typical small organic molecules. The correlation with experiment (R^2) is 0.95 and 0.96 for $\lambda^{5}_{\mathrm{coul}}$ and $\lambda^{11}_{\mathrm{coul}}$ schedules, respectively (Table 4).

In the current alchemical transformation protocol, the vdW component of the solvation free energy is staged using a 9-λ coupling parameters ($\lambda^{9}_{\mathrm{vdW}} = 0.0, 0.1069742, 0.1745536, 0.2252634, 0.2816288, 0.366175, 0.5014272, 0.7099106$, and 1.0). As a test, we added four additional λ windows near $\lambda = 0.0$, since this is believed to be the most critical part of the λ schedule due to the effective appearance of new excluded volume regions. The resulting 13-λ schedule had

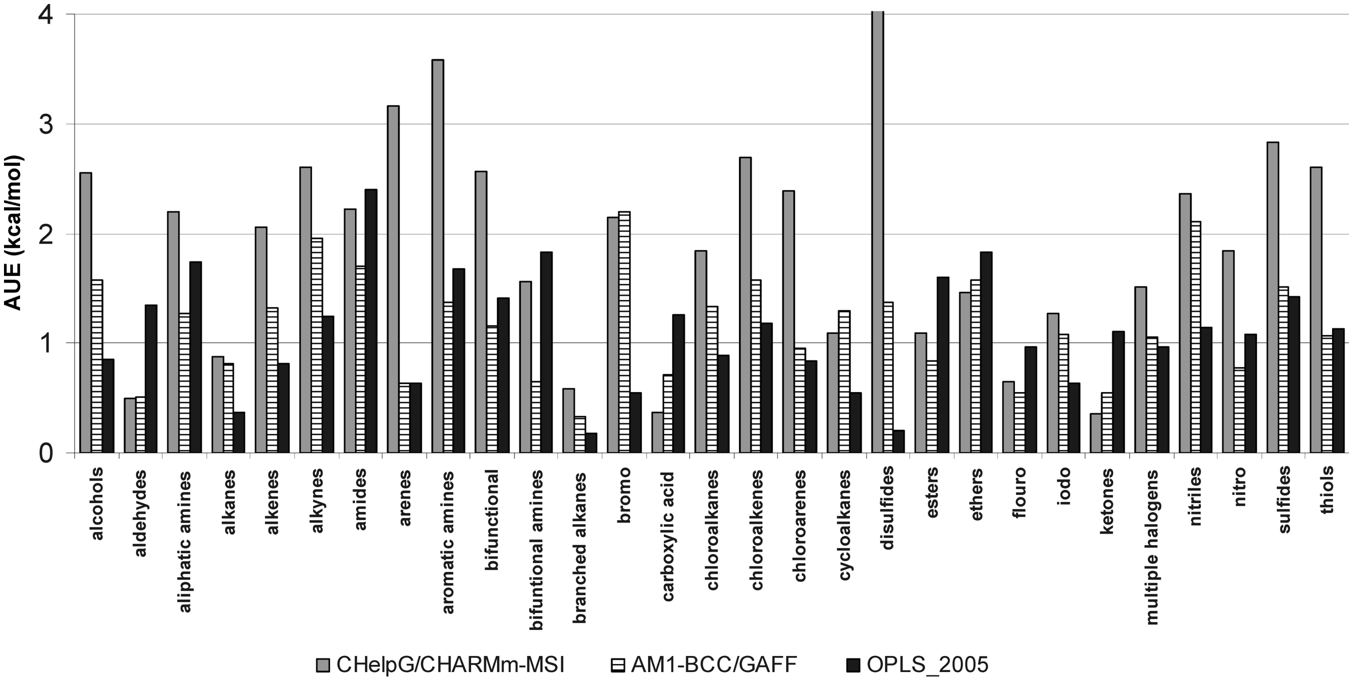


Figure 3. Comparison of the average unsigned error (AUE) for different classes of compounds using OPLS_2005, AM1-BCC/GAFF, and CHelpG/CHARMm-MSI.

λ values of $\lambda^{13}_{\mathrm{vdW}} = 0.0$, 0.01, 0.02, 0.05, 0.07, 0.1069742, 0.1745536, 0.2252634, 0.2816288, 0.366175, 0.5014272, 0.7099106, and 1.0. We ran the new λ schedule on three of the bulkier molecules in order to stress the difference in the vdW treatment, which should be more significant with larger molecules. As seen in Table 5, the results from the default ($\lambda^{9}_{\mathrm{vdW}}$) and modified ($\lambda^{11}_{\mathrm{vdW}}$) λ schedules are essential equivalent within the statistical error range.

Comparison of the Force Fields. Two different charge models were used for calculating the absolute solvation free energies on the 13 compound subset—default OPLS_2005 charges with the OPLS_2005 force field (OPLS_2005/OPLS_2005) and AM1-BCC charges with the OPLS_2005 force field (AM1-BCC/OPLS_2005). The OPLS_2005 charges were assigned using the default Schrödinger atom-typing infrastructure.[26] The AM1-BCC charges were taken from Shivakumar et al.[18] AM1-BCC was chosen for comparison because it outperformed the other charge models discussed in the work by Shivakumar et al. Finally, the best combined charge assignment method and force field from Shivakumar et al. (AM1-BCC/GAFF) was also added to the comparison. The results are shown in Table 6. The OPLS_2005/OPLS_2005 combination performs better than either AM1-BCC/GAFF or AM1-BCC/OPLS_2005, although the latter is very close. The correlation coefficient R^2 is 0.96 for OPLS_2005/OPLS_2005, whereas it is 0.95 for AM1-BCC/OPLS_2005 and 0.87 for AM1-BCC/GAFF. However, the AM1-BCC/OPLS_2005 has the lowest AUE and best slope among the three. This shows that the AM1-BCC charge model, which has been often used in conjunction with GAFF, also performs equally well with the OPLS_2005 force field. The major outlier in AM1-BCC/OPLS_2005 and AM1-BCC/

Table 6. Comparison of the Force Fields and Charge Assignment Methods (OPLS_2005/OPLS_2005, AM1-BCC/OPLS_2005, and AM1-BCC/GAFF)[a]

entry ID	title	exp	OPLS_2005/OPLS_2005	AM1-BCC/OPLS_2005	AM1-BCC/GAFF[b]
1	1,3-dioxolane	−4.10	−2.50 ± 0.21	−3.98. ± 0.18	−2.81
2	4-methyl-1H-imidazole	−10.25	−8.50 ± 0.21	−8.37 ± 0.20	−6.80
3	biphenyl	−2.64	−1.40 ± 0.24	−2.63 ± 0.21	−2.68
4	butanal	−3.18	−1.80 ± 0.18	−3.51 ± 0.18	−2.85
5	cyclopropane	0.75	2.18 ± 0.17	2.20 ± 0.11	1.10
6	dimethyldisulfide	−1.54	−1.28 ± 0.18	−0.58 ± 0.24	0.23
7	morpholine	−7.17	−5.54 ± 0.25	−5.47 ± 0.17	−6.34
8	nitrobenzene	−4.12	−2.61 ± 0.19	−4.37 ± 0.17	−3.18
9	piperazine	−7.40	−7.63 ± 0.21	−7.17 ± 0.20	−7.94
10	propionic	−6.47	−5.31 ± 0.19	−6.96 ± 0.19	−5.84
11	propane,2-methoxy-2-methyl	−0.79	−0.44 ± 0.21	−0.98 ± 0.25	−0.11
12	propiononitrile	−3.85	−3.39 ± 0.17	−0.91 ± 0.18	−1.43
13	trimethylamine	−3.42	−2.35 ± 0.17	−2.25 ± 0.17	−1.71
	average unsigned error (AUE)		1.08	0.90	1.15
	R^2		0.96	0.94	0.87
	slope		0.96	0.95	0.89
	intercept		0.89	0.49	0.60

[a] The SPC water model was used for all simulations. Energies are reported in kcal/mol. [b] Obtained from Shivakumar et al.[18]

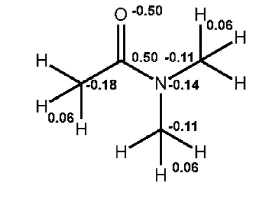
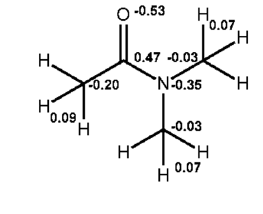
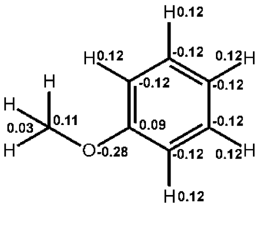
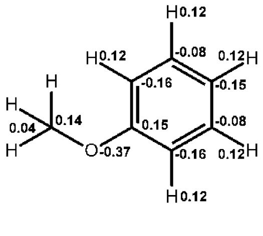
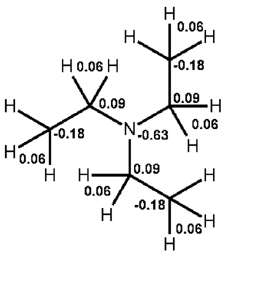
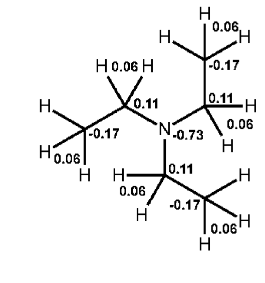
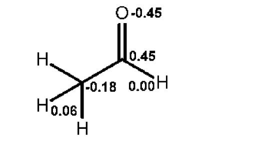
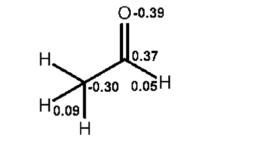
Molecule name	Experimental Solvation Free Energies	OPLS 2005 Charges and Solvation Free Energies	CM1A-BCC/OPLS_2005 Charges and Solvation Free Energie
N,N-dimethylacetamide	-8.50 kcal/mol	 O -0.50 H 0.06 H H 0.50 -0.11 -0.18 N -0.14 H H 0.06 H -0.11 H H H 0.06	 O -0.53 H 0.07 H H 0.47 -0.03 -0.20 N -0.35 H H 0.09 H -0.03 H H H 0.07
		-6.16 ± 0.19 kcal/mol	-7.31 ± 0.21 kcal/mol
methoxybenzene	-3.73 kcal/mol	 H 0.12 H 0.12 -0.12 0.12 H H -0.12 -0.12 H 0.03 0.11 0.09 -0.12 H O -0.28 -0.12 0.12 H H 0.12	 H 0.12 H 0.12 -0.08 0.12 H H -0.16 -0.15 H 0.04 0.14 0.15 -0.08 H O -0.37 -0.16 0.12 H H 0.12
		-0.72 ± 0.22 kcal/mol	-3.86 ± 0.16 kcal/mol
N,N-diethylethanamine	-4.07 kcal/mol	 H 0.06 H H -0.18 H 0.06 H 0.09 0.09 H H -0.18 N -0.63 0.06 H H 0.06 H H 0.09 H 0.06 -0.18 H H 0.06 H H	 H 0.06 H H -0.17 H 0.06 H 0.11 0.11 H H -0.17 N -0.73 0.06 H H 0.06 H H 0.11 H 0.06 -0.17 H H 0.06 H H
		-1.07 ± 0.19 kcal/mol	-4.42 ± 0.22 kcal/mol
acetaldehyde	-3.5 kcal/mol		
		-2.10 ± 0.17 kcal/mol	-3.55 ± 0.12 kcal/mol

Figure 4. Partial atomic charges and solvation free energies in kcal/mol for a selection of polar compounds (amide, ether, amine, and aldehyde) using OPLS_2005 and CM1A-BCC/OPLS_2005.

GAFF is propiononitrile, which performs much better using OPLS_2005/OPLS_2005. This suggests further improvement in the AM1-BCC charges for nitrile functional groups. A similar observation was also made in the published work of Shivakumar et al.

Absolute Solvation Free Energies of the Full Test Set

Finally, we calculate the absolute solvation free energy for the entire set of 239 molecules using the SPC water model and the OPLS_2005 force field. The AUE for OPLS_2005 is 1.10, which is consistent with the results from the 13-compound diverse subset shown in Table 6. The AUE as reported by Shivakumar et al. was 1.17 for AM1-BCC/GAFF and 1.88 for CHelpG/CHARMm-MSI. Figure 2A shows the correlation between the predicted and experimental absolute solvation free energies, with a coefficient (R^2) of 0.94, a slope of 0.86, and an intercept of 0.68 kcal/mol. The correlation is significantly better than that for AM1-BCC/GAFF (Figure 2B) or CHelpG/CHARMm-MSI (Figure 2C) on the same set of molecules. The R^2, slope, and intercept for AM1-BCC/GAFF were 0.87, 0.97, and 0.78 and for CHelpG/CHARMm-MSI were 0.72, 0.91, and 1.19, respectively.

To further compare the performance of different charge models and force fields, the compounds were analyzed based on the chemical functional group classes shown in Table 2. The AUE in the absolute solvation free energy for each class is shown in Figure 3. The OPLS_2005 force field produces

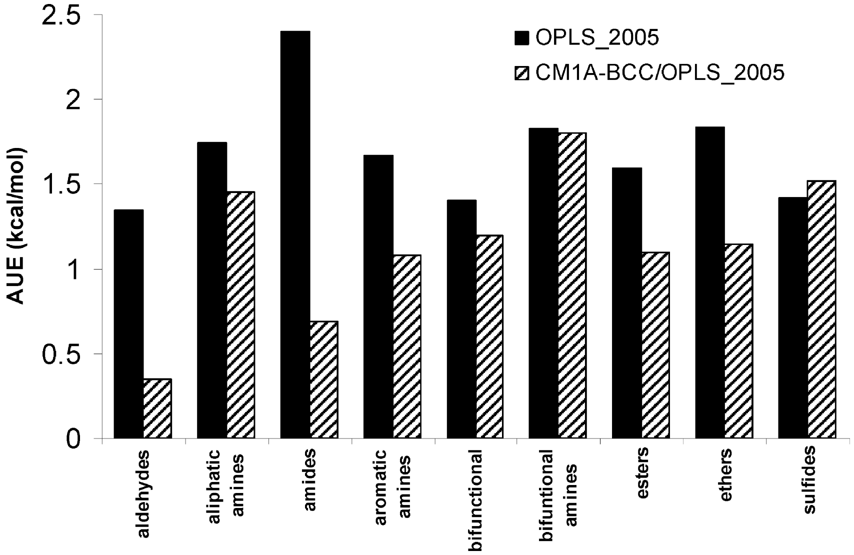


Figure 5. AUE using the OPLS_2005 and CM1A-BCC/OPLS_2005 force fields for the functional group classes that perform poorly with the OPLS_2005 force field.

the lowest AUE for a number of important functional groups, such as alcohols, alkanes, alkenes, cycloalkanes, and disulfides. Both OPLS_2005 and AM1-BCC/GAFF perform similarly for other hydrocarbon groups (branched alkanes, alkynes, and cycloalkanes), polar groups (aldehydes, ketones, carboxylic acids, esters, ethers, aliphatic amines, nitro, and sulfides), and the majority of the halogenated molecules. The worst performing classes of compounds with OPLS_2005 contain polar functional groups, such as amides, amines, esters, and ethers. CHelpG/CHARMm-MSI did not perform well on many systems, with 14 classes having an AUE of greater than 2.0 kcal/mol and 3 classes (arenes, amines, and disulfides) having an AUE greater than 3.0 kcal/mol.

OPLS_2005 with CM1A-BCC Charges. Overall, the absolute solvation free energies obtained using the OPLS_2005 force field show the highest correlation with the experimental solvation free energies compared to other two force fields mentioned in this study. However, there are classes of compounds with large errors that need improvement even with OPLS_2005. For example, the nitrogen-containing polar functional groups, such as amines and amides, did not perform well with OPLS_2005. Polar functional groups like these have previously posed challenges in predicting solvation free energies with fixed charge force fields.[57] Interestingly, these functional groups performed better using the semiempirical AM1-BCC charges in the work by Shivakumar et al.,[18] suggesting that an improved charge assignment model could lead to better OPLS_2005 predictions.

We examined the most challenging compound classes using the CM1A-BCC charge assignment method with the OPLS_2005 force field. The partial atomic charges for a subset of molecules from OPLS_2005 and CM1A-BCC are shown in Figure 4 to illustrate the subtle differences in atomic charges and how they impact the solvation free energies. For example, the experimental solvation free energy for *N*,*N*-dimethylacetamide is −8.5 kcal/mol. Using the default charges from the OPLS_2005 force field, the predicted absolute solvation free energy is −6.16 ± 0.19 kcal/mol, whereas using CM1A-BCC charges, we see an improvement in the predicted absolute solvation free energy to −7.31 ± 0.21 kcal/mol that results from a series of small charge differences spread across the molecule. The absolute solvation free energies using OPLS_2005 are systematically less negative than the experimental values for these polar molecules, whereas the CM1A-BCC charges result in a more favorable predicted solvation free energy. Significant improvements are also observed with other functional classes, such as ethers, amines, and aldehydes. A representative molecule from each of these classes is shown in Figure 4. The Figure 5 shows the AUE in the solvation free energy using CM1A-BCC charges for the compound classes that perform poorly with the default OPLS_2005 force field charge assignment. The results for CM1A-BCC charges are better in all cases, except sulfides where they are marginally worse. There are significant improvements for aldehydes, amides, aromatic amines, esters, and ethers functional groups.

Conclusions

In this study we computed the absolute solvation free energies for a diverse set of 239 neutral molecules using molecular dynamics/free energy perturbation (MD/FEP) with explicit solvent and compared the results with experimental data. The OPLS_2005 all-atom force field performed well compared with two other commonly used small molecule force fields (AM1-BCC/GAFF and CHelpG/CHARMm-MSI) that have been previously reported[18] for the same set of small molecules. While there was good correlation with experimental values and a low average unsigned error across the entire set with the OPLS_2005 force field, there were important classes of polar compounds that performed suboptimally. It was found that for most polar classes of compounds the results improved when the newly developed CM1A-BCC charge assignment methodology was used in conjunction with the rest of the OPLS_2005 force field parameters. Further work is needed to explore the effects across all classes of compounds, including nonpolar functional groups. While the CM1A-BCC charges described in this work performed well when combined with the other OPLS_2005 force field parameters, a self-consistent force field will likely perform better. We are in the process of completing the development of the next generation OPLS-AA force field, which uses the CM1A-BCC charge assignment method described in this work and also greatly expands on the parameter space coverage of small molecules compared to the OPLS_2005 force field. Validation of the complete force field for solvation free energy calculations will be reported in a future work.

The accurate characterization of solvation effects is critical in the thermodynamic process of protein−ligand binding. The prediction of solvation free energies provides a surrogate for the biologically relevant process of transferring a small molecule from solution (high-dielectric environment) to the binding site of a protein (low-dielectric region) and, therefore, is an important step toward predicting accurate binding free energies. Further work on improving force field accuracy and enhanced sampling simulation methods could help to bring the accuracy level of binding free energy predictions to the point where they can provide substantial value in the drug discovery arena.

Acknowledgment. The authors thank Drs. Huafeng Xu, Byungchan Kim, Teng Lin, and David Rinaldo for helpful discussions and suggestions and Drs. Ramy Farid and Ken Dyall for providing critical comments on the manuscript. Also, the authors thank Profs. Benoit Roux and Bill Jorgensen for helpful discussions concerning the FEP methodology.

Supporting Information Available: Variation in RMSE as a function of simulation time (Figure S1) and the 2D chemical structures for all molecules used in this study along with their experimental and calculated solvation free energies (Figure S2) is available. The coordinates of the individual molecules can be requested from the authors or downloaded from http://www.schrodinger.com.This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Jorgensen, W. L. *Science* **2004**, *303*, 1813–1818.

(2) Karplus, M. *Acc. Chem. Res.* **2002**, *35*, 321–323.

(3) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 4–16.

(4) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199–203.

(5) Kirkwood, J. G. *J. Chem. Phys.* **1935**, 300–313.

(6) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420–1426.

(7) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(8) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161–2200.

(9) Sharp, K. A.; Honig, B. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301–332.

(10) Nina, M.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 5239–5248.

(11) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.

(12) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395–2417.

(13) Jorgensen, W. L. *Acc. Chem. Res.* **1989**, *22*, 184–189.

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

(15) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2007**, *111*, 2242–2254.

(16) Mobley, D. L.; Dill, K. A.; Chodera, J. D. *J. Phys. Chem. B* **2008**, *112*, 938–946.

(17) 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*, 769–779.

(18) Shivakumar, D.; Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 919–930.

(19) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

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

(21) Momany, F. A.; Rone, R. *J. Comput. Chem.* **1992**, *13*, 888–900.

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

(23) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361–373.

(24) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *40*, 10269–10280.

(25) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87–110.

(26) *Maestro*, version 8.5; Schrodinger, Inc.: New York, NY, 2008.

(27) *Desmond Molecular Dynamics System*; D. E. Shaw Research: New York, NY, 2008.

(28) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregerson, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. *Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters*; Proceedings of theACM/IEEE Conference on Supercomputing(SC06)Tampa, FL2006.

(29) Shaw, D. E. *J. Comput. Chem.* **2005**, *26*, 1318–1328.

(30) Bowers, K. J.; Dror, R. O.; Shaw, D. E. *J. Chem. Phys.* **2006**, *124*, 184109–184111.

(31) Lippert, R. A.; Bowers, K. J.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Shaw, D. E. *J. Chem. Phys.* **2007**, *126*, 046101−046102.

(32) Essmann, U.; Perera, L.; Berkowitz, M.; Darden, T.; Lee, H.; Pedersen, L. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(33) Beutler, T. C.; Mark, A. E.; Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529–643.

(34) Lague, P.; Pastor, R. W.; Brooks, B. R. *J. Phys. Chem. B* **2003**, *108*, 363–368.

(35) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245–268.

(36) Chernick, M. R. *Bootstrap Methods: A Guide for Practitioners and Researchers*, 2nd ed.; Wiley: Hoboken, NJ, 2007.

(37) Banks, J. L.; Beard, H. S.; Cao, Y.; Cho, A. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; Maple, J. R.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Zhang, L. Y.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2005**, *26*, 1752–1780.

(38) Jacobson, M. P.; Kaminski, G. A.; Rapp, C. A.; Friesner, R. A. *J. Phys. Chem. B* **2002**, *106*, 11673–11680.

(39) Jensen, K. P.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2006**, *2*, 1499–1509.

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

(41) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(42) Damm, W.; Frontera, A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Comput. Chem.* **1997**, *18*, 1955–1970.

(43) Jorgensen, W. L.; McDonald, N. A. *J. Phys. Chem. B* **1998**, *102*, 8049–8059.

(44) Rizzo, R. C.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1999**, *121*, 4827–4836.

(45) Price, M. L. P.; Ostrovsky, D.; Jorgensen, W. L. *J. Comput. Chem.* **2001**, *22*, 1340–1352.

(46) Jakalian, A.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

(47) Chambers, C. C.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 16385–16398.

(48) Udier-Blagovic, M.; De Tirado, P. M.; Pearlman, S. A.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 1322–1332.

(49) Schlick, T. *Molecular modeling and simulation: an interdisciplinary guide*; Springer: New York, NY, 2002.

(50) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635–2643.

(51) Martyna, G. J.; Tobias, D. J.; Klein, M. L. *J. Chem. Phys.* **1994**, *101*, 4177–4189.

(52) Deng, Y.; Roux, B. *J. Phys. Chem. B* **2004**, *108*, 16567–16576.

(53) Maple, J. R.; Cao, Y.; Damm, W.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A. *J. Chem. Theory Comput.* **2005**, *1*, 694–715.

(54) Tanimoto, T. T. *IBM Internal Report*; IBM: Armonk, NY1957.

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

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

(57) Morgantini, P.-Y.; Kollman, P. A. *J. Am. Chem. Soc.* **2002**, *117*, 6057–6063.

CT900587B

JCTC Journal of Chemical Theory and Computation

Reparameterization of RNA χ Torsion Parameters for the AMBER Force Field and Comparison to NMR Spectra for Cytidine and Uridine

Ilyas Yildirim,[†] Harry A. Stern,[†] Scott D. Kennedy,[‡] Jason D. Tubbs,[†] and Douglas H. Turner[*,†,§]

Department of Chemistry and Center for RNA Biology, University of Rochester, Rochester, New York 14627, and Department of Biochemistry and Biophysics, School of Medicine and Dentistry, University of Rochester, Rochester, New York 14642

Received November 17, 2009

Abstract: A reparameterization of the torsional parameters for the glycosidic dihedral angle, χ, for the AMBER99 force field in RNA nucleosides is used to provide a modified force field, AMBER99χ. Molecular dynamics simulations of cytidine, uridine, adenosine, and guanosine in aqueous solution using the AMBER99 and AMBER99χ force fields are compared with NMR results. For each nucleoside and force field, 10 individual molecular dynamics simulations of 30 ns each were run. For cytidine with AMBER99χ force field, each molecular dynamics simulation time was extended to 120 ns for convergence purposes. Nuclear magnetic resonance (NMR) spectroscopy, including one-dimensional (1D) ^{1}H, steady-state 1D ^{1}H nuclear Overhauser effect (NOE), and transient 1D ^{1}H NOE, was used to determine the sugar puckering and preferred base orientation with respect to the ribose of cytidine and uridine. The AMBER99 force field overestimates the population of syn conformations of the base orientation and of C2′-endo sugar puckering of the pyrimidines, while the AMBER99χ force field's predictions are more consistent with NMR results. Moreover, the AMBER99 force field prefers high anti conformations with glycosidic dihedral angles around 310° for the base orientation of purines. The AMBER99χ force field prefers anti conformations around 185°, which is more consistent with the quantum mechanical calculations and known 3D structures of folded ribonucleic acids (RNAs). Evidently, the AMBER99χ force field predicts the structural characteristics of ribonucleosides better than the AMBER99 force field and should improve structural and thermodynamic predictions of RNA structures.

1. Introduction

Understanding the physical interactions governing the structure and dynamics of ribonucleosides should improve the accuracy of simulations of ribonucleic acid (RNA) molecules. Methods for simulating biological systems include residue-centered force fields (coarse-grained),[1] atom-centered force fields (AMBER,[2] CHARMM,[3,4] GROMOS),[5,6] approximate quantum mechanics,[7,8] and mixed quantum mechanics/molecular mechanics methods (QM/MM).[9−18] With advances in computer power, it is possible to run simulations at least as long as milliseconds and microseconds with coarse-grained and atom-centered potentials, respectively.[19−23] The AMBER force fields are particularly widely used for simulations of RNA. They have provided satisfactory descriptions of structural and thermodynamic properties for some RNA and deoxyribonucleic acid (DNA) systems,[24−29] while some challenging systems still provide difficulty.[30−32] Predictions for the individual ribonucleosides have not been

[*] Corresponding author. Telephone: (585) 275-3207. Fax: (585) 276-0205. E-mail: turner@chem.rochester.edu.

[†] Department of Chemistry, University of Rochester.

[‡] Department of Biochemistry and Biophysics, University of Rochester.

[§] Center for RNA Biology, University of Rochester.

10.1021/ct900604a © 2010 American Chemical Society
Published on Web 04/16/2010

extensively used as benchmarks for AMBER force fields. A fundamental understanding of nucleosides is crucial to simulate the behavior of residues in single strands, noncanonical base pairs, and hairpins. Mimicking the real behavior of ribonucleosides in simulations should improve predictions of RNA properties.

Due to limitations of computer power, small model systems were used to parametrize the glycosidic dihedral angle in the AMBER94 force field.[2] In this article, the glycosidic dihedral angle, χ, of ribonucleic acids is reparameterized by extending the quantum mechanical (QM) fitting protocol, and new parameters are used in a revised force field, AMBER99χ. Structural and thermodynamic results are extracted from molecular dynamics (MD) simulations using AMBER99[33] and AMBER99χ force fields.

Previous experimental work on nucleosides and nucleotides has classified the behavior of individual torsion angles.[34−41] Structures of modified and unmodified nucleosides/nucleotides have been interrogated by one-dimensional (1D) ^{1}H nuclear magnetic resonance (NMR) and steady-state 1D ^{1}H nuclear Overhauser effect (NOE) difference spectroscopy (SSNOE).[42−50] In this work, transient 1D ^{1}H NOE spectroscopy[51] and sugar proton coupling constants extracted from 1D ^{1}H NMR spectra for cytidine (C) and uridine (U) are used to quantitatively deduce the preferred conformations of the glycosidic dihedral angle and the sugar pucker, respectively. These results are compared to computational predictions. The AMBER99 force field overestimates the fraction of syn conformations for the base orientation and of C2′-endo sugar puckering of the pyrimidines, while the results of AMBER99χ are more consistent with that of the experimental NMR data. Simulations on adenosine (A) and guanosine (G) show that AMBER99 prefers high anti conformations around 310°, while AMBER99χ prefers anti conformations around 185°. The latter is more consistent with QM energy profiles and is the typical anti region seen in crystal structures of nucleic acids.

2. Methods

2.1. NMR. C, U, A, and G were purchased from Sigma Aldrich. Solutions of 0.2, 1, and 5 mM nucleosides were made in H_2O with an NMR buffer consisting of 80 mM NaCl, 10 mM sodium phosphate, and 0.5 mM disodium EDTA at a pH of 7.0. Two lyophilizations were performed on each sample, reconstituting each time with 99.9% D_2O (Cambridge Isotopes Laboratories). One final lyophilization was performed, and each sample was reconstituted with 99.990% D_2O (Sigma Aldrich).

NMR experiments were performed with Varian Unity Inova 500 and 600 MHz spectrometers. Chemical shift data were extracted from 1D ^{1}H NMR (see Supporting Information). For A, the chemical shifts of H8, H2, H1′, and H2′ protons vary with concentration, implying that there is base stacking and/or base pairing interactions (see Supporting Information).[52] The 5′-guanosine monophosphate is known to form quadruplex structures and other kinds of aggregates in solution,[53−55] and presumably, guanosine does the same. Aggregation and even precipitation was seen in 5 mM G solutions. Thus, the NMR spectra for nucleosides of A and G were not interpreted, except that ^3J spin−spin couplings of 0.2 mM samples were measured as a function of temperature (see Supporting Information).

For C and U, transient 1D NOE measurements were performed with a selective inversion−recovery experiment in which the frequency of the selective inversion pulse was alternated between on resonance with the H6 proton and 2000 Hz downfield, where no resonances are present. The on/off resonance spectra were subtracted, and the integral of the resulting NOE peaks was divided by peak integrals in a 1D spectrum to obtain percent enhancement. Steady-state 1D NOE spectra were acquired in a similar manner with the inversion−recovery replaced by low-power irradiation for 10 s that was on/off the H6 resonance.

2.2. Ab Initio Potential Energy Surface Scan of χ. Initial geometries were chosen to represent experimental conformations. The γ dihedral angle (O5′−C5′−C4′−C3′) was set to 54°, which is the observed γ value for A-form RNA. The δ dihedral angle (C5′−C4′−C3′−O3′) was set to either 140° or 81°, which is C2′- or C3′-endo sugar pucker, respectively. The O4′−C1′−C2′−C3′ dihedral was set to either 32° or −24° to force the sugar pucker to stay in C2′- or C3′-endo conformations, respectively. In ribonucleosides, there are three OH groups (5′, 3′, and 2′) that are free to rotate in solution. The 3′ OH group will not interact with the base as much as the 5′ and 2′ OH groups. Thus, different conformations of 5′ and 2′ OH groups were included in the fitting.

For each nucleoside (Figure 1), four different sugar conformations (Table 1) were chosen for QM calculations with Gaussian03.[56] For each sugar conformation, a potential energy surface (PES) scan was done around the glycosidic dihedral angle with increments of 5°, yielding $4 \times 72 = 288$ conformations for each nucleoside. For each conformation in the PES scan, the structures were first optimized with HF/6-31G* level of theory. During the optimization, most dihedrals were frozen in order to have a smooth energy profile with respect to the χ torsion angle (see Supporting Information). Then, QM energies, E_{QM}, were calculated with MP2/6-31G* level of theory.

2.3. Force Field Fitting of χ Torsions. The molecular mechanics (MM) energies, $E_{MM}^{(noCHI)}$, of each conformation were calculated by restraining the dihedral angles to the values of the optimized QM geometries with a force constant of 1500 kcal/mol·A^2 using the AMBER99[33] force field parameters, except χ torsion parameters were set to zero (see Supporting Information). AMBER9[57] was used to calculate the MM energies, which use the default 1−4 vdW and electrostatic screening factors of 2.0 and 1.2, respectively.

The energy difference, $E_{QM} - E_{MM}^{(noCHI)}$, represents the potential energy due to χ torsion:

$$E_{QM} - E_{MM}^{(noCHI)} = E_{CHI} \quad (1)$$

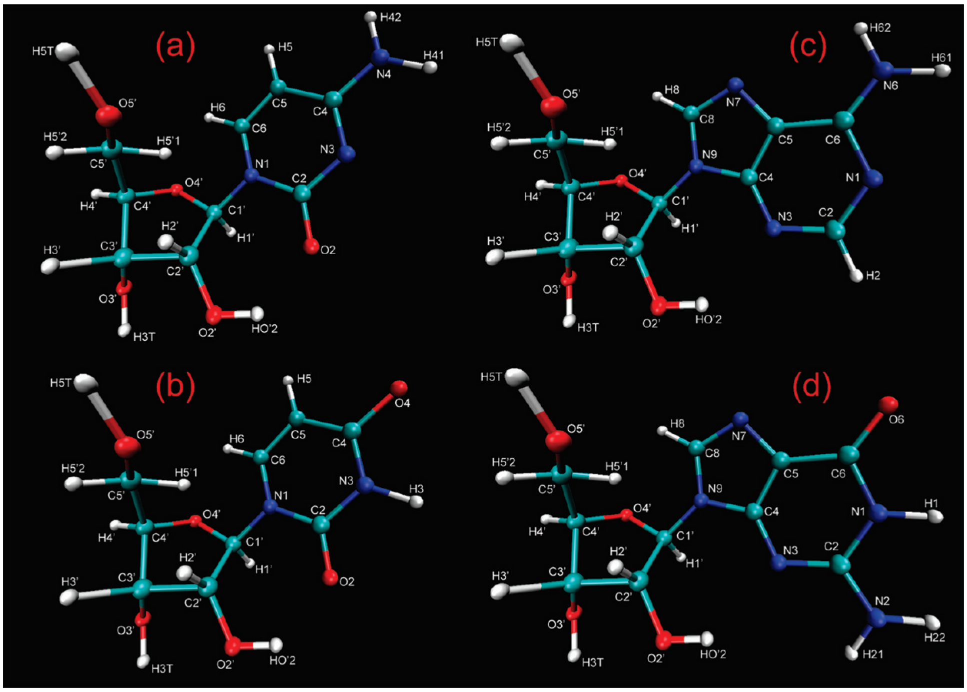


Figure 1. Atom notations of nucleosides: (a) cytidine, (b) uridine, (c) adenosine, and (d) guanosine. For C and U, χ is the dihedral angle defined by O4′−C1′−N1−C2, and for A and G, χ is defined by O4′−C1′−N9−C4. These particular structures in a−d have anti χ angles and C2′-endo sugar conformations.

Table 1. Dihedral Angles Used to Create the Four Sugar Conformations (sc) for Each Nucleoside

	C2′-endo		C3′-endo	
dihedral	sc 1	sc 2	sc 3	sc 4
H5T−O5′−C5′−C4′	60	60	174	174
O5′−C5′−C4′−C3′	54	54	54	54
C5′−C4′−C3′−O3′	140	140	81	81
C4′−C3′−O3′−H3T	−148	−148	−148	−148
O4′−C1′−C2′−C3′	32	32	−24	−24
C1′−C2′−O2′−HO′2	−61	21	−153	93

For each nucleoside, the 4 × 72 = 288 data points from eq 1 were fitted by linear least-squares to the Fourier series shown in eq 2.

$$E^{\text{fit}}_{\text{CHI}}(\phi_1, \phi_2) = \sum_{n=1}^{4} V_{n1}(1 + \cos(n\phi_1)) + V_{n2}(1 + \cos(n\phi_2)) \tag{2}$$

Here, ϕ_1 and ϕ_2 are the dihedral angles of O4′−C1′−N1−C6 (O4′−C1′−N9−C8) and C2′−C1′−N1−C6 (C2′−C1′−N9−C8), respectively. V_{n1} and V_{n2} are the potential energy barriers of O4′−C1′−N1−C6 (O4′−C1′−N9−C8) and C2′−C1′−N1−C6 (C2′−C1′−N9−C8) torsions. For each nucleoside, a separate fitting was done to calculate the χ torsion energy barriers, V_{n1} and V_{n2}. The new χ torsion parameters are listed in Table 2.

2.4. MD Simulations of Cytidine, Uridine, Adenosine, and Guanosine. Each structure was created with the xleap module of AMBER9.[57] Two conformations were used as initial structures: C3′-endo sugar puckering with base orientations of anti or syn. C, U, A, and G were solvated with TIP3P water molecules[58] in a truncated octahedral box, having 458, 451, 427, and 430 water molecules, respectively.

The structures were minimized in two steps: (i) With the nucleoside held fixed with a restraint force of 500 kcal/mol·Å^2, steepest descent minimization of 500 steps was followed by a conjugate gradient minimization of 500 steps. (ii) With all restraints removed, steepest descent minimization of 1000 steps was followed by a conjugate gradient minimization of 1500 steps. The long-range cutoff for nonbonded interactions during the minimization was 8 Å.

After minimization, two steps of pressure equilibration were done with the SANDER module in AMBER9: (i) Nucleosides were held fixed with a restraint force of 10 kcal/mol·Å^2. Constant volume dynamics with a long-range cutoff of 8 Å was used. SHAKE[59] was turned on for bonds involving hydrogen atoms. The temperature was raised from 0 to 300 K in 20 ps. Langevin dynamics with a collision frequency of 1 ps^{-1} was used. A total of 20 ps of MD were run with a 2 fs time step. (ii) The above conditions were chosen, except the constant pressure dynamics with isotropic position scaling was turned on. The reference pressure was 1 atm with a pressure relaxation time of 2 ps. A total of 100 ps of MD were run with a 2 fs time step. The particle mesh Ewald (PME) method was used for all simulations.

The production run was similar to the second step of the pressure equilibration described above. Constant pressure dynamics was chosen with a long-range cutoff of 8 Å. SHAKE was turned on for bonds involving hydrogen atoms. For each nucleoside, a total of 30 ns of MD were run with a 1 fs time step. For cytidine with AMBER99χ force field, the simulation time was 120 ns for convergence purposes. In production runs, simulations were carried out with the PMEMD module in AMBER9.[57] Trajectory files were written at each 250 fs time step.

Table 2. New χ Torsion Parameters for Adenosine, Guanosine, Cytidine, and Uridine

nucleoside	torsion	n	V_n	nucleoside	torsion	n	V_n
adenosine	O4′−C1′−N9−C8	1	1.355570	cytidine	O4′−C1′−N1−C6	1	0.331762
		2	0.504875			2	0.592225
		3	−1.699430			3	−3.108180
		4	0.152425			4	−0.116806
	C2′−C1′−N9−C8	1	1.603540		C2′−C1′−N1−C6	1	1.724800
		2	−0.278197			2	−0.62684
		3	1.267980			3	2.287890
		4	0.228818			4	0.0664267
guanosine	O4′−C1′−N9−C8	1	0.835436	uridine	O4′−C1′−N1−C6	1	0.0409516
		2	0.789849			2	0.604617
		3	0.351892			3	−2.686990
		4	0.183535			4	−0.0104774
	C2′−C1′−N9−C8	1	1.047920		C2′−C1′−N1−C6	1	1.235900
		2	−0.0516452			2	−0.683638
		3	−0.905523			3	2.277010
		4	0.131907			4	0.147500

Simulations were performed for systems prepared with the AMBER99 and AMBER99χ force fields. For C, U, A, and G, and each force field, 10 separate simulations of 30 ns each were run at 300 K yielding a total of 300 ns of explicit solvent MD simulation (see Supporting Information). Five of the 10 MD simulations had a starting structure of anti type, while the other five had a starting structure of syn type (see Supporting Information). For C with AMBER99χ force field, the simulations were extended to 11 separate simulations with 120 ns each (see Supporting Information). The fractions of anti and syn conformations observed were essentially independent of the starting structure as were values obtained for C when the time for each of the 11 simulations was extended from 30 to 60 ns and 120 ns (see Supporting Information).

Ultrasonic relaxation studies in aqueous solution revealed a relaxation time of 3 ns for A and no relaxation signal for pyrimidines.[60,61] The relaxation signal is attributed to the syn→anti transformation of the χ torsion. Evidently, 300 ns of MD simulations of the nucleosides is sufficient to sample adequately the syn→anti transformation.

3. Results

3.1. NMR Results for Cytidine and Uridine. In solution, nucleosides have two important regions that describe their structures: (i) the glycosidic dihedral angle, and (ii) the sugar pucker. NMR NOE experiments were done to analyze the structures of C and U.

The magnitudes of NOEs are proportional to $1/(r_{ij})^6$, where r_{ij} is the distance between the protons of i and j. When the base of a pyrimidine is oriented in an anti conformation, the H6 proton is about 3.5 Å from the H1′ proton, essentially independent of sugar pucker.[62] Thus, irradiation of H6 yields a moderate NOE to H1′. When the base of a pyrimidine is oriented in a syn conformation, however, the H6 proton is about 2.1 Å from H1′, yielding a strong NOE to H1′ when H6 is irradiated.[62] In pyrimidines, the distance between the H5 and H6 protons is constant at 2.48 Å, which can be used as a reference for calculating interproton distances from NOESY or transient NOE experiments according to eq 3:[63]

$$\mathrm{NOE}_{ij} = \mathrm{NOE}_{\mathrm{H5H6}}\frac{(r_{\mathrm{H5H6}})^6}{(r_{ij})^6} \tag{3}$$

Here, NOE_{ij} is the NOE between protons i and j, $\mathrm{NOE}_{\mathrm{H5H6}}$ is the NOE between H5 and H6 protons, and r_{H5H6} is the distance between the H5 and H6 protons, i.e. 2.48 Å.

Transient NOE spectroscopy[51] with different mixing times was used to quantitatively analyze the preferences for anti/syn populations, and the results are presented in Table 3 (also see Supporting Information). Transient NOE is similar to NOESY NMR except that it is 1D. To minimize spin diffusion effects and maximize signal-to-noise ratio, mixing times in the linear region of intensity vs mixing time plots were used to estimate distances between protons (see Supporting Information). A two-state model described by the following equation, which assumes that the structure is in either syn or anti conformations, was used to determine the proportions of each conformation:

$$\frac{\mathrm{NOE}_{\mathrm{H1'H6}}}{\mathrm{NOE}_{\mathrm{H5H6}}(r_{\mathrm{H5H6}})^6} = \frac{F_{\mathrm{anti}}}{(r_{\mathrm{H1'H6,anti}})^6} + \frac{F_{\mathrm{syn}}}{(r_{\mathrm{H1'H6,syn}})^6} \tag{4}$$

Here, $\mathrm{NOE}_{\mathrm{H1'H6}}$ is the NOE between the protons of H1′ and H6, F_{anti} and F_{syn} are the fractions of anti and syn conformations satisfying $F_{\mathrm{anti}} + F_{\mathrm{syn}} = 1$, $r_{\mathrm{H1'H6,anti}}$ and $r_{\mathrm{H1'H6,syn}}$ are the distances between the protons of H1′ and H6 when the structures are in anti and syn conformations, respectively, which are 3.48 Å and 2.12 Å, corresponding to the distances extracted from the minimum energy structures of the PES scans for C and U (see Methods Section). As can be seen from Table 3, the anti orientation is favored over syn. Comparison of NMR results for C at 2 and 10 °C show that the fraction of anti base orientation is essentially independent of temperature (Supporting Information). Higher temperature could not be used because of the overlap of the H1′ and H5 peaks (see Supporting Information). SSNOE spectroscopy confirms that anti is favored over syn base orientation (see Supporting Information).

Table 3. Experimentally Deduced and Force Field Predicted Base Orientation and Sugar Puckering for C, U, A, and G, and $\Delta G°$ (in kcal/mol) of Syn→Anti and C2′-endo→C3′-endo Transformations for C and U[a]

	base orientation, % anti			sugar pucker, % C3′-endo		
	($\Delta G°_{\text{syn}\rightarrow\text{anti}}$, kcal/mol)			($\Delta G°_{\text{C2}'\rightarrow\text{C3}'}$, kcal/mol)		
	AMBER99	AMBER99χ	NMR[b]	AMBER99	AMBER99χ	NMR[c]
C	30	66	87	27	54	60
	(0.49)	(−0.45)	(−1.07)	(0.58)	(−0.11)	(−0.24)
U	28	83	93	35	55	56
	(0.55)	(−0.95)	(−1.45)	(0.36)	(−0.13)	(−0.15)
A	15[e]	13[f]	–	24	32	37[d]
G	11[e]	24[f]	–	35	54	41[d]

[a] For a transformation of A→B, $\Delta G°_{\text{A}\rightarrow\text{B}} = -RT\ln(K)$, where $R = 1.987$ cal K^{-1} mol^{-1}, T is the temperature in kelvins, and K is the ratio of the concentrations of each species, [B]/[A] (see Supporting Information). [b] Measurements of the syn/anti proportions of pyrimidines were extracted from transient NOE experiments at 10 °C, while the simulations were done at 300 K (27 °C). NMR spectra for C at 2 and 10 °C indicate essentially no temperature dependence for the syn→anti equilibrium (see Supporting Information), so all $\Delta G°$'s were calculated at 300 K. [c] These values are for 30 °C (see Supporting Information). [d] These values are for 0.2 mM samples of A and G at 30 °C where there may be some association (see Supporting Information). [e] These values represent populations of high anti conformations with $\chi \approx 310°$ (see Supporting Information and Figure 7). [f] These values represent populations of anti conformations with $\chi \approx 185°$ (see Supporting Information and Figure 7).

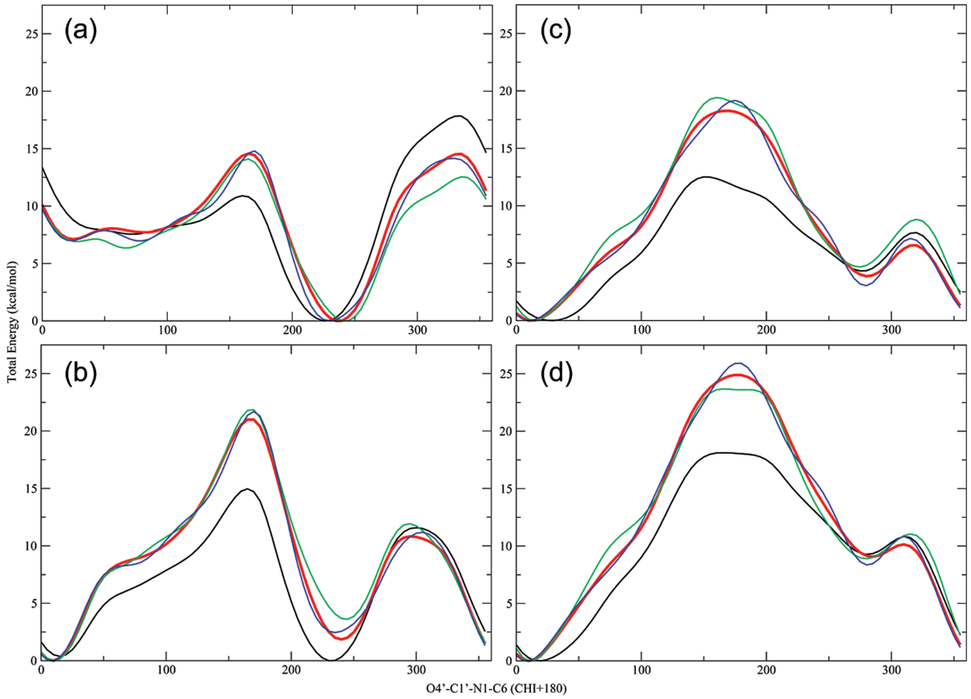


Figure 2. Total energy (in kcal/mol) vs O4′−C1′−N1−C6 of cytidine with AMBER99 (black), AMBER99χ (red), QM (green), and Ode force field (blue) for: (a) sc 1, (b) sc 2, (c) sc 3, and (d) sc 4 (see Table 1). For visualization purposes, minimum energies of each curve are set to zero. Anti, high anti, and syn base orientations correspond to *x*-axis ranges of 0−70°, 100−180°, and 200−300°, respectively, because the *x*-axis is $\chi + 180°$ to be consistent with the AMBER94 force field.[2]

Sugar proton coupling constants extracted from 1D ^{1}H NMR spectra were used to determine the sugar puckering on the basis of the following equation:[64]

$$\%\text{C3}'\text{endo} = 100\left(\frac{^3J_{3'4'}}{^3J_{1'2'} + {}^3J_{3'4'}}\right) \tag{5}$$

where $^3J_{1'2'}$ and $^3J_{3'4'}$ are ^3J spin−spin couplings between H1′ and H2′ and between H3′ and H4′ protons, respectively. The proportion of C2′-endo sugar puckering is equal to (1 − fraction of C3′-endo). Sugar pucker (±2%) is independent of temperature from 5 to 40 °C (Supporting Information), and results at 30 °C are presented in Table 3.

3.2. Comparison of Force Field to QM Energies. Figures 2−5 show the QM, MM$_{\text{AMBER99}}$, MM$_{\text{AMBER99}\chi}$, and MM$_{\text{Ode}}$[65] energy profiles with respect to the glycosidic dihedral angle of all the structures used in the fitting protocol for the nucleosides, where AMBER99, AMBER99χ, and Ode[65] force fields were used to calculate MM$_{\text{AMBER99}}$, MM$_{\text{AMBER99}\chi}$, and MM$_{\text{Ode}}$ energies, respectively. In all the plots, energy profiles of the AMBER99χ force field describe the QM energy profiles best, although the Ode force field's energy profile is also similar to the QM energy profiles. The differences between the predictions of the AMBER99χ and Ode force fields is likely due to the Ode force field using CH_3, $H_2C{-}CH_3$ and $H_2C{-}O{-}CH_3$ as model systems to

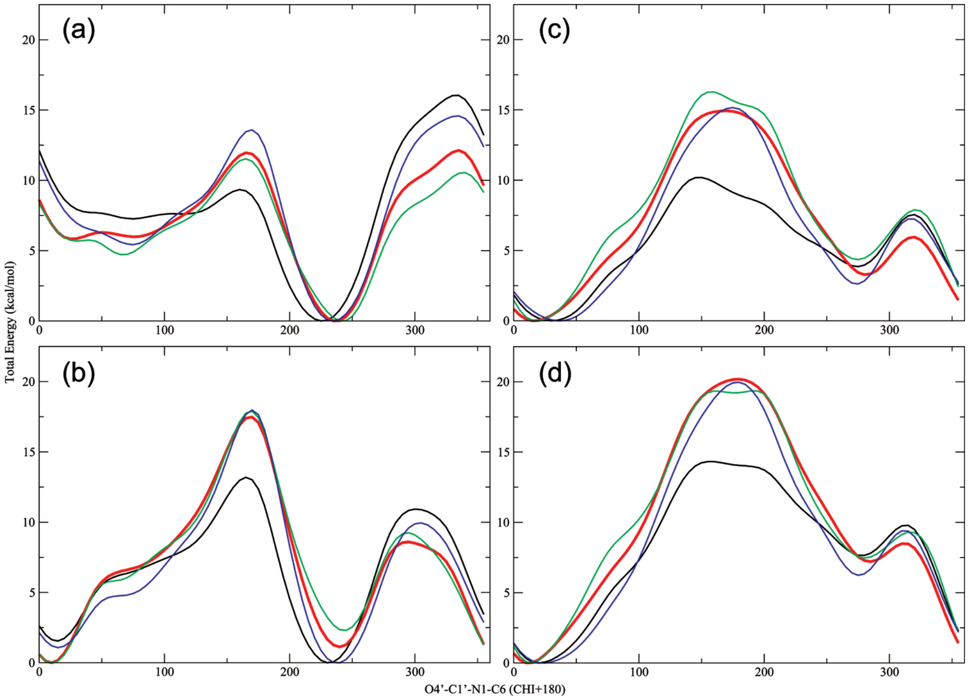


Figure 3. Total energy (in kcal/mol) vs O4′−C1′−N1−C6 of uridine with AMBER99 (black), AMBER99χ (red), QM (green), and Ode force field (blue) for: (a) sc 1, (b) sc 2, (c) sc 3, and (d) sc 4 (see Table 1). For visualization purposes, minimum energies of each curve are set to zero. Anti, high anti, and syn base orientations correspond to *x*-axis ranges of 0−70°, 100−180°, and 200−300°, respectively, because the *x*-axis is $\chi + 180°$ to be consistent with the AMBER94 force field.[2]

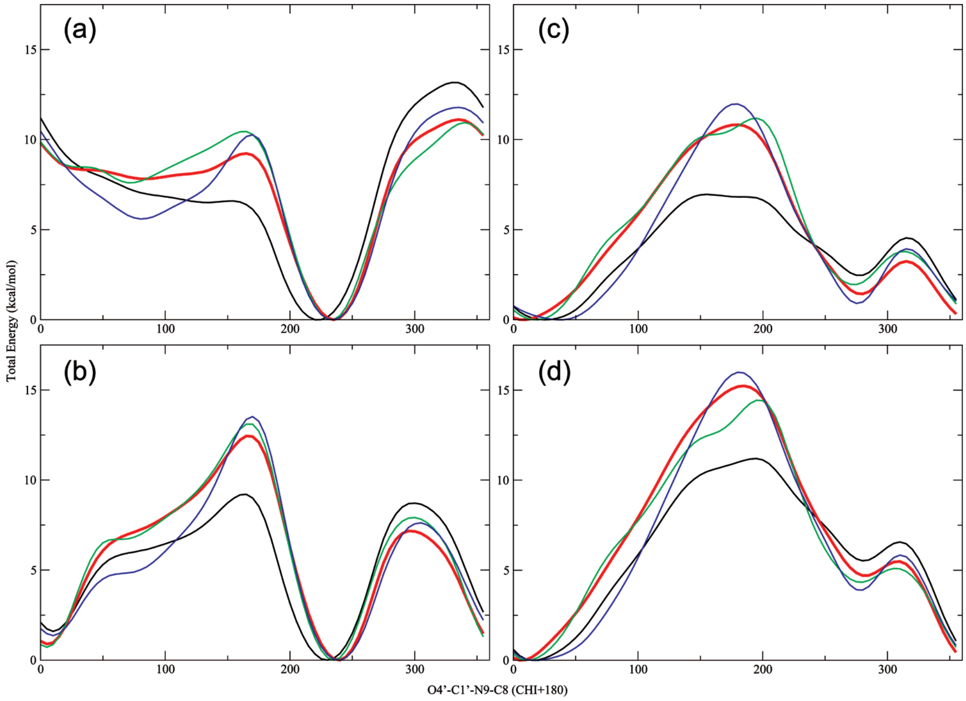


Figure 4. Total energy (in kcal/mol) vs O4′−C1′−N9−C8 of adenosine with AMBER99 (black), AMBER99χ (red), QM (green), and Ode force field (blue) for: (a) sc 1, (b) sc 2, (c) sc 3, and (d) sc 4 (see Table 1). For visualization purposes, minimum energies of each curve are set to zero. Anti, high anti, and syn base orientations correspond to *x*-axis ranges of 0−70°, 100−180°, and 200−300°, respectively, because the *x*-axis is $\chi + 180°$ to be consistent with the AMBER94 force field.[2]

represent the sugar, while the AMBER99χ force field used the entire ribose with four different sugar conformations to calculate the χ torsional parameters. The Ode force field also uses more parameters. Yet, both Ode and AMBER99χ force fields should provide similar predictions for structural and/or dynamical properties of RNA. Comparisons of the force fields to QM calculations on eight sugar conformations not included in the fitting showed that AMBER99χ also describes those QM energy profiles better than AMBER99 and Ode force fields (see Supporting Information).

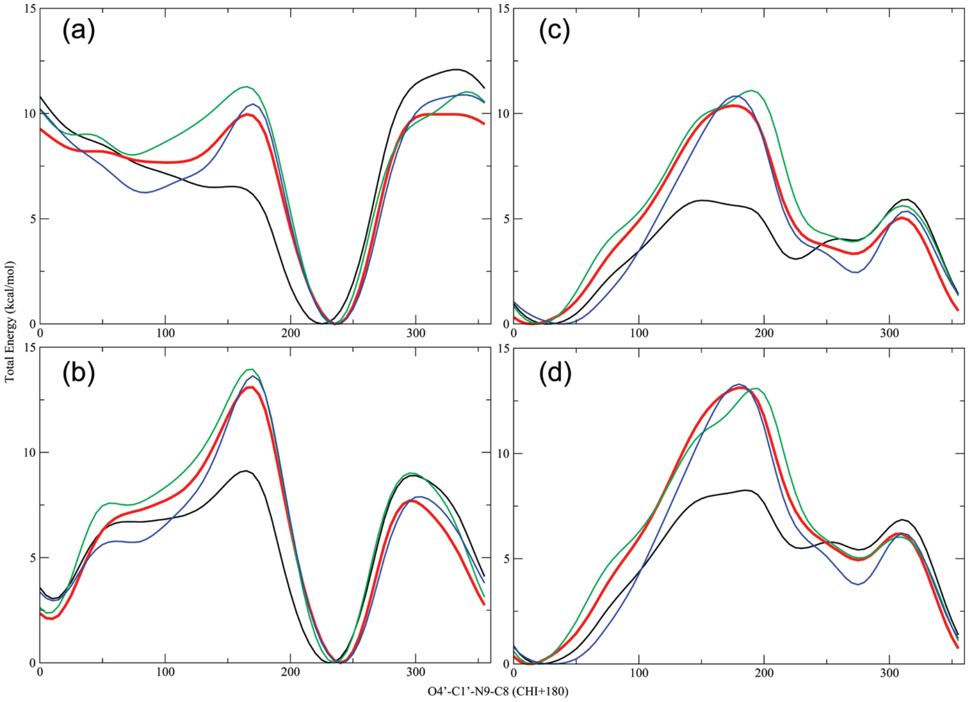


Figure 5. Total energy (in kcal/mol) vs O4′−C1′−N9−C8 of guanosine with AMBER99 (black), AMBER99χ (red), QM (green), and Ode force field (blue) for: (a) sc 1, (b) sc 2, (c) sc 3, and (d) sc 4 (see Table 1). For visualization purposes, minimum energies of each curve are set to zero. Anti, high anti, and syn base orientations correspond to *x*-axis ranges of 0−70°, 100−180°, and 200−300°, respectively, because the *x*-axis is $\chi + 180°$ to be consistent with the AMBER94 force field.[2]

3.3. MD Simulations of Cytidine, Uridine, Adenosine, and Guanosine with AMBER99 and AMBER99χ. For comparison with NMR results, predictions of population distributions of χ dihedral angle and sugar pucker were analyzed for C, U, A, and G using the combined trajectories of the 10 individual MD simulations with AMBER99 and AMBER99χ force fields (see Methods). Population distribution plots in 2D of χ dihedral and pseudorotation angles for each nucleoside are shown in Figures 6 and 7. Table 3 shows the force field predictions of base orientation and sugar pucker for each nucleoside (also see Table 4 and Supporting Information). Analyses of the individual MD simulations show at least seven syn↔anti transformations for each (see Supporting Information).

4. Discussion

Table 3 shows the experimental results for C and U as well as the predictions of AMBER99 and AMBER99χ force fields of the base orientation and the sugar pucker for C, U, A, and G. For the syn→anti equilibrium of C and U, NMR indicates 87% and 93% anti conformation, respectively, corresponding to $\Delta G°_{\text{syn}\to\text{anti}}$ of −1.07 and −1.45 kcal/mol. The AMBER99 force field predicts 30% and 28% anti conformation, respectively, corresponding to $\Delta G°_{\text{syn}\to\text{anti}}$ of 0.49 and 0.55 kcal/mol. In comparison, the AMBER99χ force field predicts 66% and 83% anti conformation, respectively, corresponding to $\Delta G°_{\text{syn}\to\text{anti}}$ of −0.45 and −0.95 kcal/mol, closer to the NMR results. Evidently, AMBER99 overestimates the syn conformations of C and U (see Figure 6).

For the C2′-endo→C3′-endo equilibrium of C and U, NMR indicates 60% and 56% C3′-endo sugar puckering at 30 °C, respectively, corresponding to free energy differences, $\Delta G°_{C2'\to C3'}$, of −0.24 and −0.15 kcal/mol (Table 3). The percentages are essentially independent of temperature from 5 to 40 °C (see Supporting Information). The AMBER99 force field predicts 27% and 35% C3′-endo sugar pucker at 27 °C, respectively, corresponding to $\Delta G°_{C2'\to C3'}$ of 0.58 and 0.36 kcal/mol. In comparison, the AMBER99χ force field predicts 54% and 55% C3′-endo sugar pucker at 27 °C, respectively, corresponding to $\Delta G°_{C2'\to C3'}$ of −0.11 and −0.13 kcal/mol, which is close to the experimental values. Evidently, AMBER99 underestimates C3′-endo sugar puckering of C and U (see Figure 6).

The AMBER99χ force field predicts A and G to have 13% and 24% anti conformation (Table 3), respectively, with a χ dihedral angle around 185°, which is consistent with QM calculations and typical of the anti region seen in crystal structures of RNA.[66] The AMBER99 force field predicts 15% and 11% anti conformation (Table 3), respectively, but with a χ dihedral angle around 310° (Figure 7), which is the high anti region. QM PES scans did not find any minimum around 310° but rather between 180−250° for three different sugar puckers for A and G (Figures 4−5 and Supporting Information, where the *x*-axis, however, is $\chi + 180°$).

The concentration dependence of chemical shifts for A and G indicated aggregation at concentrations required to determine NOEs with enough signal-to-noise to determine the base orientation quantitatively. Pioneering studies of 2′- and 3′- AMP and GMP at high concentrations, however, indicated syn populations well over 50%.[67,68]

The AMBER99 force field predicts A and G to have 24% and 35% C3′-endo sugar puckering, respectively, while AMBER99χ predicts 32% and 54%. Chemical shift data of 0.2, 1.0, and 5.0 mM A implies base stacking that differs

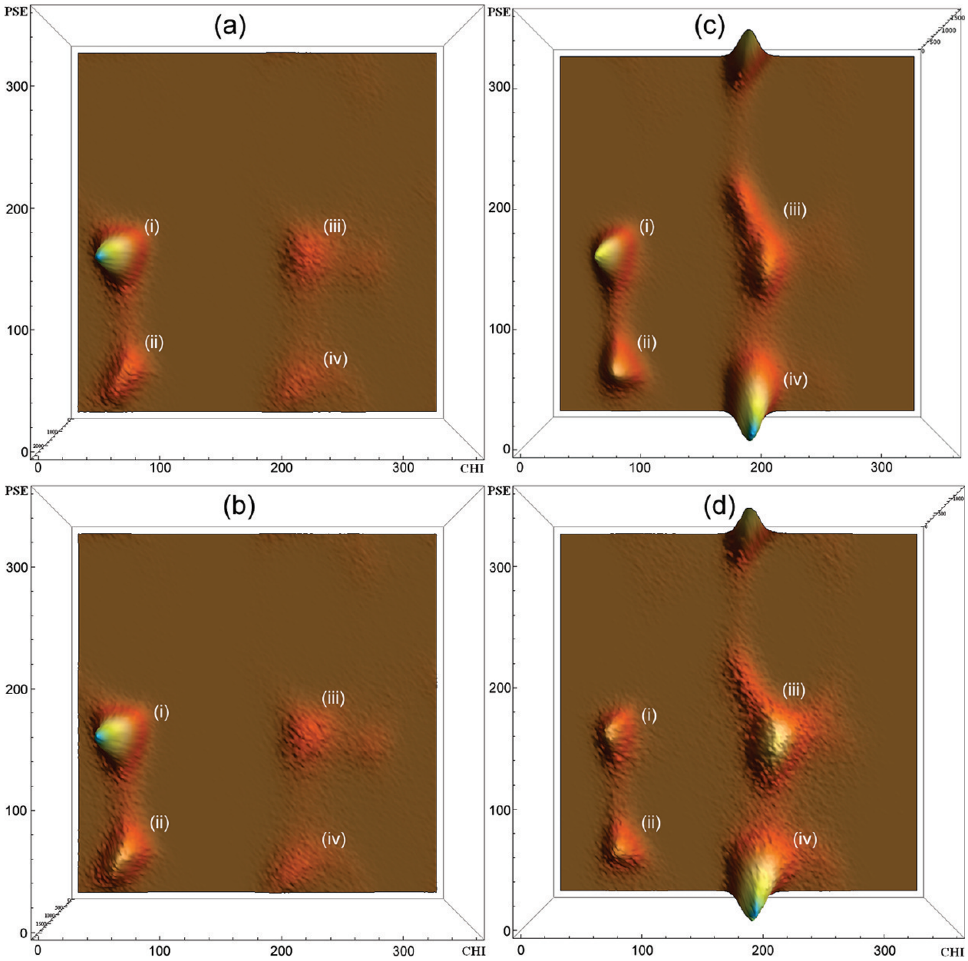


Figure 6. Population distribution of cytidine and uridine using AMBER99 (a and b, respectively) and AMBER99χ (c and d, respectively) force fields. PSE (*y*-axis) and CHI (*x*-axis) stand for the pseudorotation and χ dihedral angles. Table 4 shows the predicted populations of (i−iv). PSE angles of 18° and 162° correspond to C3′-endo and C2′-endo sugar pucker, respectively. χ angles of 200°, 300°, and 60° correspond to anti, high-anti, and syn conformations, respectively.[70]

with concentration. The differences of the chemical shifts between 0.2 and 1.0 mM samples are small, however. Therefore, the 0.2 mM samples of A and G were used to calculate ^3J spin−spin couplings to estimate the sugar puckering (see Supporting Information). At room temperature, the C3′-endo sugar puckering of A and G is about 40% (Table 3 and Supporting Information). For A, both force fields' predictions are similar to the experimental results. For G, the AMBER99 force field apparently predicts better than AMBER99χ does. It is known, however, that guanosine monophosphate forms quadruplex structures and other aggregates in solution.[53−55] Aggregation and precipitation were seen by eye in the 5 mM G NMR samples. Thus, it is not conclusive whether 0.2 mM G can be used to reveal the sugar puckering of monomer G.

There are several reasons why the AMBER99χ force field improves predictions for nucleosides. When the χ torsions were parametrized for AMBER99, model systems for adenosine and thymidine were used, and the results were generalized for all DNA/RNA residues.[2] Moreover, the model systems mimicked deoxyribose C2′-endo sugar puckering. At that time, QM calculations were limited by computer power and only 8−9 data points were used in the QM fitting. Also, in the AMBER99 force field, the original Cornell force field parameters for χ torsions were changed without doing any fitting. The V_2 term of χ torsion parameters was zeroed to improve the C2′-endo sugar puckering phase angle for DNA residues.[69] This effect, however, changes the whole predicted potential energy surface of the nucleosides, which, therefore, does not represent the QM energy surface well.

For the AMBER99χ force field, the χ torsions of C, U, A, and G were reparameterized individually. A multiconformational fitting that included the entire nucleoside with different sugar puckering was done to provide the χ torsion parameters. In the PES scan, a total of 4 × 72 = 288 data points were used in the fitting protocol for each nucleoside. The new parameter set was tested on 12 different sugar conformations (four separate conformations for each of C2′-endo, C3′-endo, and O4′-endo sugar puckering) for each nucleoside and shown to predict well the QM energy surface for these conformations (see Figures 2−5 and Supporting Information). The shape of the QM energy surfaces of these conformations is also predicted well by the Ode force field,[65] although not quite as well as by AMBER99χ (see Figures 2−5 and Supporting Information). As a result, there should not be any big difference between AMBER99χ and Ode

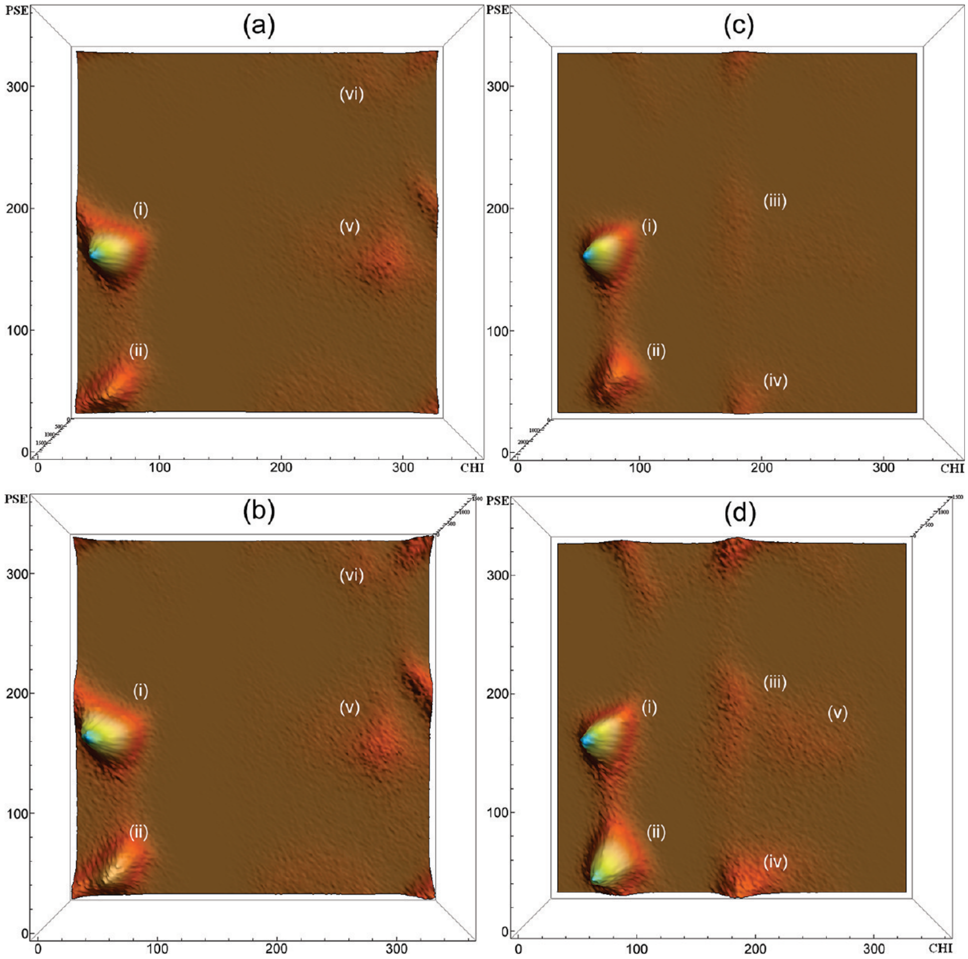


Figure 7. Population distribution of adenosine and guanosine using AMBER99 (a and b, respectively), and AMBER99χ (c and d, respectively) force fields. PSE (*y*-axis) and CHI (*x*-axis) stand for the pseudorotation and χ dihedral angles. Table 4 shows the predicted populations of (i), (ii), (iii), (iv), (v), and (vi). PSE angles of 18° and 162° correspond to C3′-endo and C2′-endo sugar pucker, respectively. χ angles of 200°, 300°, and 60° correspond to anti, high-anti, and syn conformations, respectively.[70]

Table 4. Population Analysis Results for C, U, A and G of the AMBER99 and AMBER99χ Force Fields[a]

	i (%)	ii (%)	iii (%)	iv (%)	v (%)	vi (%)
AMBER99						
cytidine	52	16	19	11	–	–
uridine	47	24	17	11	–	–
adenosine	57	21	–	–	12	3
guanosine	54	31	–	–	7	4
AMBER99χ						
cytidine	20	11	23	43	–	–
uridine	9	8	36	47	–	–
adenosine	59	24	5	8	–	–
guanosine	33	39	9	15	2	–

[a] Regions of (i) syn/C2′-endo, (ii) syn/C3′-endo, (iii) anti/C2′-endo, (iv) anti/C3′-endo, (v) high anti/C2′-endo, and (vi) high-anti/C3′-endo (Figures 6 and 7).

force field[65] predictions for structural and thermodynamic properties of nucleosides.

Many reasonable combinations of parameters were tested for approximating the QM PES representing the four major conformations of each nucleoside. For instance, we tried fitting to two dihedrals with three cosine terms, four dihedrals with two cosine terms, and four dihedrals with three cosine terms. Two dihedrals with four cosine terms provided excellent fits, and more terms gave minimal improvement. As a comparison, the Ode force field[65] uses 3 dihedrals (a total of 13 V_i parameters) to represent the χ torsions, while we use 2 dihedrals (a total of 8 V_i parameters), but comparisons of the force fields to the QM potential energy surfaces shown in Figures 2−5 and Supporting Information reveal that AMBER99χ provides a better fit. This may be because the calculations for AMBER99χ included the entire ribose group.

It is crucial to use a force field that appropriately models the true behavior of RNA systems. Otherwise, during MD simulations, sampling space will include unphysical regions, which will cause errors in predictions. With the AMBER99χ modification, significant improvements are seen in the structural and thermodynamic predictions for cytidine and uridine in solution (Table 3). This modification should be particularly important for non-Watson−Crick regions and terminal base pairs because sampling will not include unrealistic populations of syn conformations or of C2′-endo sugar puckering. In Watson−Crick regions, the χ torsion is restricted by hydrogen bonding in base pairs, so little effect should be seen there. Thus, the AMBER99χ force field should improve structural and thermodynamic predictions for RNA.

Acknowledgment. We are grateful to the Center for Research Computing at the University of Rochester for providing the necessary computing systems (BlueHive cluster) and to the personnel to enable the research presented in this manuscript. This work was supported by NIH grant GM22939 (D.H.T.).

Supporting Information Available: Frozen dihedrals during QM optimization in PES scan; restrained dihedral angles and sample restraint file used in MM minimization; chemical shift data of C, U, A and G; sugar pucker conformations used to test AMBER99, AMBER99χ, and Ode force fields; NOE data from transient NOE and SSNOE for C and U; details of population analysis results for C, U, A, and G; details of $\Delta G°$ values of C2′→C3′ and syn→anti transformations for C and U; ^3J spin−spin couplings and experimentally deduced sugar puckering of C, U, A, and G; detailed analysis of Table 3; analysis of individual simulations of C, U, A, and G; comparison of AMBER99, AMBER99χ, and Ode force fields to QM energy surface predictions for C, U, A, and G; intensity vs mixing time plots of C and U; rmsd vs time plots of MD simulations of C, U, A, and G; convergence analysis of the population distributions of C, U, A, and G. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Malhotra, A.; Harvey, S. C. A quantitative model of the Escherichia coli 16S RNA in the 30S ribosomal subunit. *J. Mol. Biol.* **1994**, *240*, 308.

(2) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

(3) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM - A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187.

(4) MacKerell, J., A. D.; Brooks, B.; Brooks, C. L., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M., CHARMM: The Energy Function and Its Parametrization with an Overview of the Program. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, I., H. F., Schreiner, P. R., Eds. John Wiley & Sons: Chichester, U.K., 1998; Vol. 1, pp 271.

(5) Scott, W. R. P.; Hunenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Kruger, P.; van Gunsteren, W. F. The GROMOS biomolecular simulation program package. *J. Phys. Chem. A* **1999**, *103*, 3596.

(6) 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*, Verlag der Fachvereine: Zürich, Switzerland, 1996.

(7) Hobza, P.; Sponer, J. Toward true DNA base-stacking energies: MP2, CCSD(T), and complete basis set calculations. *J. Am. Chem. Soc.* **2002**, *124*, 11802.

(8) Kratochvil, M.; Sponer, J.; Hobza, P. Global minimum of the Adenine•Thymine base pair corresponds neither to Watson-Crick nor to Hoogsteen structures. Molecular dynamic/Quenching/AMBER and ab initio beyond Hartree-Fock studies. *J. Am. Chem. Soc.* **2000**, *122*, 3495.

(9) Nam, K.; Gao, J. L.; York, D. M. Electrostatic interactions in the hairpin ribozyme account for the majority of the rate acceleration without chemical participation by nucleobases. *RNA* **2008**, *14*, 1501.

(10) Nam, K. H.; Gao, J. L.; York, D. M. Quantum mechanical/molecular mechanical simulation study of the mechanism of hairpin ribozyme catalysis. *J. Am. Chem. Soc.* **2008**, *130*, 4680.

(11) Bash, P. A.; Field, M. J.; Karplus, M. Free energy perturbation method for chemical reactions in the condensed phase - A dynamical approach based on a combined Quantum and Molecular Mechanics potential. *J. Am. Chem. Soc.* **1987**, *109*, 8092.

(12) Eichinger, M.; Tavan, P.; Hutter, J.; Parrinello, M. A hybrid method for solutes in complex solvents: Density functional theory combined with empirical force fields. *J. Chem. Phys.* **1999**, *110*, 10452.

(13) Freindorf, M.; Gao, J. L. Optimization of the Lennard-Jones parameters for a combined ab initio quantum mechanical and molecular mechanical potential using the 3-21G basis set. *J. Comput. Chem.* **1996**, *17*, 386.

(14) Gao, J. L.; Xia, X. F. A priori evaluation of aqueous polarization effects through Monte-Carlo QM/MM simulations. *Science* **1992**, *258*, 631.

(15) Murphy, R. B.; Philipp, D. M.; Friesner, R. A. A mixed Quantum Mechanics/Molecular Mechanics (QM/MM) method for large-scale modeling of chemistry in protein environments. *J. Comput. Chem.* **2000**, *21*, 1442.

(16) Stanton, R. V.; Hartsough, D. S.; Merz, K. M. Calculation of solvation free energies using a density functional Molecular Dynamics coupled potential. *J. Phys. Chem.* **1993**, *97*, 11868.

(17) Tunon, I.; MartinsCosta, M. T. C.; Millot, C.; RuizLopez, M. F.; Rivail, J. L. A coupled density functional-molecular mechanics Monte Carlo simulation method: The water molecule in liquid water. *J. Comput. Chem.* **1996**, *17*, 19.

(18) Warshel, A.; Levitt, M. Theoretical studies of enzymatic reactions - Dielectric, electrostatic and steric stabilization of Carbonium-Ion in reaction of Lysozyme. *J. Mol. Biol.* **1976**, *103*, 227.

(19) Duan, Y.; Kollman, P. A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **1998**, *282*, 740.

(20) Russell, R.; Millettt, I. S.; Tate, M. W.; Kwok, L. W.; Nakatani, B.; Gruner, S. M.; Mochrie, S. G. J.; Pande, V.; Doniach, S.; Herschlag, D.; Pollack, L. Rapid compaction during RNA folding. *Proc. Natl. Acad. Sci. U. S.A.* **2002**, *99*, 4266.

(21) Trylska, J.; Tozzini, V.; McCammon, J. A. Exploring global motions and correlations in the ribosome. *Biophys. J.* **2005**, *89*, 1455.

(22) Zagrovic, B.; Pande, V. Solvent viscosity dependence of the folding rate of a small protein: Distributed computing study. *J. Comput. Chem.* **2003**, *24*, 1432.

(23) Zagrovic, B.; Sorin, E. J.; Pande, V. Beta-hairpin folding simulations in atomistic detail using an implicit solvent model. *J. Mol. Biol.* **2001**, *313*, 151.

(24) Cheatham, T. E.; Young, M. A. Molecular dynamics simulation of nucleic acids: Successes, limitations, and promise. *Biopolymers* **2000**, *56*, 232.

(25) Giudice, E.; Lavery, R. Simulations of nucleic acids and their complexes. *Acc. Chem. Res.* **2002**, *35*, 350.

(26) Orozco, M.; Perez, A.; Noy, A.; Luque, F. J. Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.* **2003**, *32*, 350.

(27) Perez, A.; Blas, J. R.; Rueda, M.; Lopez-Bes, J. M.; de la Cruz, X.; Orozco, M. Exploring the essential dynamics of B-DNA. *J. Chem. Theory Comput.* **2005**, *1*, 790.

(28) Beveridge, D. L.; McConnell, K. J. Nucleic acids: theory and computer simulation, Y2K. *Curr. Opin. Struct. Biol.* **2000**, *10*, 182.

(29) Cheatham, T. E. Simulation and modeling of nucleic acid structure, dynamics and interactions. *Curr. Opin. Struct. Biol.* **2004**, *14*, 360.

(30) Fadrna, E.; Spackova, N.; Sarzynska, J.; Koca, J.; Orozco, M.; Cheatham, T. E.; Kulinski, T.; Sponer, J. Single Stranded Loops of Quadruplex DNA As Key Benchmark for Testing Nucleic Acids Force Fields. *J. Chem. Theory Comput.* **2009**, *5*, 2514.

(31) Yildirim, I.; Stern, H. A.; Sponer, J.; Spackova, N.; Turner, D. H. Effects of restrained sampling space and non-planar amino groups on free energy predictions for RNA with imino and sheared tandem GA base pairs flanked by GC, CG, iGiC or iCiG base pairs. *J. Chem. Theory Comput.* **2009**, *5*, 2088.

(32) Yildirim, I.; Turner, D. H. RNA challenges for computational chemists. *Biochemistry* **2005**, *44*, 13225.

(33) Wang, J. M.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules. *J. Comput. Chem.* **2000**, *21*, 1049.

(34) Davies, D. B. Conformations of nucleosides and nucleotides. *Prog. Nucl. Magn. Reson. Spectrosc.* **1978**, *12*, 135.

(35) Sundaralingam, M.; Arora, S. K. Stereochemistry of nucleic acids and their constituents. IX. The conformation of the antibiotic puromycin dihydrochloride pentahydrate. *Proc. Natl. Acad. Sci. U.S.A.* **1969**, *64*, 1021.

(36) Altona, C.; Sundaralingam, M. Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. *J. Am. Chem. Soc.* **1972**, *94*, 8205.

(37) Lai, T. F.; Marsh, R. E. The crystal structure of adenosine. *Acta Cryst. B* **1972**, *28*, 1982.

(38) Kraut, J.; Jensen, L. H. Crystal structure of adenosine-5′-phosphate. *Nature* **1960**, *186*, 798.

(39) Green, E. A.; Rosenstein, R. D.; Shiono, R.; Abraham, D. J.; Trus, B. L.; Marsh, R. E. The crystal structure of uridine. *Acta Cryst. B* **1975**, *31*, 102.

(40) Altona, C. Conformational analysis of nucleic acids. Determination of backbone geometry of single-helical RNA and DNA in aqueous solution. *Recl. Trav. Chim. Pays-Bas* **1982**, *101*, 413.

(41) Vandeven, F. J. M.; Hilbers, C. W. Nucleic acids and nuclear magnetic resonance. *Eur. J. Biochem.* **1988**, *178*, 1.

(42) Chapman, G. E.; Abercrombie, B. D.; Cary, P. D.; Bradbury, E. M. Measurement of small nuclear overhauser effects in H1 spectra of proteins, and their application to lysozyme. *J. Magn. Reson.* **1978**, *31*, 459.

(43) Richarz, R.; Wuthrich, K. NOE difference spectroscopy: A novel method for observing individual multiplets in proton nmr spectra of biological macromolecules. *J. Magn. Reson.* **1978**, *30*, 147.

(44) Gracz, H. S.; Guenther, R. H.; Agris, P. F.; Folkman, W.; Golankiewicz, B. Structure and conformation of the hypermodified purine nucleoside wyosine and its isomers: A comparison of coupling constants and distance geometry solutions. *Magn. Reson. Chem.* **1991**, *29*, 885.

(45) Desaulniers, J. P.; Chui, H. M. P.; Chow, C. S. Solution conformations of two naturally occurring RNA nucleosides: 3-Methyluridine and 3-methylpseudouridine. *Bioorg. Med. Chem.* **2005**, *13*, 6777.

(46) Rosemeyer, H.; Toth, G.; Golankiewicz, B.; Kazimierczuk, Z.; Bourgeois, W.; Kretschmer, U.; Muth, H. P.; Seela, F. Syn-Anti conformational analysis of regular and modified nucleosides by 1D ^{1}H NOE difference spectroscopy: A simple graphical method based on conformationally rigid molecules. *J. Org. Chem.* **1990**, *55*, 5784.

(47) Chang, Y. C.; Herath, J.; Wang, T. H. H.; Chow, C. S. Synthesis and solution conformation studies of 3-substituted uridine and pseudouridine derivatives. *Bioorg. Med. Chem.* **2008**, *16*, 2676.

(48) Geraldes, C. F. G. C.; Santos, H.; Xavier, A. V. A proton relaxation study of the conformations of some purine mononucleotides in aqueous solution. *Can. J. Chem.* **1982**, *60*, 2976.

(49) Hart, P. A. Conformation of mononucleotides and dinucleoside monophosphates. P[H] and H[H] nuclear overhauser effects. *Biophys. J.* **1978**, *24*, 833.

(50) Santos, H.; Xavier, A. V.; Geraldes, C. F. G. C. Conformation of purine mononucleotides by H{H} and P{H} nuclear overhauser effects. *Can. J. Chem.* **1983**, *61*, 1456.

(51) Stott, K.; Stonehouse, J.; Keeler, J.; Hwang, T. L.; Shaka, A. J. Excitation sculpting in high-resolution nuclear magnetic resonance spectroscopy: Application to selective NOE experiments. *J. Am. Chem. Soc.* **1995**, *117*, 4199.

(52) Ts'o, P. O. P., Bases, Nucleosides and Nucleotides. In *Basic Principles in Nucleic Acid Chemistry*, Ts'o, P. O. P., Ed.; Academic: New York, 1974; Vol. I, pp 453.

(53) Pinnavaia, T. J.; Marshall, C. L.; Mettler, C. M.; Fisk, C. I.; Miles, H. T.; Becker, E. D. Alkali metal ion specificity in the solution ordering of a nucleotide, 5′-guanosine monophosphate. *J. Am. Chem. Soc.* **1978**, *100*, 3625.

(54) Gellert, M.; Lipsett, M. N.; Davies, D. R. Helix formation by guanylic acid. *Proc. Natl. Acad. Sci. U. S. A.* **1962**, *48*, 2013.

(55) Davis, J. T. G-quartets 40 years later: From 5 ′-GMP to molecular biology and supramolecular chemistry. *Angew. Chem., Int. Ed. Engl.* **2004**, *43*, 668.

(56) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.;

Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(57) Case, D. A.; Darden, T. A.; Cheatham, T. E. I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*, University of California San Francisco: San Francisco, CA, 2006.

(58) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926.

(59) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical-Integration of cartesian equations of motion of a system with constraints: Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327.

(60) Hemmes, P. R.; Oppenheimer, L.; Jordan, F. Ultrasonic relaxation evaluation of the thermodynamics of syn-anti glycosidic isomerization in adenosine. *J. Am. Chem. Soc.* **1974**, *96*, 6023.

(61) Rhodes, L. M.; Schimmel, P. R. Nanosecond relaxation processes in aqueous mononucleoside solutions. *Biochemistry* **1971**, *10*, 4426.

(62) Wuthrich, K. *NMR of Proteins and Nucleic Acids*; John Wiley & Sons: New York, 1986.

(63) Wemmer, D. Structure and Dynamics by NMR. In *Nucleic Acids: Structures, Properties, and Functions*; Bloomfield, V. A., Crothers, D. M., Tinoco, I., Jr., Eds.; University Science Books: Sausalito, CA, 2000; pp 111.

(64) Altona, C.; Sundaralingam, M. Conformational analysis of the sugar ring in nucleosides and nucleotides. Improved method for the interpretation of proton magnetic resonance coupling constants. *J. Am. Chem. Soc.* **1973**, *95*, 2333.

(65) Ode, H.; Matsuo, Y.; Neya, S.; Hoshino, T. Force field parameters for rotation around chi torsion axis in nucleic acids. *J. Comput. Chem.* **2008**, *29*, 2531.

(66) Richardson, J. S.; Schneider, B.; Murray, L. W.; Kapral, G. J.; Immormino, R. M.; Headd, J. J.; Richardson, D. C.; Ham, D.; Hershkovits, E.; Williams, L. D.; Keating, K. S.; Pyle, A. M.; Micallef, D.; Westbrook, J.; Berman, H. M. RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA* **2008**, *14*, 465.

(67) Son, T. D.; Guschibauer, W.; Gueron, M. Flexibility and conformations of guanosine monophosphates by the Overhauser effect. *J. Am. Chem. Soc.* **1972**, *94*, 7903.

(68) Gueron, M.; Chachaty, C.; Son, T. D. Properties of purine nucleotides studied by the Overhauser effect: conformations, flexibility, aggregation. *Ann. N.Y. Acad. Sci.* **1973**, *222*, 307.

(69) Cheatham, T. E.; Cieplak, P.; Kollman, P. A. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845.

(70) Schneider, B.; Moravek, Z.; Berman, H. M. RNA conformational classes. *Nucleic Acids Res.* **2004**, *32*, 1666.

CT900604A

JCTC Journal of Chemical Theory and Computation

Assessment of Functionals for TD-DFT Calculations of Singlet−Triplet Transitions

Denis Jacquemin,*,† Eric A. Perpète,† Ilaria Ciofini,‡ and Carlo Adamo*,‡

Groupe de Chimie-Physique Théorique et Structurale, Facultés Universitaires Notre-Dame de la Paix, Rue de Bruxelles, 61, B-5000 Namur, Belgium, and Ecole Nationale Supérieure de Chimie de Paris, Laboratoire Electrochimie et Chimie Analytique, UMR CNRS-ENSCP no. 7575, 11, Rue Pierre et Marie Curie, F-75321 Paris Cedex 05, France

Received January 4, 2010

Abstract: The calculation of transition energies for electronically excited states remains a challenge in quantum chemistry, for which time-dependent density functional theory (TD-DFT) is often viewed as a balanced (computational effort/obtained accuracy) technique. In this study, we benchmark 34 DFT functionals in the specific framework of TD-DFT calculations for singlet−triplet transitions. The results are compared to accurate wave function data reported for the same set of 63 excited-states, and it turns out that, within the selected TD-DFT framework, BMK and M06−2X emerge as the most efficient hybrids. This investigation clearly illustrates that the conclusions drawn for singlet excited states do not necessarily hold for triplet states, even for similar molecular structures.

1. Introduction

Time-dependent density functional theory (TD-DFT)[1−7] is probably the most extensively used theoretical tool for the computing electronic transition energies in organic or inorganic compounds. In the recent years, a series of benchmark calculations aimed at appraising the relative qualities of DFT functionals in the TD-DFT framework have appeared.[8−12] Although the detailed conclusions might significantly differ from one work to the other (e.g., see the discussion in ref 12), the typical mean absolute deviation (MAE) obtained with the most efficient functionals is in the range of 0.20−0.30 eV for singlet excited states. In the present article, we aim at evaluating the *pros* and *cons* of an extended panel of functionals for the calculation of singlet−triplet transitions, as the number of previous benchmarks remains limited, despite several specific TD-DFT applications for these states.[13−17] On the one hand, Grimme and Neese compared the B3LYP and double-hybrid singlet−triplet energies for a series of small molecules,[18] for which they obtained an average deviation limited to 0.25 eV with B3LYP and 0.18 eV with B2PLYP. On the other hand, Thiel and co-workers used their high-level ab initio estimates for 63 singlet−triplet transitions to assess the efficiency of TD-DFT calculations relying on the BP86, B3LYP, BHHLYP functionals as well as the DFT/multireference−configuration interaction (DFT/MR-CI) procedure.[10] The average deviations range from 0.4 to 0.6 eV for the three functionals, significantly exceeding the errors observed for the singlet−singlet transition energies in the same molecules. This is in full agreement with other works, hinting that the results produced with TD-DFT could be less accurate and significantly more functional-dependent for triplet states than for singlet states.[15,19]

One should note that the results presented in this contribution have been obtained by "traditional" TD-DFT, requiring extreme care when considering spin-flipping transitions if the target open-shell excited state cannot be described with a single-determinant model. Specific models have been developed to overcome such difficulties.[20,21] In this framework, it is worth pinpointing the investigation by Nguyen and co-workers who computed the $S_0 - T_1$ energy difference, for a large set of organic molecules, using UB3LYP calculations.[22] The results appeared pretty accurate, at the

* Corresponding authors. E-mail: denis.jacquemin@fundp.ac.be (D.J.) and carlo-adamo@enscp.fr (C.A.).

† FUNDP, Namur.

‡ ENSCP, Paris.

10.1021/ct100005d © 2010 American Chemical Society
Published on Web 04/14/2010

price of the limitation of the computational procedure to the lowest-lying triplet.

In this article, we investigate the efficiency of more than 30 DFT functionals within the TD-DFT formalism, considering the same set of transitions as in ref 10 as references. Therefore, we assess the functional's qualities using accurate wave function estimates, rather than experimental values. This choice is motivated, on the one hand, by the intention to ensure perfectly meaningful comparisons and, on the other hand, by the comparatively (with respect to the singlet case) limited number of accurate experimental data available in literature.

2. Method

As benchmark set, we have selected the list of singlet−triplet transitions recently gathered together by Thiel's group.[10,23] This set includes 20 small- and medium-sized compounds and a total of 63 excited states that have been treated with CAS-PT2, CC2, and CC3, theoretical best estimates have also been determined. For the transitions under scrutiny in this letter, these best estimates generally correspond to the CC3 values. All calculations reported here have been performed with the Gaussian suite of programs, with a tight self-consistent field convergence threshold ($10^{-8}-10^{-10}$ au), using both commercial and development versions.[24−26] The MP2/6-31G(d) geometries were recovered from ref 23, and the eight-to-twenty first vertical triplet excited states have been computed with TD-DFT/TZVP. As demonstrated in Section 4, in which EOM-CCSD and TD-DFT calculations are performed with larger basis sets [6-311+G(2d,p), 6-311++G(3df,3pd), and aug-cc-pVTZ], several transition energies are significantly altered by including diffuse functions, but the statistical impact of using larger basis sets remains limited.

As we assess the performances of DFT approaches, we have selected a large panel of functionals, including local density approximation (LDA), generalized gradient approximation (GGA), meta-GGA, global hybrids (GH), and range-separated hybrids (RSH): SVWN5, BLYP, BP86, OLYP, PBE, M06-L, VSCX, τ-HCTCH, TPSS, TPSSh, O3LYP, τ-HCTCH-hyb, B3LYP, B3P86, X3LYP, B98, PBE0, mPW1PW91, M06, M05, BMK, BHHLYP, M06-2X, M05-2X, M06-HF, LC-ωPBE (20), LC-BLYP, LC-OLYP, LC-PBE, LC-τHCTH, LC-TPSS, LC-ωPBE, and CAM-B3LYP. We have also included the CIS approach for comparison purposes. We refer the reader to ref 12 for appropriate bibliographic informations for these functionals.

3. Results and Discussion

The transition energies obtained for all molecules and functionals as well as statistical analysis are catalogued in the Supporting Information. For the extensive benchmark calculations presented here, it is probably useless to discuss the computed spectra molecule-per-molecule, as such specific analysis is already available for three typical functionals in ref 10. Consequently, we will focus here on global results, allowing to unravel general trends, thanks to statistical analysis.

Table 1. Statistical Analysis for the Full Set of Transitions, Using the Theoretical Best Estimates As References[a]

functional	MSE	MAE	SD	rms	*b*	R^2
CIS	0.12	0.56	0.37	0.67	1.11	0.82
SVWN5	0.35	0.42	0.37	0.56	0.95	0.88
BLYP	0.48	0.49	0.33	0.58	0.96	0.93
BP86	0.49	0.49	0.32	0.58	0.97	0.93
OLYP	0.46	0.46	0.32	0.50	0.97	0.93
PBE	0.50	0.50	0.33	0.60	0.96	0.93
M06-L	0.37	0.38	0.24	0.45	0.98	0.96
VSXC	0.43	0.43	0.25	0.49	0.98	0.96
τ-HCTH	0.55	0.55	0.25	0.61	0.98	0.96
TPSS	0.49	0.49	0.26	0.55	0.98	0.96
TPSSh	0.51	0.51	0.23	0.55	0.98	0.97
O3LYP	0.43	0.43	0.22	0.49	0.98	0.97
τ-HCTH-hyb	0.42	0.42	0.20	0.46	0.99	0.97
B3LYP	0.44	0.44	0.19	0.48	0.99	0.98
B3P86	0.45	0.45	0.19	0.49	0.99	0.98
X3LYP	0.44	0.44	0.19	0.48	0.99	0.98
B98	0.37	0.37	0.20	0.42	1.00	0.98
PBE0	0.49	0.49	0.21	0.54	0.99	0.97
mPW1PW91	0.51	0.51	0.22	0.55	0.99	0.97
M06	0.44	0.44	0.18	0.47	0.99	0.98
M05	0.70	0.70	0.34	0.78	0.98	0.95
BMK	0.18	0.24	0.16	0.28	0.99	0.97
BHHLYP	0.54	0.58	0.46	0.73	0.95	0.90
M06-2X	0.07	0.23	0.17	0.28	0.98	0.94
M05-2X	0.22	0.27	0.20	0.33	0.98	0.96
M06-HF	−0.06	0.44	0.26	0.51	1.02	0.87
LC-ωPBE (20)	0.45	0.45	0.21	0.49	1.00	0.97
LC-BLYP	0.34	0.36	0.19	0.41	1.05	0.97
LC-OLYP	0.32	0.35	0.20	0.40	1.06	0.97
LC-PBE	0.37	0.40	0.21	0.44	1.08	0.97
LC-τ-HCTH	0.49	0.50	0.31	0.58	1.08	0.95
LC-TPSS	0.39	0.42	0.25	0.49	1.10	0.97
LC-ωPBE	0.50	0.55	0.38	0.66	1.17	0.93
CAM-B3LYP	0.41	0.42	0.24	0.48	1.05	0.97

[a] MSE is the mean signed error (reference-TD-DFT), MAE is the mean absolute error, SD is the standard deviation, and RMS is the residual mean squared error. MSE, MAE, SD, and RMS are in eV, and *b* and R^2 are the slope and the squared correlation coefficient, respectively, obtained through an unconstrained least-square linear fit.

The computed mean signed error (MSE), MAE, standard deviation (SD), root-mean-square deviation (rms) as well as the slope (*b*) and R^2 determined by unconstrained linear fittings that can be found in Table 1, whereas Figure 1 provides error profiles for a selection of functionals (VSXC, PBE0, M06-2X, and CAM-B3LYP). In the Supporting Information, the reader will find tables with MSE, MAE, rms, and R^2, using CAS-PT2 or CC3 values as benchmarks rather than the "best theoretical estimates". While the average errors and correlation coefficients differ from the results listed in Table 1, the discrepancies remain limited to a typical ±0.04 eV for the MSE, MAE, and rms and ±0.02 for the R^2. The similarities between the CAS-PT2 and CC3 transition energies are clear in ref 10, and we consequently only use the "best estimates" as reference values in the following, except when explicitly noted.

As can be seen in Table 1, the MSE values are systematically positive, but for M06-HF, that includes 100% of exact exchange, indicating that DFT functionals tend to almost systematically underestimate the transition energies. For the traditional GH, incorporating 20% and 30% of HF-like exchange, the errors are large (e.g., PBE0 in Figure 1) with typical deviations larger than 0.40 eV, except for B98 (0.37

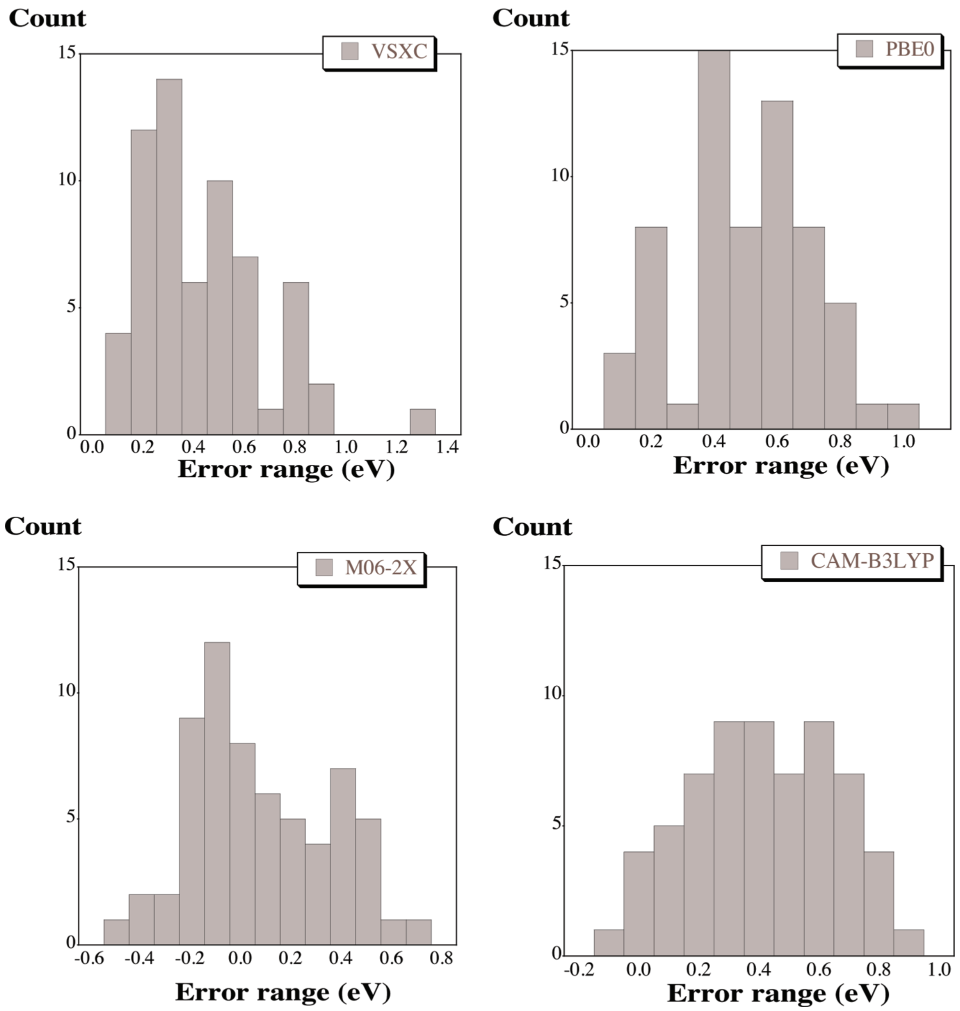


Figure 1. Histogram of the errors (eV) computed for four representative functionals.

eV). In fact, the M06-L meta-GGA outperforms most traditional hybrids on the MSE criterion. In the present case, the selection of range-separated methodologies does not help improve the estimates; the MSE values are systematically larger than 0.35 eV for the eight RSH functionals (see CAM-B3LYP histogram in Figure 1). It turns out that only four schemes yield |MSE| below the 0.2 eV threshold: BMK, M06-2X, M06-HF, and CIS. This finding strongly contrasts with singlet excited-states for which GH with 22−25% of exact exchange produces the smallest MSE.[12] The MAE delivered by most functionals are similar to that of the MSE, owing to the systematic overestimation effect. Only three theoretical schemes lead to average absolute errors smaller than 0.3 eV, namely: BMK (0.24), M06-2X (0.23), and M05-2X (0.27 eV). For the record, the more refined and computationally demanding DFT/MR-CI approach provides a similar MAE (0.25 eV) for the same set of molecules.[10] It is striking that the CIS scheme, characterized by a small MSE, yields the largest MAE, whereas the BHHLYP scheme produces much larger deviations (MAE of 0.58 eV) than functionals including a similar amount of exact exchange, like BMK or M06-2X. This indicates that the HF/DFT mixing is not the only major parameter governing the response of hybrids contrary to singlet−singlet transitions for which this parameter mostly guides the final answer. Such a statement is confirmed by comparing the B98/X3LYP and M06/M05 columns in the Supporting Information; the computed transition energies significantly differ though these functionals rely on very similar exact exchange ratios. In what concerns the consistency of the computed values, one notes that CIS provides the poorest agreement with the reference data ($R^2 = 0.82$). The SVWN5 LDA also yields a poor R^2 (0.88), while all GGA (meta-GGA) deliver correlation coefficients of 0.93 (0.96), clearly indicating that climbing Jacob's ladder of functionals improves the consistency of the results, although it might be detrimental for the average errors. Most tested GH values are characterized by large correlation coefficients (0.97−0.98) and by slopes reasonably close to 1. It is worth noting that the M06-2X R^2 of 0.94 increases to 0.96 when CAS-PT2 and CC3 reference data are selected. Nevertheless, BMK apparently grants slightly more consistent singlet−triplet energies than M06-2X. Eventually, we note that the R^2 of all RSH are similar to that of global hybrids, except for LC-ωPBE ($R^2 = 0.93$). This result is consistent with the larger damping parameter (0.40) used in LC-ωPBE; it delivers results almost identical to GH, including a large share of exact exchange. In short, this investigation clearly demonstrates that the optimal functionals within the TD-DFT framework significantly differ for singlet and triplet excited states. Indeed, for the latter, BMK (see Figure 2) and M06-2X emerge as the two most promising approaches, whereas the performance of BMK is

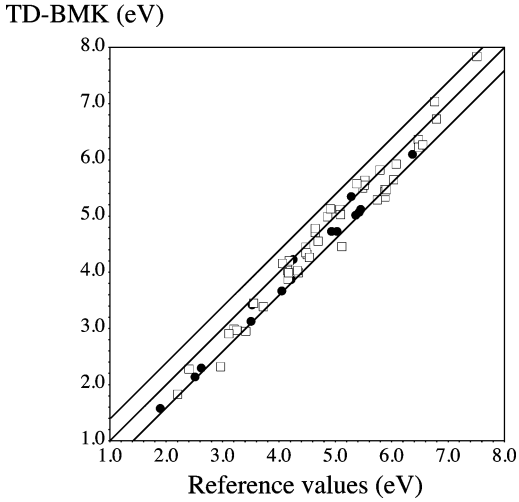


Figure 2. Comparisons between TD-BMK and theoretical best estimates (ref 10) for singlet−triplet transitions. The open squares (closed circles) correspond to $\pi \rightarrow \pi^\star$ ($n \rightarrow \pi^\star$) transitions. The central line indicates a perfect match, whereas the two side lines are border for ±0.4 eV deviations.

relatively modest for singlet states.[12] These two functionals provide a precision that is not far from that obtained with refined TD-DFT approaches including spin-flipping transitions,[27] as it has been shown that also in this case, large HF exchange contributions are needed to get accurate values. Therefore, it could be argued that the best-performing functionals, all containing a very large percent of HF exchange, could somehow handle such problems.

We have also performed a statistical analysis limited to the first singlet−triplet transition of each molecule, as TD-DFT is often viewed as particularly well-suited for describing low-lying states. For the present set of molecules, the R^2 improves for the majority of functionals, but no systematic decrease (wrt the full set) of the MAE or the rms is oberved. Indeed, the MAE computed for pure (RSH) functionals significantly decrease (increase) with relatively trifling variations for GH (but M05 and BHHLYP). For instance, as the BLYP MAE goes from 0.49 eV (full set) to 0.38 eV (first states), the B3LYP MAE changes by a negligible +0.01 eV, whereas the LC-BLYP MAE raises from 0.36 eV to 0.44 eV. In fact, the two most effective functionals remain BMK and M06-2X, with respective MAE of 0.26 and 0.20 eV. In the Supporting Information, tables specific to the $n \rightarrow \pi^\star$ and $\pi \rightarrow \pi^\star$ states are also given. For the first family of transitions, estimates are particularly sensitive to the actual form of the functional, as expected. For instance, the MAE of two meta-GGA, M06-L, and TPSS are similar (0.40 eV and 0.44 eV) for the full set but differ by more than 50% (0.37 eV and 0.63 eV) for $n \rightarrow \pi^\star$ transitions. For these states, a much larger correlation coefficient is obtained. Indeed, all hybrids (but BHHLYP) present a R^2 of 0.98 or 0.99 for the $n \rightarrow \pi^\star$ transitions. This is fully consistent with our previous works for singlet states.[12,28] Surprisingly, the BHHLYP MAE is minimal (0.21 eV), though BMK and M06-2X still perform satisfactorily (MAE of 0.27 and 0.26 eV, respectively, see also Figure 2). The most striking difference wrt the full set is probably the improved accuracy of RSH for $n \rightarrow \pi^\star$ excitations. The error patterns for $\pi \rightarrow \pi^\star$ are similar to that of the full set, with the M06−2X, BMK, and M05-2X leading to the smallest average deviations.

Table 2. Investigation of the Basis Set Effects for Three Typical Functionals[a]

		EOM-CCSD			CC3	
functional		TZVP	6-311+G (2d,p)	6-311++G (3df,3pd)	TZVP	BE
B3LYP	MSE	0.50	0.47	0.47	0.47	0.44
	MAE	0.50	0.47	0.47	0.47	0.44
	SD	0.22	0.18	0.18	0.18	0.19
	rms	0.54	0.50	0.50	0.51	0.48
	b	0.97	0.98	0.99	1.00	0.99
	R^2	0.97	0.98	0.98	0.98	0.98
M06-2X	MSE	0.13	0.13	0.13	0.11	0.07
	MAE	0.26	0.23	0.22	0.23	0.23
	SD	0.17	0.16	0.16	0.18	0.17
	rms	0.31	0.28	0.27	0.29	0.28
	b	0.97	0.98	0.99	1.00	0.99
	R^2	0.95	0.96	0.96	0.96	0.94
CAM-B3LYP	MSE	0.47	0.45	0.43	0.45	0.41
	MAE	0.47	0.45	0.43	0.45	0.42
	SD	0.18	0.18	0.18	0.24	0.24
	rms	0.50	0.48	0.47	0.51	0.48
	b	1.04	1.05	1.06	1.06	1.05
	R^2	0.98	0.98	0.98	0.97	0.97

[a] The same basis set is used for wavefunction and TD-DFT calculations. BE is the theoretical best estimates. See caption of Table 1 for more details.

4. Basis Set Effects

The results presented in Section 3 rely on the TZVP basis set for both the tested TD-DFT approaches and the wave function references. For sure, one expects that using larger diffuse-containing basis sets would induce significant variations of the computed transition energies, especially for the wave function schemes. For this reason, we have used Gaussian09 to perform EOM-CCSD calculations with three more extended basis sets, namely 6-311+G(2d,p), 6-311++G(3df,3pd), and aug-cc-pVTZ[29] (see Table 4 in the Supporting Information). This choice of EOM-CCSD as a reference method is justified because, according to Thiel,[23] all CC schemes behave well for singlet−triplet states. This statement is confirmed by Table 2 in which the statistical data collected using CC3/TZVP or EOM-CCSD/TZVP references are typically within 0.03 eV of each other.

As expected, using diffuse-containing basis sets tends to decrease the computed transition energies, though in most cases, the effect is relatively limited. Indeed, taking the aug-cc-pVTZ values as reference, we have computed EOM-CCSD mean absolute variation of 0.07, 0.02, and 0.01 eV for TZVP, 6-311+G(2d,p), and 6-311++G(3df,3pd), respectively (see Table 4 in the Supporting Information). Basically, 6-311+G(2d,p) provides converged EOM-CCSD results, and the largest deviation with respect to the aug-cc-pVTZ results is limited to −0.12 eV (A″ state of imidazole); the second largest discrepancy being as small as −0.06 eV (B_{2u} state of benzene). It is certainly appropriate to state that the EOM-CCSD/6-311++G(3df,3pd) energies are almost free of any basis set influence, at least for the transitions investigated herein. In Table 2, we perform a statistical analysis for three hybrid functionals using larger basis sets for both TD-DFT and wave function approaches. As can be seen, using larger basis sets tends to slightly decrease the average errors of all three functionals, but this effect remains small and does not affect the relative performances of each functional, the variations being similar for the three hybrids. For instance the B3LYP, M06-2X and CAM-B3LYP MAE are 0.50, 0.26, and 0.47 eV, respectively, with TZVP and become 0.47, 0.22, and 0.43 eV with 6-311++G(3df,3pd). These minor variations originate in almost parallel evolution of the transition energies at the TD-DFT and EOM-CCSD levels. The states that are strongly basis set dependent (or unaffected by the size of the basis set) at the EOM-CCSD level are the same at the TD-DFT level. For instance, the sensitive (insensitive) A″ state of imidazole (B_2 state of pyrrole) varies by −0.74 eV (−0.04 eV) when shifting from TZVP to aug-cc-pVTZ at the EOM-CCSD level, while the shift is −0.76 (−0.03), −0.64 (−0.10), and −0.48 eV (−0.03 eV) at the B3LYP, M06−2X, and CAM-B3LYP levels, respectively.

5. Conclusions

Using the set of “best available” wave function values provided by Thiel, we have benchmarked more than 30 density functional theory (DFT) functionals in the framework of time-dependent density functional theory (TD-DFT) evaluations of singlet−triplet transition energies. It turned out that: (i) for most functionals, the average deviations are larger than for singlet excited states; (ii) hybrids relying on similar exact exchange proportion but with different exchange−correlation forms might deliver significantly different values, especially for $n \rightarrow \pi^\star$ transitions; (iii) BMK and M06-2X allow taking the inner track to accurate estimates with MAE close to 0.25 eV in all cases, whereas B3LYP and PBE0 deviations are typically larger than 0.40 eV; (iv) range-separated formalism yield large errors for $\pi \rightarrow \pi^\star$ transitions; (v) these noted trends hold for the lowest-lying states, whereas substantial differences between $n \rightarrow \pi^\star$ and $\pi \rightarrow \pi^\star$ have been unravelled; and (vi) though a few transition energies are strongly affected by basis set effects, these “statistical” conclusions pertain for larger basis sets.

We are currently considering larger molecules and experimental benchmarking of DFT functionals.

Acknowledgment. D.J. and E.A.P. thank the Belgian National Fund for Scientific Research for their research associate and senior research associate positions, respectively. The authors thank Prof. G.E. Scuseria for the use of LCH. G03 and G09 calculations have been performed on the Interuniversity Scientific Computing Facility (ISCF), installed at the Facultés Universitaires Notre-Dame de la Paix (Namur, Belgium), for which the authors gratefully acknowledge the financial support of the FNRS-FRFC and the “Loterie Nationale” for the convention number 2.4578.02 and of the FUNDP. The collaboration between the Belgian and French group is supported by the Wallonie-Bruxelles International, the Belgian Fonds de la Recherche Scientfique, the Ministere Français des Affaires étrangeres et européennes, and the Ministere de l’Enseignement supérieur et de la Recherche in the framework of Hubert Curien Partnership.

Supporting Information Available: Tables with all transition energies, statistical data, and basis set study. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997–1000.

(2) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218–8224.

(3) Casida, M. E. In *Time-dependent density-functional response theory for molecules*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; Vol. 1, pp 155−192.

(4) Perdew, J. P.; Ruzsinsky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. *J. Chem. Phys.* **2005**, *123*, 062201.

(5) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009–4037.

(6) Barone, V.; Polimeno, A. *Chem. Soc. Rev.* **2007**, *36*, 1724–1731.

(7) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C. *Acc. Chem. Res.* **2009**, *42*, 326–334.

(8) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 044118.

(9) Rohrdanz, M. A.; Herbert, J. M. *J. Chem. Phys.* **2008**, *129*, 034107.

(10) Silva-Junior, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *129*, 104103.

(11) Goerigk, L.; Moellmann, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4611–4620.

(12) Jacquemin, D.; Wathelet, V.; Perpete, E. A.; Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420–2435.

(13) Paddon-Row, M. N.; Shephard, M. J. *J. Phys. Chem. A* **2002**, *106*, 2395–2944.

(14) Chong, D. P. *J. Electron Spectrosc. Relat. Phenom.* **2005**, *148*, 115–121.

(15) Santoro, F.; Improta, R.; Lami, A.; Bloino, J.; Barone, V. *J. Chem. Phys.* **2007**, *126*, 184102.

(16) Lanzo, I.; Russo, N.; Sicilia, E. *J. Phys. Chem. B* **2008**, *112*, 4123–4130.

(17) Caricato, M.; Trucks, G. W.; Frisch, M. J.; Wiberg, K. B. *J. Chem. Theory Comput.* **2010**, *6*, 370–383.

(18) Grimme, S.; Neese, F. *J. Chem. Phys.* **2007**, *127*, 154116.

(19) Tozer, D. J.; Handy, N. C. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2117–2121.

(20) Shao, Y.; Head-Gordon, M.; Krylov, A. I. *J. Chem. Phys.* **2003**, *118*, 4807–4818.

(21) Wang, F.; Ziegler, T. *J. Chem. Phys.* **2004**, *121*, 12191–12196.

(22) Nguyen, K. A.; Kennel, J.; Pachter, R. *J. Chem. Phys.* **2002**, *117*, 7128–7136.

(23) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110.

(24) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.;

Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revisions D.02 and E.01, Gaussian, Inc.: Wallingford, CT, 2004.

(25) Gaussian DV, Revision G.01; Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian, Inc.: Wallingford, CT, 2009.

(26) Frisch, M. J.; Trucks, G. W.; Schlegel, H.s B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. et al. *Gaussian 09* , revision A.2, Gaussian, Inc.: Wallingford, CT, 2009.

(27) Wang, F.; Ziegler, T. *J. Chem. Phys.* **2005**, *122*, 074109.

(28) Jacquemin, D.; Perpete, E. A.; Vydrov, O. A.; Scuseria, G. E.; Adamo, C. *J. Chem. Phys.* **2007**, *127*, 094102.

(29) For the latter basis sets, the two largest molecules (octatetraene and naphthalene) were beyond computational reach.

CT100005D

JCTC Journal of Chemical Theory and Computation

Updated Branching Plane for Finding Conical Intersections without Coupling Derivative Vectors

Satoshi Maeda,[†,‡,§] Koichi Ohno,[*,‡] and Keiji Morokuma[*,†,§]

Fukui Institute for Fundamental Chemistry, Kyoto University, Kyoto 606-8103, Japan, Toyota Physical and Chemical Research Institute, Nagakute, Aichi 480−1192, Japan, and Department of Chemistry and Cherry L. Emerson Center for Scientific Computation, Emory University, Atlanta, Georgia 30322

Received January 14, 2010

Abstract: The conical intersections (CIs) form a (*f*-2)-dimensional hyperspace on which two diabatic potential energy surfaces (PESs) belonging to the same symmetry cross, where *f* is the internal degree of freedom. The branching plane (BP) is a (two-dimensional) plane defined by the difference gradient vector (DGV) and the coupling derivative vector (CDV), and on the BP, the degeneracy of the two adiabatic PESs is lifted. The properties of the BP are often used in the exploration of the conical intersection hyperspace, such as determination of the minimum energy CI or the first-order saddle point in CI. Although both DGV and CDV are necessary to construct the BP in general, CDV is not always available depending on ab initio methods and programs. Therefore, we developed an approach for optimizing critical points on the CI hypersurface without CDV by using a BP updating method, which was shown to be accurate and very useful for minimum energy and saddle point optimization and for the minimum energy path following within the CI hypersurface in numerical tests for C_6H_6 and $C_5H_8N^+$.

1. Introduction

The conical intersections (CIs) form a (*f*-2)-dimensional hyperspace on which two diabatic potential energy surfaces (PESs) belonging to the same symmetry cross, where *f* is the internal degree of freedom. Conical intersections have been explored in a number of studies on photochemical and ion−molecule reactions, as regions where nonadiabatic transitions take place efficiently.[1−4] Especially, the minimum energy conical intersection (MECI) point is considered to be a critical point for nonadiabatic transition. Hence, there have been considerable efforts for developing efficient MECI optimization algorithms.[5−11] Furthermore, an automated systematic exploration method for MECIs has very recently been developed.[12] Recently, the first-order saddle point in CI hypersurface and the corresponding minimum energy path (MEP) were proposed to be important in dynamical trajectory simulations, and an optimization method was developed for such high-energy points within the CI hypersruface.[9]

The branching plane (BP) is a (two-dimensional) plane defined by the difference gradient vector (DGV) and the coupling derivative vector (CDV), and on the BP, the degeneracy of the two adiabatic PESs is lifted. The properties of BP are often used in the exploration of the CI hyperspace. Some of MECI optimizers use BP to keep degeneracy of two adiabatic states during optimizations.[6−9] The method for characterizing higher energy CI points also uses BP.[9] BP is required for finding transition directions in the generalized trajectory surface hopping method based on the Zhu−Nakamura theory,[13−15] in which the direction of CDV is estimated as the maximum eigenvalue direction of the difference Hessian matrix when CDV is not available.[13−15]

In order to use BP for optimization, both DGV and CDV vectors are necessary in every optimization step. DGV can be obtained easily from gradient vectors for two adiabatic PESs. If an analytical gradient is not available, it can be evaluated easily by numerical energy differentiation. However, CDV is

[*] Corresponding authors e-mail: ohnok@mail.tains.tohoku.ac.jp (K.O.); morokuma@emory.edu (K.M.).

[†] Kyoto University.

[‡] Toyota Physical and Chemical Research Institute.

[§] Department of Chemistry and Cherry L. Emerson Center for Scientific Computation.

10.1021/ct1000268 © 2010 American Chemical Society
Published on Web 04/07/2010

not available for all ab initio methods, and programs since implementation of an analytical derivative method are required. When a new accurate ab initio method is developed, typically only the energy can be calculated, and the BP-based optimization method cannot be employed. In addition, it is sometimes better to avoid CDV calculations, if possible, because the cost for computing CDV is not negligible, especially for correlated ab initio methods.

To avoid CDV calculations, penalty function methods[10,11] have been developed and have been very useful for finding MECI regions using ab initio methods without CDV codes. However, convergence of these methods is, in general, slower than the BP-based method, especially if tight optimization for ($E_1 - E_2$) is desired.[16] Therefore, often a very loose convergence criterion allowing large energy differences (usually larger than 1 kJ/mol) is employed for these methods. On the other hand, in the constrained optimization methods,[5,6,8] the constraint of the CDV direction can be omitted when CDV is not available, and this approach has also been employed together with correlated multireference methods without CDV codes,[17−19] although its convergence is slow when the mean energy gradient vector is highly coupling with the CDV direction.

Another possible approach for finding CIs without CDV is updating BP by using DGV, as has been done with the Hessian updating for single energy optimization. Many excellent Hessian updating methods have been developed for conventional augmented Hessian geometry optimizers.[20−24] BP is composed of only two vectors, and its updating should be much easier than updating Hessian of N atom systems with $3N - 6$ vectors. In the present paper, we propose a very simple method for updating BP. Although this BP-updating method can be combined with any BP-based optimization approach, we use it in combination with the gradient projection method, developed by Bearpark et al.[7] as an MECI optimizer and extended very recently by Sicilia et al.[9] for finding higher energy CI points. We demonstrate that the present update method gives very accurate BPs in CI regions and is very useful and efficient in locating minimum energy points, saddle points, and minimum energy paths in numerical tests for C_6H_6 and $C_5H_8N^+$.

2. Methods

2.1. Updating Branching Plane. The adiabatic energies for state 1 (E_1) and state 2 (E_2) can be written with the diabatic energies (U_{11} and U_{22}) and with their coupling U_{12}.

$$\begin{aligned} E_1 &= \frac{1}{2}(U_{11} + U_{22}) - \frac{1}{2}\sqrt{(U_{11} - U_{22})^2 + 4U_{12}^2} \\ E_2 &= \frac{1}{2}(U_{11} + U_{22}) + \frac{1}{2}\sqrt{(U_{11} - U_{22})^2 + 4U_{12}^2} \end{aligned} \tag{1}$$

Either when $U_{11} - U_{22} = U_{12} = 0$ is satisfied (on a CI) or when one assumed that the diabatic energies and their coupling is linearly dependent on any coordinate at the point of interest (in the first-order approximation of U_{nm}), the second derivative of $(E_1 - E_2)^2$ is given by

$$\frac{\partial^2(E_1 - E_2)^2}{\partial x_i \partial x_j} = 2\left(\frac{\partial U_{11}}{\partial x_i} - \frac{\partial U_{22}}{\partial x_i}\right)\left(\frac{\partial U_{11}}{\partial x_j} - \frac{\partial U_{22}}{\partial x_j}\right) + 8\frac{\partial U_{12}}{\partial x_i}\frac{\partial U_{12}}{\partial x_j} \tag{2}$$

From eq 2, the second derivative matrix **H** of $(E_1 - E_2)^2$ can be written with two vectors **p** and **q** as

$$\mathbf{H} = 2\mathbf{p}\mathbf{p}^T + 8\mathbf{q}\mathbf{q}^T \tag{3}$$

where **p** is the DGV for diabatic energy with the component $\partial U_{11}/\partial x_i - \partial U_{22}/\partial x_i$ and **q** is the CDV for diabatic energy with the component $\partial U_{12}/\partial x_i$. Thus, **H** is given as a function of two vectors **p** and **q** and defines the BP. Here, adiabatic DGV and CDV can be written in terms of a linear combination of **p** and **q**, since BP does not change by a diabatic to adiabatic transformation. Although diabatic DGV and CDV and adiabatic DGV and CDV are used together in the following explanation, BP can be defined by any combination of these four vectors because they all are vectors on a common BP.

We express BP at the kth optimization step by two vectors $\mathbf{x}_k$ and $\mathbf{y}_k$, where these are an unit vector parallel to DGV for adiabatic energy at the kth step and an unit vector on BP perpendicular to $\mathbf{x}_k$, respectively. Here, at the kth step, $\mathbf{x}_{k-1}$, $\mathbf{y}_{k-1}$, and $\mathbf{x}_k$ are known, and $\mathbf{y}_k$ is an unknown vector to be estimated by the BP updating method. In the first-order approximation ($\partial^2 U_{nm}/\partial x_i \partial x_j = 0$), it is obvious that BP (i.e., **p** and **q** in eq 3) does not change by any geometry displacement because $\partial U_{nm}/\partial x_i$, which are elements of **p** and **q** shown in eq 2, is independent of x_i. Thus the first-order BP at the kth step is nothing but a plane defined by $\mathbf{x}_{k-1}$ and $\mathbf{y}_{k-1}$. Let us consider $\mathbf{x}_k$ has been obtained exactly without the first-order approximation. Thus $\mathbf{x}_k$ may have a component not contained in $\mathbf{x}_{k-1}$ or $\mathbf{y}_{k-1}$ because of the higher order terms in determining $\mathbf{x}_k$. The value of $\mathbf{y}_k$ can be estimated by the unchanged first-order BP: such a $\mathbf{y}_k$ should be written by a linear combination of $\mathbf{x}_{k-1}$ and $\mathbf{y}_{k-1}$ as $\mathbf{y}_k = \alpha\mathbf{x}_{k-1} + \beta\mathbf{y}_{k-1}$. Since $\mathbf{y}_k$ is a unit vector orthogonal to $\mathbf{x}_k$, we get the following simultaneous equations for α and β:

$$\begin{aligned} \alpha(\mathbf{x}_{k-1}\cdot\mathbf{x}_k) + \beta(\mathbf{y}_{k-1}\cdot\mathbf{x}_k) &= 0 \\ \alpha^2 + \beta^2 &= 1 \end{aligned} \tag{4}$$

Then, by solving eq 4, we obtain $\mathbf{y}_k$ as

$$\mathbf{y}_k = \frac{(\mathbf{y}_{k-1}\cdot\mathbf{x}_k)\mathbf{x}_{k-1} - (\mathbf{x}_{k-1}\cdot\mathbf{x}_k)\mathbf{y}_{k-1}}{\sqrt{(\mathbf{y}_{k-1}\cdot\mathbf{x}_k)^2 + (\mathbf{x}_{k-1}\cdot\mathbf{x}_k)^2}} \tag{5}$$

This $\mathbf{y}_k$ is used together with $\mathbf{x}_k$ for constructing the updated BP at the kth step, and they are stored either in memory or in disk for the next step. At the initial step, one does not have the first-order BP (**x** and **y** at the last step). A plane for $\mathbf{x}_0$ and the mean energy gradient vector was used as an initial BP in this study. Such a BP is exact at stationary points in CIs, since the mean energy gradient vector does not contain any components perpendicular to BP at such points. Thus CDV is no longer necessary at every optimization step when this BP updating algorithm is employed. Although this scheme partly assumes the first-order approximation of U_{nm}, higher-order effects are accounted for by using the exact $\mathbf{x}_k$, and it worked very well in numerical tests shown below.

As shown in the following test calculations, *updated* $\mathbf{y}_k$ using eq 5 converges to a very accurate one around CI regions, even if an initial $\mathbf{y}_0$ is very poor. This can be understood by considering an optimization step around an apex of cone. When *updated* $\mathbf{y}_{k-1}$ is not very accurate and the mean energy gradient vector has a component of *true* $\mathbf{y}_{k-1}$, an optimization step there will have a component of *true* $\mathbf{y}_{k-1}$, in addition to a component of $\mathbf{x}_{k-1}$. The component of $\mathbf{x}_{k-1}$ minimizes the energy difference, whereas the component of *true* $\mathbf{y}_{k-1}$ increases the energy difference. In other words, the component of $\mathbf{x}_{k-1}$ heads toward the apex, whereas the component of *true* $\mathbf{y}_{k-1}$ leaves the apex behind. Consequently, $\mathbf{x}_k$ becomes very similar to *true* $\mathbf{y}_{k-1}$, and *true* $\mathbf{y}_k$ becomes very similar to $\mathbf{x}_{k-1}$. Here, except for the case where *updated* $\mathbf{y}_{k-1}$ is orthogonal to *true* $\mathbf{y}_{k-1}$, *updated* $\mathbf{y}_k$ by eq 5 becomes very similar to $\mathbf{x}_{k-1}$ and *true* $\mathbf{y}_k$. In the final stage, optimizations keep walking around the apex of cone, and $\mathbf{y}_k$ can be purified to a very accurate one by eq 5 because of accompanying rotations of $\mathbf{x}$ around the apex of cone. Here, based on this discussion, one can assume a simpler algorithm using a plane of $\mathbf{x}_{k-1}$ and $\mathbf{x}_k$ as an updated BP. However, this algorithm was numerically unstable in our tests when $\mathbf{x}_k$ is very similar to $\mathbf{x}_{k-1}$, and it can be stabilized by adding $\mathbf{y}_{k-1}$ to the plane of $\mathbf{x}_{k-1}$ and $\mathbf{x}_k$ with a certain amount as eq 5.

As can be seen in the above discussion, the present BP updating is a technique to obtain an accurate BP using rotations of $\mathbf{x}$ induced by a walk at a finite distance from CI. Hence, this does not work for special systems, such as symmetry required Jahn−Teller systems, because one can walk exactly on CI using symmetry. However, in such systems, BP is not required in optimizations because there is no need to keep geometry on CI by using BP. Moreover, the present BP updating should work in Jahn−Teller systems too when optimizations start from a lower symmetry point and when no symmetry related constraint is employed. On the other hand, no rotation of $\mathbf{x}$ occurs when a seam between two states with different symmetry or when spin-multiplicity as the norm of CDV is zero. Although the BP updating does not work in this case too, BP is not necessary because $\mathbf{x}$ is only direction to increase the energy difference. It follows that the present BP updating is expected to work many systems in which BP is required in optimizations.

2.2. The Gradient Projection Method. In the gradient projection method, the following gradient vector is employed in optimizations:[7,9]

$$\mathbf{g} = \mathbf{g'}_{\text{diff}} + \mathbf{P}\mathbf{g}_{\text{mean}} \tag{6}$$

where $\mathbf{g'}_{\text{diff}} = 2(E_1 - E_2)\mathbf{x}$, $\mathbf{g}_{\text{mean}}$ is the mean energy gradient vector, and $\mathbf{P}$ is the following projection matrix:

$$\mathbf{P} = \mathbf{1} - \mathbf{x}\mathbf{x}^T - \mathbf{y}\mathbf{y}^T \tag{7}$$

Here, a unit vector parallel to CDV is used instead of $\mathbf{y}$ in the original expression of $\mathbf{P}$,[7,9] which is identical to $\mathbf{P}$ of eq 7. On the CI hypersurface, the condition $|\mathbf{g'}_{\text{diff}}| = 0$ is fulfilled at all points, and points with $|\mathbf{g'}_{\text{diff}}| = |\mathbf{P}\mathbf{g}_{\text{mean}}| = 0$ correspond to stationary points. A geometry displacement toward the inverse direction of the gradient vector minimizes the $(E_1 + E_2)/2$ function in the $3N - 8$ dimensional intersection space and minimizes the $(E_1 - E_2)^2/\alpha$ function in the two-dimensional branching space, where α is the norm of DGV. The use of $(E_1 - E_2)^2/\alpha$, rather than $(E_1 - E_2)^2$ itself, is better because $1/\alpha$ serves as a parameter that scales energy units and weight and improves the performance.

Treatment of Hessian is very tricky when the gradient of eq 6 is combined to augmented Hessian methods, such as the Newton−Raphson method.[9] Here, we describe only a procedure we used in this study, which is similar to a treatment of Sicilia et al.[9] in CI regions. Among normal modes for augmented Hessian methods, $3N - 8$ are nonzero eigenvalue modes of the projected mean energy Hessian $\mathbf{P}\mathbf{H}_{\text{mean}}\mathbf{P}$. Another is $\mathbf{x}$ with corresponding eigenvalue 2α. Only these $3N - 7$ modes are used because optimization steps are always perpendicular to the remaining $\mathbf{y}$, as the gradient of eq 6 does not contain components of $\mathbf{y}$. This set of modes and eigenvalues gives exact second-order steps in CIs, as long as $\mathbf{H}_{\text{mean}}$ is exact.

It is easy to combine the gradient projection method with the minimum energy path (MEP) following methods for single PES.[25−28] We employed the second-order algorithm which was proposed by Page and McIver[26] and is employed in Gaussian09[29] for predictor steps.[28] Although combining the corrector step in Gaussian09 to the gradient projection method will improve performance of the MEP following, it is beyond the scope of this study. We simply used the above-mentioned set of normal modes and eigenvalues as normal modes for the second-order algorithm. Since such second-order steps sometime caused unacceptably large energy differences of >5 kJ/mol, $(E_1 - E_2)^2$ was minimized along DGV until $|E_1 - E_2| < 0.1$ kJ/mol was met.

2.3. Computation. In the present study, the rational function optimization (RFO) method[30] was employed in combination with the gradient projection method[7,9] for obtaining each optimization step. Such optimizations were performed in the Cartesian coordinates throughout. $\mathbf{H}_{\text{mean}}$ was updated by using $\mathbf{g}_{\text{mean}}$ at the current and last optimization steps, where combined BFGS[20] and SR1[21] methods[24] were used in the minimization and combined Powell[22] and SR1 methods[23] were employed in the saddle point optimization and the MEP following. All the test calculations have been performed at the SA-CASSCF level of theory, where a two π electron and two π orbital active space and an STO-3G basis set were employed unless mentioned. Energy, gradient, and CDV were computed by using the Gaussian09 programs.[29] Optimizations were considered to be converged when the following four conditions are met simultaneously: (i) the maximum gradient is smaller than 3.0×10^{-4} hartree Å^{-1}; (ii) the root-mean-square (rms) gradient is smaller than 2.0×10^{-4} hartree Å^{-1}; (iii) the maximum displacement is smaller than 1.5×10^{-3} Å; and (iv) the rms displacement is smaller than 1.0×10^{-3} Å. These algorithms were implemented in the GRRM program developed by the authors for

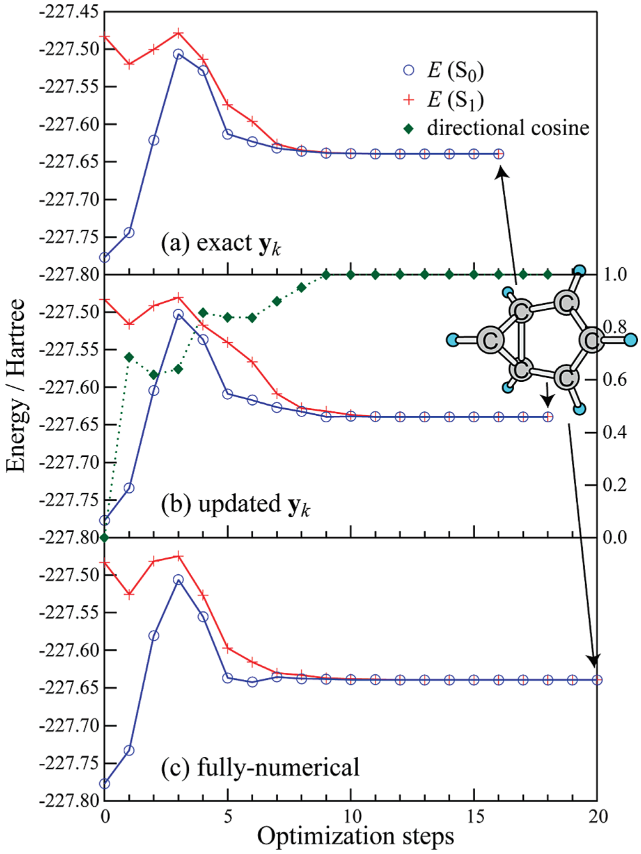


Figure 1. Energy profiles along three different optimizations for MECI of C_6H_6 (benzene): (a) using analytical DGV and exact $\mathbf{y}_k$, (b) using analytical DGV and updated $\mathbf{y}_k$, and (c) using numerical DGV and updated $\mathbf{y}_k$ (fully numerical). Energies for S_1 and S_0 states are plotted by + and ○, respectively. Plots (♦) of directional cosine values between exact and updated $\mathbf{y}_k$ during the optimization with updated $\mathbf{y}_k$ are superimposed on (b). Energy gaps at the final points in optimizations (a−c) are 0.0000, 0.0003, and 0.1187 kJ/mol, respectively.

automated global reaction route mapping on PESs,[31−33] and all geometry displacements were treated by the GRRM program.

3. Numerical Tests

3.1. Optimization of MECI. Tests are made for three known S_0/S_1 MECIs for C_6H_6 (benzene[34] and fulbene)[35] and $C_5NH_8^+$ (a model species of protonated retinal,[36] $CH_2{=}CH{-}CH{=}CH{-}CH{=}NH^+$).

Figure 1 shows energy profiles along three different optimizations for MECI of C_6H_6 (benzene): (a) using analytical DGV and exact $\mathbf{y}_k$; (b) using analytical DGV and updated $\mathbf{y}_k$; and (c) using numerical DGV and updated $\mathbf{y}_k$ (fully numerical), where numerical DGVs were evaluated by forward and backward single-point energy samplings in the Cartesian coordinates with a step size of 0.005 Å. The initial structure was prepared by a RHF/STO-3G constrained optimization, fixing one CCCC dihedral angle of benzene at 75°. The initial $\mathbf{H}_{\text{mean}}$ is the ground-state Hessian of the RHF/STO-3G method at the initial structure. Although (a) converged most quickly among the three, (b) and (c) are not very slow compared to (a) with only a few extra optimization steps. To see the accuracy of the updated BP in each optimization step, we plotted values of directional cosine between exact and updated $\mathbf{y}_k$ in Figure 1b, along the optimization profile of (b), where the value is unity if an updated $\mathbf{y}_k$ is exact, whereas it is zero if perpendicular to an exact $\mathbf{y}_k$. Although updated $\mathbf{y}_k$ was not very accurate at the initial stage because of a poor initial BP, it was improved substantially with optimization steps. Degeneracy of the two states was almost reached around the 10th step, and updated $\mathbf{y}_k$ was especially accurate from the 10th step with the values larger than 0.99. This is because eq 5 was formulated from eq 2 which is a second-order formula only at CIs. Inaccurate $\mathbf{y}_k$ in the initial stage of (b) remained only in two more optimization steps compared to (a). This is because $\mathbf{y}_k$ is not very important far from CIs, since adiabatic PESs have the cone-shaped square-root topology only around CI regions. The fully numerical optimization (c) is also performed to demonstrate that the present method allows BP-based optimizations using single-point energy only calculations, although this calculation gave a larger energy gap than those in (a) and (b) because of numerical errors in the numerical differentiations of single-point energies.

Fulbene, an isomer of benzene, is a good benchmark of CI optimizers since there are many low-lying critical points in its CI.[35] Although many early works[37,38] failed to locate true MECI and reported first- and second-order saddle-points to be MECI. The true MECI was discovered by very careful analyses in the CI hyperspace using the intersection-space Hessian approach.[35] Hence, in this test, we prepared different initial structures in the potential basin of fulbene to see the radius of convergence of the present optimizer. Here, we employed a six π electron and six π orbital active-space and the Dunning cc-pVDZ basis set so that obtained structures can be compared to the MECI reported in ref 35. In both optimizations, the present BP updating method was employed. Figure 2 shows energy profiles along two optimizations starting from two different structures. In the optimization (a), the initial structure was prepared by a RHF/STO-3G constrained optimization fixing a CC═CH dihedral angle at 75°, where CC of the CC═CH are atoms in the five-membered ring and the remaining CH are atoms in the ═CH$_2$ group of fulbene. In the optimization (b), the initial structure was prepared by a RHF/STO-3G constrained optimization fixing the CC═CH dihedral angle at 5°. The initial $\mathbf{H}_{\text{mean}}$ are ground-state Hessian of the RHF/STO-3G method at these initial structures. As seen in Figure 2, both optimizations converged to the same MECI with C_1 symmetry reported in ref 35, where average electronic energies of the final structures in optimizations (a) and (b) are −230.6513948 and −230.6513947 hartree, respectively, which are very similar to the reported value (−230.6513957) in ref 35. Although there is an energy bump in the profile (b), due to a significant change in molecular orbitals of CASSCF at the point, the optimization finally converged to the correct MECI. Among the two initial structures, one with the CC═CH dihedral angle equal to 75° is closer to the MECI than that of the other. Consequently, the optimization (a) converged much more quickly than (b). Although there is a planar second-order saddle-point with C_{2v} symmetry

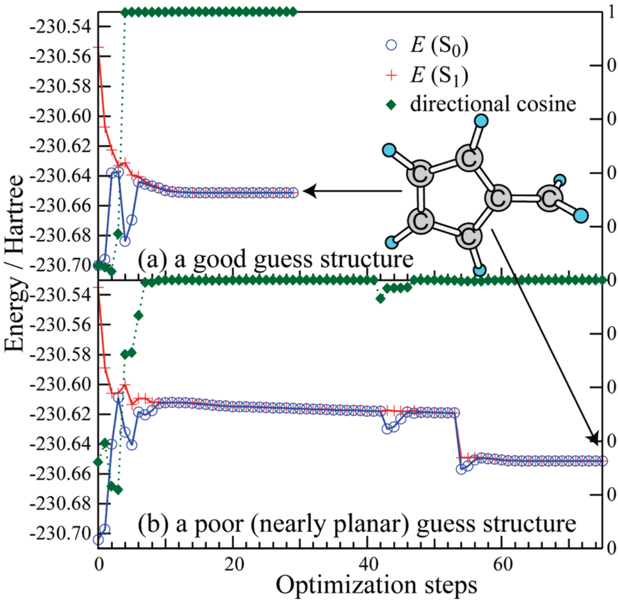


Figure 2. Energy profiles along two different optimizations for MECI of C_6H_6 (fulbene) using analytical DGV and updated $\mathbf{y}_k$: (a) starting from a good initial structure and (b) starting from a poor (nearly planar) initial structure. Energies for S_1 and S_0 states are plotted by + and ○, respectively. Plots (♦) of directional cosine values between exact and updated $\mathbf{y}_k$ are superimposed. Energy gaps at the final points in optimizations (a) and (b) are 0.0016 and 0.0026 kJ/mol, respectively.

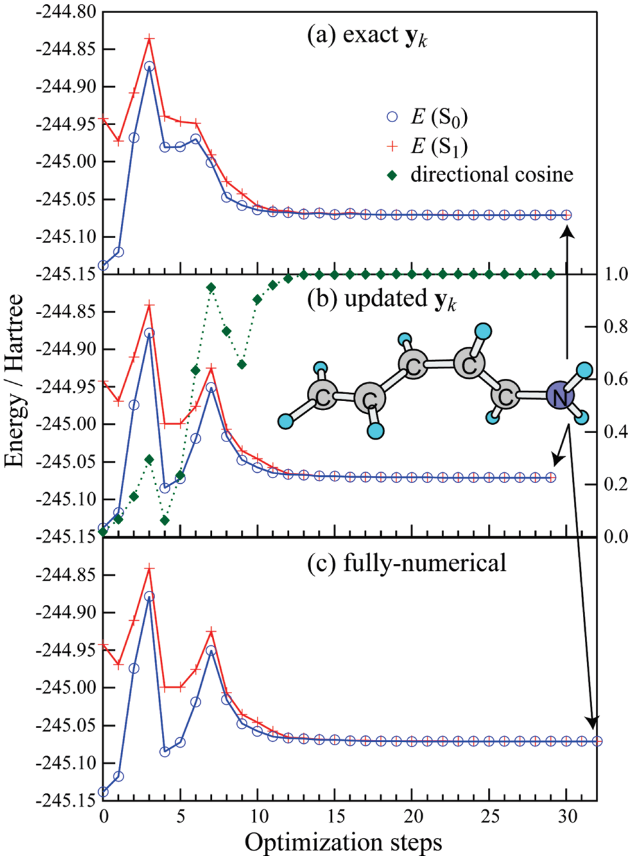


Figure 3. Energy profiles along three different optimizations for MECI of $C_5H_8N^+$: (a) using analytical DGV and exact $\mathbf{y}_k$, (b) using analytical DGV and updated $\mathbf{y}_k$, and (c) using numerical DGV and updated $\mathbf{y}_k$ (fully numerical). Energies for S_1 and S_0 states are plotted by + and ○, respectively. Plots (♦) of directional cosine values between exact and updated $\mathbf{y}_k$ during the optimization with updated $\mathbf{y}_k$ are superimposed on (b). Energy gaps at the final points in optimizations (a−c) are 0.0005, 0.0045, and 0.0819 kJ/mol, respectively.

close to the second initial structure,[35] the optimization (b) converged to the true MECI structure because of the proper step contorting by the RFO method. As seen in the figure, values of directional cosine between exact and updated $\mathbf{y}_k$ are very close to unity in CI regions.

Figure 3 shows results for $C_5NH_8{}^+$. Optimizations (a−c) are similar to those for benzene (Figure 1). The initial structure was prepared by RHF/STO-3G constrained optimizations fixing the central CCCC dihedral angle of $CH_2{=}CH{-}CH{=}CH{-}CH{=}NH^+$ at 75°. The initial $\mathbf{H}_{\mathrm{mean}}$ is ground-state Hessian of the RHF/STO-3G method at the initial structure. The numbers of optimization steps are similar among the three different optimizations. Although updated $\mathbf{y}_k$ was not accurate in the initial stage, it was very accurate after degeneracy was reached around the 15th step. The errors in the initial stage did not strongly affect the total number of optimization steps.

3.2. Determination of Saddle-Point and MEP within the CI Hypersurface. We performed a saddle-point optimization and a corresponding MEP calculation within the CI hypersurface for C_6H_6 to show that the BP updating method works also for higher energy CI points. The initial structure for the saddle-point optimization was prepared by a RHF/STO-3G constrained optimization fixing a CC bond length at 1.853 Å and a CCCC dihedral angle at 90°, where the CC bond is a newly generated bond for the three-membered ring in the MECI of Figure 1, and the CCCC dihedral angle is the one fixed in the initial structure preparation for the MECI optimization. Although this structure was prepared intending to find a saddle-point located in a reaction coordinate related to a deformation of C_6 backbone (mainly related to the CC dimer rotation), there is no such saddle-point, and another saddle-point was discovered as shown below. The initial $\mathbf{H}_{\mathrm{mean}}$ is the ground-state Hessian of the RHF/STO-3G method at the initial structure.

Figure 4 shows energy profiles along two different optimizations for a saddle point in CI of C_6H_6: (a) using analytical DGV and exact $\mathbf{y}_k$, and (b) using analytical DGV and updated $\mathbf{y}_k$. In this case, (b) accidentally converged slightly faster than (a). Although the initial $\mathbf{H}_{\mathrm{mean}}$ did not have any negative eigenvalue modes, these optimizations finally converged to a saddle point by walking along the lowest frequency mode. Hence, most parts of the profiles are uphill, and consequently, these optimizations needed as much as 130 steps. The gradient projection method found the CI region very quickly within 10 optimization steps, after which the updated $\mathbf{y}_k$ was very accurate. It should be noted that this example is demonstrating robustness of the RFO method in first-order saddle-point optimizations in the CI hypersurface, starting from a poor initial guess and an inaccurate $\mathbf{H}_{\mathrm{mean}}$. As has been shown in optimizations of single PESs, augmented Hessian methods, such as the RFO

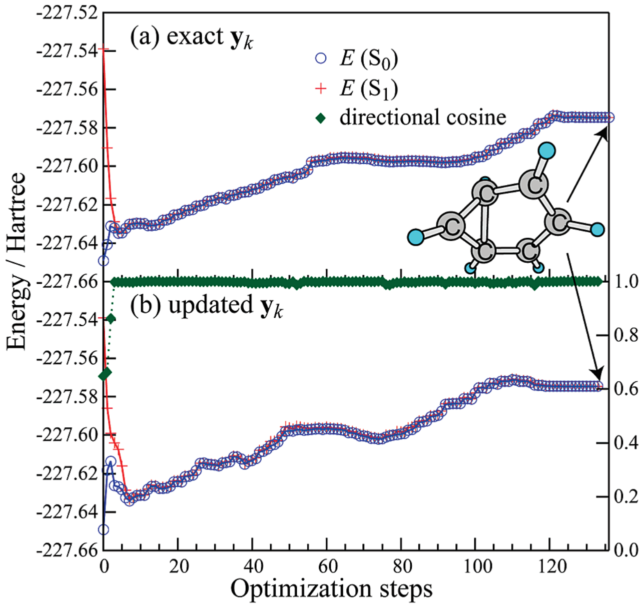


Figure 4. Energy profiles along two different optimizations for a saddle point in CI of C_6H_6: (a) using analytical DGV and exact $\mathbf{y}_k$ and (b) using analytical DGV and updated $\mathbf{y}_k$. Energies for S_1 and S_0 states are plotted by + and ○, respectively. Plots (♦) of directional cosine values between exact and updated $\mathbf{y}_k$ during the optimization with updated $\mathbf{y}_k$ are superimposed on (b). Energy gaps at the final points in optimizations (a) and (b) are 0.0021 and 0.0037 kJ/mol, respectively.

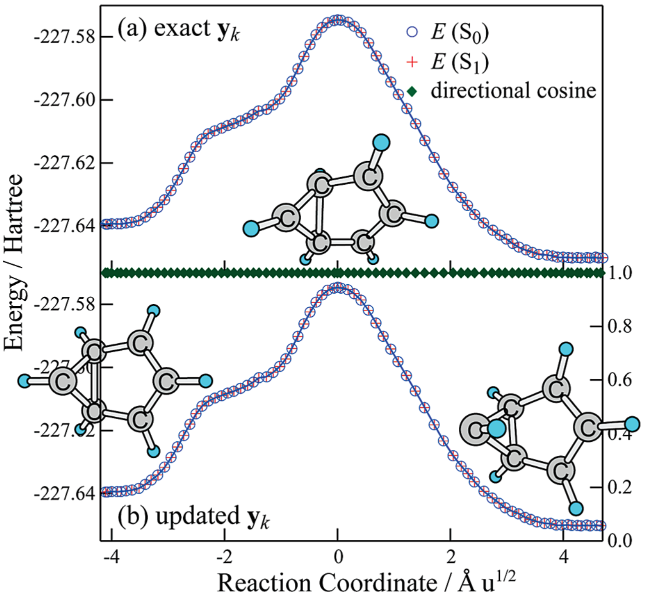


Figure 5. Energy profiles along a MEP in CI of C_6H_6 traced by two different algorithms: (a) using analytical DGV and exact $\mathbf{y}_k$ and (b) using analytical DGV and updated $\mathbf{y}_k$. Energies for S_1 and S_0 states are plotted by + and ○, respectively. Plots (♦) of directional cosine values between exact and updated $\mathbf{y}_k$ during the MEP calculation with updated $\mathbf{y}_k$ are superimposed on (b). Energy gaps at the end points of the MEPs are (a-left) 0.0018, (a-right) 0.0005, (b-left) 0.0029, and (b-right) 0.0011 kJ/mol.

method, greatly expand the radius of convergence of transition-state optimizations.[39]

Starting from the first-order saddle point within the CI hypersurface obtained by this optimization, a MEP was calculated in the mass weighted Cartesian coordinate, where a step size (in Å u$^{1/2}$) was adjusted so that a simple linear displacement along the inverse gradient vector is equal to 0.1 Å in each step. In the MEP following, $\mathbf{H}_{\mathrm{mean}}$ was computed once at the initial structure by numerical differentiations of $\mathbf{g}_{\mathrm{mean}}$ to define the negative eigenvalue mode to be followed. Here, the projected $\mathbf{H}_{\mathrm{mean}}$ ($\mathbf{PH}_{\mathrm{mean}}\mathbf{P}$) has only one negative eigenvalue mode, which is confirming that the point is a first-order saddle point in the CI hypersurface. Here, this procedure using $\mathbf{PH}_{\mathrm{mean}}\mathbf{P}$ is based on the first-order approximation of CI hyperspace, which may cause a large energy gap in the initial MEP integration step. Use of the intersection-space Hessian defined by Sicilia et al.[35,40] gives a more accurate initial reaction coordinate direction. Nevertheless, the present procedure gave a reasonable MEP passing through a space with a very small energy gap, as shown below because of the minimization of $(E_1 - E_2)^2$ in every integration point.

Figure 5 shows energy profiles along an MEP in the CI hypersurface of C_6H_6 by two different calculations, starting from the saddle point of Figure 4: (a) using analytical DGV and exact $\mathbf{y}_k$, and (b) using analytical DGV and updated $\mathbf{y}_k$. These two MEP calculations gave exactly the same connection which is between the MECI in Figure 1 and a lower energy MECI with a different CH direction. Energy differences between two states are always smaller than 0.1 kJ/mol along these profiles (overlapping in the figure) because of the minimizations of $(E_1 - E_2)^2$. The updated $\mathbf{y}_k$ was very accurate, and the directional cosine between exact and updated $\mathbf{y}_k$ was larger than 0.9999 at all points. Here, $\mathbf{y}_0$ at the initial step (at 0 Å u$^{1/2}$) was already very accurate with the directional cosine of 0.99994 because the initial BP for $\mathbf{x}_0$, and the mean gradient vector is exact at stationary points in the CI hypersurface.

4. Conclusion

We proposed a very simple method for updating branching plane (BP) by using a difference gradient vector (DGV) at the current position and (either exact or approximate) BP at the last position. In this study, we combined it with the gradient projection method to look into its performance not only around minimum energy conical intersections (MECIs) but also in higher energy CI points, although it is straightforward to combine it with other BP-based MECI optimizers in principle. In spite of a very simple assumption in the BP updating formula, updated BPs were very accurate in both low- and high-energy CI regions, as shown in numerical tests for C_6H_6 and $C_5H_8N^+$. Thus, the present method can be a powerful tool for finding CIs when the coupling derivative vector (CDV) is not available. Since the use of updated BPs did not increase the total numbers of optimization steps in the numerical tests, it can reduce computation demands for CDV calculations as well.

Acknowledgment. S.M. is supported by a Research Fellowship from the Japan Society for Promotion of Science for Young Scientists. This work is partly supported by a grant from the Japan Science and Technology Agency with a Core

Research for Evolutional Science and Technology (CREST) in the area of high performance computing for multiscale and multiphysics phenomena as well as a grant from AFOSR (Grant no. FA9550-07−1-0395). The computational resources at the Research Center for Computational Science at Institute for Molecular Science are gratefully acknowledged.

References

(1) Bernardi, F.; Olivucci, M.; Robb, M. A. Potential Energy Surface Crossings in Organic Photochemistry. *Chem. Soc. Rev.* **1996**, *25*, 321.

(2) Yarkony, D. R. Conical Intersections: Diabolical and Often Misunderstood. *Acc. Chem. Res.* **1998**, *31*, 511.

(3) Sobolewski, A. L.; Domcke, W.; Dedonder-Lardeux, C.; Jouvet, C. Excited-State Hydrogen Detachment and Hydrogen Transfer Driven by Repulsive $1\pi\sigma^*$ States: A New Paradigm for Nonradiative Decay in Aromatic Biomolecules. *Phys. Chem. Chem. Phys.* **2002**, *4*, 1093.

(4) Levine, B. G.; Martínez, T. J. Isomerization through Conical Intersections. *Annu. Rev. Phys. Chem.* **2007**, *58*, 613.

(5) Koga, N.; Morokuma, K. Determination of the Lowest Energy Point on the Crossing Seam between Two Potential Surfaces Using the Energy Gradient. *Chem. Phys. Lett.* **1985**, *119*, 371.

(6) Manaa, M. R.; Yarkony, D. R. On the Intersection of Two Potential Energy Surfaces of the Same Symmetry. Systematic Characterization Using a Lagrange Multiplier Constrained Procedure. *J. Chem. Phys.* **1993**, *99*, 5251.

(7) Bearpark, M. J.; Robb, M. A.; Schlegel, H. B. A Direct Method for the Location of the Lowest Energy Point on a Potential Surface Crossing. *Chem. Phys. Lett.* **1994**, *223*, 269.

(8) Anglada, J. M.; Bofill, J. M. A Reduced-Restricted-Quasi-Newton-Raphson Method for Locating and Optimizing Energy Crossing Points between Two Potential Energy Surfaces. *J. Comput. Chem.* **1997**, *18*, 992.

(9) Sicilia, F.; Blancafort, L.; Bearpark, M. J.; Robb, M. A. New Algorithms for Optimizing and Linking Conical Intersection Points. *J. Chem. Theor. Comput.* **2008**, *4*, 257.

(10) Ciminelli, C.; Granucci, G.; Persico, M. The Photoisomerization Mechanism of Azobenzene: A Semiclassical Simulation of Nonadiabatic Dynamics. *Chem.—Eur. J.* **2004**, *10*, 2327.

(11) Levine, B. G.; Coe, J. D.; Martínez, T. J. Optimizing Conical Intersections without Derivative Coupling Vectors: Application to Multistate Multireference Second-Order Perturbation Theory (MS-CASPT2). *J. Phys. Chem. B* **2008**, *112*, 405.

(12) Maeda, S.; Ohno, K.; Morokuma, K. Automated Global Mapping of Minimal Energy Points on Seams of Crossing by the Anharmonic Downward Distortion Following Method: A Case Study of H_2CO. *J. Phys. Chem. A* **2009**, *113*, 1704.

(13) Oloyede, P.; Mil'nikov, G.; Nakamura, H. Generalized Trajectory Surface Hopping Method Based on the Zhu-Nakamura Theory. *J. Chem. Phys.* **2006**, *124*, 144110.

(14) Nakamura, H. Dynamics of Nonadiabatic Chemical Reactions. *J. Phys. Chem. A* **2006**, *110*, 10929.

(15) Nakamura, H. Nonadiabatic Chemical Dynamics: Comprehension and Control of Dynamics, and Manifestation of Molecular Functions. *Adv. Chem. Phys.* **2008**, *138*, 95.

(16) Keal, T. W.; Koslowski, A.; Thiel, W. Comparison of Algorithms for Conical Intersection Optimisation Using Semiempirical Methods. *Theor. Chem. Acc.* **2007**, *118*, 837.

(17) Zhang, P.; Irle, S.; Morokuma, K.; Tschumper, G. S. Ab Initio Theoretical Studies of Potential Energy Surfaces in the Photodissociation of the Vinyl Radical. I. Ã State Dissociation. *J. Chem. Phys.* **2003**, *119*, 6524.

(18) Serrano-Andrés, L.; Merchán, M.; Lindh, R. Computation of Conical Intersections by Using Perturbation Techniques. *J. Chem. Phys.* **2005**, *122*, 104107.

(19) Zhang, P.; Maeda, S.; Morokuma, K.; Braams, B. J. Photochemical Reactions of the Low-Lying Excited States of Formaldehyde: T_1/S_0 Intersystem Crossings, Characteristics of the S_1 and T_1 Potential Energy Surfaces, and a Global T_1 Potential Energy Surface. *J. Chem. Phys.* **2009**, *130*, 114304.

(20) (a) Broyden, C. G. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *J. Inst. Math. Its Appl.* **1970**, *6*, 76. (b) Fletcher, R. A New Approach to Variable Metric Algorithms. *Comput. J. (Switzerland)* **1970**, *13*, 317. (c) Goldfarb, D. A Family of Variable-Metric Methods Derived by Variational Means. *Math. Comput.* **1970**, *24*, 23. (d) Shanno, D. F. Conditioning of Quasi-Newton Methods for Function. *Math. Comput.* **1970**, *24*, 647.

(21) Murtagh, B.; Sargent, R. W. H. Computational Experience with Quadratically Convergent Minimisation Methods. *Comput. J. (Switzerland)* **1972**, *13*, 185.

(22) Powell, M. J. D. Recent Advances in Unconstrained Optimization. *Math. Program.* **1971**, *1*, 26.

(23) Bofill, J. M. Updated Hessian Matrix and the Restricted Step Method for Locating Transition Structures. *J. Comput. Chem.* **1994**, *15*, 1.

(24) Farkas, Ö.; Schlegel, H. B. Methods for Optimizing Large Molecules. II. Quadratic Search. *J. Chem. Phys.* **1999**, *111*, 10806.

(25) Ishida, K.; Morokuma, K.; Komornicki, A. The Intrinsic Reaction Coordinate. An Ab Initio Calculation for HNC→HCN and $H^-+CH_4 \rightarrow CH_4+H^-$. *J. Chem. Phys.* **1977**, *66*, 2153.

(26) Page, M.; McIver, J. W., Jr. On Evaluating the Reaction Path Hamiltonian. *J. Chem. Phys.* **1988**, *88*, 922.

(27) Gonzalez, C.; Schlegel, H. B. An Improved Algorithm for Reaction Path Following. *J. Chem. Phys.* **1989**, *90*, 2154.

(28) Hratchian, H. P.; Schlegel, H. B. Using Hessian Updating to Increase the Efficiency of a Hessian based Predictor-Corrector Reaction Path Following Method. *J. Chem. Theory Comput.* **2005**, *1*, 61.

(29) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision A.1; Gaussian, Inc.: Wallingford, CT, 2009.

(30) Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. Search for Stationary Points on Surfaces. *J. Phys. Chem.* **1985**, *89*, 52.

(31) Ohno, K.; Maeda, S. A Scaled Hypersphere Search Method for the Topography of Reaction Pathways on the Potential Energy Surface. *Chem. Phys. Lett.* **2004**, *384*, 277.

(32) Maeda, S.; Ohno, K. Global Mapping of Equilibrium and Transition Structures on Potential Energy Surfaces by the Scaled Hypersphere Search Method: Applications to ab Initio Surfaces of Formaldehyde and Propyne Molecules. *J. Phys. Chem. A* **2005**, *109*, 5742.

(33) Ohno, K.; Maeda, S. Global Reaction Route Mapping on Potential Energy Surfaces of Formaldehyde, Formic Acid, and Their Metal-Substituted Analogues. *J. Phys. Chem. A* **2006**, *110*, 8933.

(34) Palmer, I. J.; Ragazos, I. N.; Bernardi, F.; Olivucci, M.; Robb, M. A. An MC-SCF Study of the S_1 and S_2 Photochemical-Reactions of Benzene. *J. Am. Chem. Soc.* **1993**, *115*, 673.

(35) Sicilia, F.; Bearpark, M. J.; Blancafort, L.; Robb, M. A. An Analytical Second-Order Description of the S_0/S_1 Intersection Seam: Fulvene Revisited. *Theor. Chem. Acc.* **2007**, *118*, 241.

(36) Garavelli, M.; Celani, P.; Bernardi, F.; Robb, M. A.; Olivucci, M. The $C_5H_6NH_2^+$ Protonated Shiff Base: An Ab Initio Minimal Model for Retinal Photoisomerization. *J. Am. Chem. Soc.* **1997**, *119*, 6891.

(37) Dreyer, J.; Klessinger, M. Excited States and Photochemical Reactivity of Fulvene. A Theoretical Study. *J. Chem. Phys.* **1994**, *101*, 10655.

(38) Deeb, O.; Cogan, S.; Zilberg, S. The Nature of the S_1/S_0 Conical Intersection of Fulvene. *Chem. Phys.* **2006**, *325*, 251.

(39) Schlegel, H. B. Exploring Potential Energy Surfaces for Chemical Reactions: An Overview of Some Practical Methods. *J. Comput. Chem.* **2003**, *24*, 1514.

(40) Sicilia, F.; Bearpark, M. J.; Blancafort, L.; Robb, M. A. Quadratic Description of Conical Intersections: Characterization of Critical Points on the Extended Seam. *J. Phys. Chem. A* **2007**, *111*, 2182.

CT1000268

JCTC Journal of Chemical Theory and Computation

Benchmark of Electronically Excited States for Semiempirical Methods: MNDO, AM1, PM3, OM1, OM2, OM3, INDO/S, and INDO/S2

Mario R. Silva-Junior and Walter Thiel*

*Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1,
45470 Mülheim an der Ruhr, Germany*

Received January 15, 2010

Abstract: Semiempirical configuration interaction (CI) calculations with eight different Hamiltonians are reported for a recently proposed benchmark set of 28 medium-sized organic molecules. Vertical excitation energies and one-electron properties are computed using the same geometries as in our previous ab initio benchmark study on electronically excited states. The CI calculations for the standard methods (MNDO, AM1, PM3) and for the orthogonalization-corrected methods (OM1, OM2, OM3) include single, double, triple, and quadruple excitations (CISDTQ) using the graphical unitary group approach (GUGA) as implemented in the MNDO code. The CIS calculations for the established INDO/S method and the reparametrized INDO/S2 variant employ a modified version of the ZINDO program. As compared to the best theoretical reference data from the ab initio benchmark, all currently applied semiempirical methods tend to underestimate the vertical excitation energies, but the errors are much larger in the case of the standard methods (MNDO, AM1, PM3). Overall, the mean absolute deviations relative to the theoretical best estimates are lowest for OM3, and only slightly higher for OM1 and OM2 (in the range of 0.4−0.5 eV). INDO/S performs similar to OM2 for the vertical excitation energies of singlet states, but deteriorates considerably for triplet states. The INDO/S2 reparametrization for oxygen improves the results for low-lying singlet states of oxygen-containing compounds, but makes them worse for high-lying singlets as well as for triplets. The ab initio reference data for oscillator strengths and excited-state dipole moments are again best reproduced by the orthogonalization-corrected approaches (OM1, OM2, OM3), which thus emerge as the most favorable semiempirical methods overall for treating valence excited states of large organic chromophores.

1. Introduction

In recent years, there has been much progress in the research on electronically excited states. Elaborate experimental techniques are available to study photophysical processes in the nanosecond or femtosecond regime. Concomitantly, improved theoretical methods have been developed that allow realistic calculations on excited states and may thus provide guidance for the experimental work. On the ab initio side, MS-CASPT2 (multistate complete-active-space second-order perturbation theory)[1−4] and coupled cluster methods (CC2, CCSD, CC3)[5−7] are well established and offer high accuracy for small molecules. Time-dependent density functional theory (TD-DFT)[8] has become popular for calculations on medium-sized molecules, giving reasonable results for various (but not all) types of excited states at relatively low computational cost.[9,10] An alternative DFT-based method makes use of Kohn−Sham orbitals in a multireference configuration interaction (MRCI) framework, modified by incorporating five universal empirical parameters to alleviate problems with the double counting of dynamic electron correlation.[11]

* Corresponding author e-mail: thiel@mpi-muelheim.mpg.de.

10.1021/ct100030j © 2010 American Chemical Society
Published on Web 03/31/2010

For a quantitative assessment of different theoretical approaches, reliable reference data are needed for benchmarking. Standard test sets are widely used for ground-state properties, for example, the G2 and G3 sets for thermochemistry.[12,13] We have recently introduced an ab initio benchmark set for electronically excited states of 28 medium-sized organic molecules with a total of 223 excitations.[14] On the basis of MS-CASPT2 and CC3 calculations on these molecules and of high-level ab initio data from the literature, we have proposed theoretical best estimates for the vertical excitation energies of 104 singlet and 63 triplet excited states. These reference data have been used to evaluate the performance of standard TD-DFT and DFT/MRCI approaches,[15] of the coupled cluster variant CCSDR(3) with noniterative triples corrections,[16] and of TD-DFT with a large number of different functionals[17,18] including double-hybrid functionals.[18]

Despite recent advances, the reliable description of electronically excited states in large molecules is still a challenging problem. Accurate ab initio methods such as MS-CASPT2 and CC3 are restricted to small molecules, and the computational cost for simpler treatments such as CC2 or DFT/MRCI still rises steeply with molecular size. TD-DFT is an attractive choice because of its computational efficiency and the availability of analytical gradients, but there are a number of well-documented problems of TD-DFT,[9,10] for example, with regard to change-transfer states[19] and singlet or triplet instabilities.[20] Moreover, the overall accuracy of TD-DFT is limited, with vertical excitation energies that typically show mean absolute deviations in the range of 0.3−0.5 eV from the theoretical best estimates in our benchmark set.[15] Given this situation, it seems worthwhile to explore the performance of semiempirical quantum-chemical methods for electronically excited states of large organic molecules.

Standard semiempirical molecular orbital (MO) methods such as MNDO,[21] AM1,[22] and PM3[23] are based on the NDDO (neglect of diatomic differential overlap) integral approximation and have been parametrized against ground-state properties, in particular heats of formation and geometries. They have been widely applied in computational studies of ground-state processes (for reviews, see, for example, refs 24−26). Applications to electronically excited states[27,28] are rare, however, mainly because these standard methods normally underestimate their energies strongly, as a result of the integral approximations and the ground-state parametrization. A straightforward remedy for this shortcoming would be a system-specific reparameterization[29] for a given application (see, for example, ref 30), which is, however, cumbersome in practice and also unsatisfactory from a conceptual point of view.

For the semiempirical calculation of vertical excitation energies, INDO/S (intermediate neglect of differential overlap for spectroscopy)[31,32] has been the method of choice for a long time. INDO/S describes excited states by CIS (configuration interaction with single excitations) and has been parametrized at this level. It has been widely used in studies of organic molecules[31,32] as well as transition metal complexes[33] and even lanthanides.[34] The INDO/S2 variant[35] is a reparametrization designed to improve the results for oxygen-containing compounds. The lack of higher excitations in the INDO/S CI treatment effectively restricts applications to states dominated by single excitations. Another limitation is the focus on vertical processes: by its design, INDO/S targets spectroscopy rather than photochemistry, and it is thus not made for the exploration of excited-state potential energy surfaces (PES).

The orthogonalization-corrected OM*x* methods (OM1,[36,37] OM2,[38,39] and OM3[40]) employ the NDDO integral approximation, but go beyond the standard methods (MNDO, AM1, PM3) by including additional terms in the Fock matrix that represent Pauli exchange repulsions in an approximate manner. These terms effectively raise the energy of antibonding virtual MOs and of the associated excited states.[37,39] Therefore, one would expect an improved performance of the OM*x* methods not only for ground-state properties,[41] but also for excited-state properties, which had not been taken into account during the OM*x* parametrization. The applications published so far support this view, for example, the OM2 studies on the electronically excited states of butadiene,[42] retinal model systems,[43] and the rhodopsin chromophore.[44] In addition, OM2 predicts reasonable geometries for a set of 12 typical conical intersections[45] (as compared to ab initio reference data). Finally, OM2 has successfully been applied in excited-state surface-hopping dynamics calculations for several small molecules,[46] for all nucleobases,[47−49] and for retinal models.[50] These promising indications call for a more comprehensive assessment, with detailed comparisons to established semiempirical treatments.

In this Article, we present a systematic evaluation of the performance of the standard NDDO-based semiempirical methods (MNDO, AM1, PM3), the commonly used INDO-based approaches (INDO/S, INDO/S2), and the orthogonalization-corrected methods (OM1, OM2, OM3) for electronically excited states. Reference data are taken from our previous benchmark work and comprise theoretical best estimates as well as MS-CASPT2/TZVP and CC3/TZVP data.[14] This Article is structured as follows: Section 2 describes the computational methods used. Section 3 presents some general considerations on the current benchmarking. Sections 4 and 5 discuss the individual results for vertical excitation energies and one-electron properties, respectively. Section 6 is devoted to statistical evaluations, and section 7 offers a brief summary and outlook.

2. Computational Methods

All calculations were carried out at the optimized ground-state equilibrium geometries reported previously.[14,51] The standard semiempirical Hamiltonians with default parameters were used for MNDO, AM1, PM3, OM1, OM2, and OM3, as implemented in the current version of the MNDO99 code.[52] In the case of INDO/S, singlet states were computed using the default parameters, with $f_{\pi\pi} = 0.585$ and the Mataga−Nishimoto expression for the two-center two-electron repulsion integrals, while triplet states were treated using the recommended special parametrization and the Pariser−Parr formula for the Coulomb integrals.[32] The

Table 1. OM2 Results for the Two Lowest Singlet Excited States of the Linear Polyenes with *k* Double Bonds, for Different Types of CI Treatment[a]

			MR-CISD[a]			
threshold (%)	FCI	CISDTQ	90	85	80	75
$k = 3$						
2^1A_g (eV)	4.86	4.86	4.86	4.87	4.87	4.89
1^1B_u (eV)	5.33	5.33	5.33	5.33	5.33	5.32
CSFs (references)[b]	175	165	162 (10)	150 (7)	146 (6)	126 (5)
CPU time (s)[c]	0.03	0.03	0.03	0.03	0.03	0.02
$k = 4$						
2^1A_g (eV)	4.19	4.20	4.21	4.22	4.25	4.25
1^1B_u (eV)	4.79	4.79	4.78	4.78	4.78	4.78
CSFs (references)[b]	1764	1195	916 (12)	758 (9)	651 (6)	651 (6)
CPU time (s)[c]	0.35	0.25	0.22	0.17	0.14	0.15
$k = 5$						
2^1A_g (eV)	3.68	3.72	3.74	3.75	3.78	3.79
1^1B_u (eV)	4.40	4.41	4.39	4.38	4.37	4.38
CSFs (references)[b]	19 404	6601	4067 (16)	3079 (11)	2207 (7)	2003 (6)
CPU time (s)[c]	10.80	2.22	1.21	1.06	0.48	0.42
$k = 6$						
2^1A_g (eV)	3.32	3.41	3.41	3.43	3.48	3.50
1^1B_u (eV)	4.11	4.13	4.10	4.09	4.07	4.08
CSFs (references)[b]	226 512	28 278	13 254 (20)	8781 (13)	5931 (8)	4868 (6)
CPU time (s)[c]	446.09	17.90	6.98	4.05	2.15	1.52
$k = 7$						
2^1A_g (eV)	3.06	3.20	3.19	3.22	3.27	3.30
1^1B_u (eV)	3.75	3.87	3.88	3.87	3.85	3.85
CSFs (references)[b]	2 760 615	98 785	35 336 (26)	24 531 (16)	15 982 (9)	11 380 (7)
CPU time (s)[c]	144 070.91[d]	120.37	32.42	17.01	8.30	5.10

[a] See text. Active space composed of all π and π^* orbitals. Geometries optimized at the B3LYP/TZVP level. [b] Total number of configuration state functions in A_g and B_u symmetry (total number of reference configurations in MR-CISD given in parentheses). [c] Computation times refer to one AMD Opteron(tm) 845 2.8 GHz processor. [d] A semidirect algorithm is used, with all coupling coefficients being recomputed as needed.

INDO/S and INDO/S2 calculations were done with the ZINDO-MN program, version 1.2.[53]

Consistent with the underlying parametrization procedure, the INDO/S and INDO/S2 results were obtained at the CIS level. Following standard INDO/S conventions, the active space generally included the 10 highest occupied MOs and the 10 lowest unoccupied MOs, yielding a total of 101 configuration state functions (CSF).[31] In small molecules with less MOs, all occupied and unoccupied MOs were normally included, but high-lying states were treated with caution: when large spurious $\sigma \rightarrow \sigma^*$ contributions were encountered, the corresponding σ^* MOs were deleted from the active space. The excited-state dipole moments were computed using the recommended class IV charge model 2 (CM2).[35]

The NDDO-based semiempirical methods considered presently (MNDO, AM1, PM3, OM1, OM2, OM3) have been parametrized against ground-state reference data at the SCF (self-consistent-field) level, so that the effects of dynamic ground-state correlation should conceptually be taken into account in an average manner through the parametrization (and through the use of damped two-electron integrals). In electronically excited states, however, there are often static (near-degeneracy) correlation effects, which call for an explicit treatment also in a semiempirical framework, using a suitably chosen (small) active space. To be as unbiased as possible, we adopted a canonical active space with *m* electrons in *n* orbitals (*mn*) for each molecule, in analogy to our previous MS-CASPT2 benchmark study.[14] This active space includes all occupied and unoccupied π-MOs in the case of $\pi \rightarrow \pi^*$ excitations, and in addition the occupied lone-pair MOs in the case of $n \rightarrow \pi^*$ excitations. For a given active space, the least biased correlation treatment is full CI, which, however, quickly becomes too expensive even at the semiempirical level. Therefore, our standard approach in the present benchmark was chosen to be the single-reference CISDTQ treatment, which includes all single, double, triple, and quadruple excitations relative to the closed-shell SCF determinant and which is expected to provide a balanced description of all relevant states (close to the full CI limit).

To check the performance and efficiency of different CI approaches, we performed test calculations on linear polyenes with *k* double bonds using the CI implementation in the MNDO99 code that is based on the graphical unitary group approach (GUGA).[54] Table 1 lists the OM2 results obtained from full CI (FCI), CISDTQ, and various MR-CISD treatments (multireference CI with single and double excitations), which are approximations to CISDTQ. It is obvious that the excitation energies from FCI calculations are reproduced very well by CISDTQ; the deviations (which increase with molecular size) are mostly much smaller than 0.1 eV. At the same time, the computational costs are much reduced for CISDTQ, for example, by a factor of 25 in the case of $k = 6$ where an in-core FCI treatment is still feasible; the much larger factor for $k = 7$ is caused by the switch to a less efficient semidirect FCI algorithm in the MNDO99 code (due to the large size of the CI matrix). In the MR-CISD

Table 2. Statistical Results and Computation Times (s) for the Set of 222 Vertical Excitation Energies (in eV) Using Different Levels of Excitation in the CI Treatment for the OM1, OM2, and OM3 Methods[a]

			MR-CISD[b]			
threshold (%)	FCI	CISDTQ	90	85	80	75
OM1						
count[c]		222	222	222	222	222
mean		0.01	0.03	0.03	0.03	0.03
abs. mean		0.01	0.03	0.03	0.03	0.03
std. dev.		0.01	0.05	0.05	0.06	0.06
max. (+) dev.		0.04	0.27	0.28	0.30	0.28
max. (−) dev.			0.14	0.13	0.13	0.13
CPU time (s)[d]	513.84	54.13	20.04	16.61	15.53	13.96
OM2						
count[c]		222	222	222	222	222
mean		0.00	0.02	0.02	0.03	0.03
abs. mean		0.00	0.02	0.02	0.03	0.03
std. dev.		0.01	0.04	0.04	0.05	0.05
max. (+) dev.		0.03	0.16	0.18	0.19	0.19
max. (−) dev.			0.01	0.01	0.02	0.02
CPU time (s)[d]	510.65	52.53	18.09	15.74	14.88	13.62
OM3						
count[c]		222	222	222	222	222
mean		0.00	0.02	0.02	0.02	0.03
abs. mean		0.00	0.02	0.02	0.03	0.03
std. dev.		0.01	0.04	0.04	0.05	0.05
max. (+) dev.		0.04	0.16	0.17	0.16	0.21
max. (−) dev.			0.03	0.03	0.02	0.02
CPU time (s)[d]	485.64	52.42	17.62	15.30	14.02	13.42

[a] Results from a full CI within the selected active space are taken as reference. [b] See text. [c] Total number of states considered. [d] Computation times refer to one Intel Pentium 4-EMT64T 3.40 GHz processor.

calculations, the reference configurations are selected by an iterative procedure to fulfill the requirement that their combined weight in the CI wave function must exceed a given threshold value (for example, 90% or 85%); starting from a single-reference calculation, this is accomplished by adding the next most important reference configurations until this condition is satisfied (normally within one or two iterations). It is again evident that the MR-CISD results are essentially identical to the CISDTQ results and that they can be obtained at significantly less cost. This conclusion is corroborated by the corresponding data for all other benchmark molecules that are documented in the Supporting Information (see Tables S1−S6) and by the statistical data collected in Table 2. For the full set of vertical excitation energies in our benchmark, the mean absolute deviations from the OM*x*/FCI reference values amount to 0.00−0.01 eV for OM*x*/CISDTQ and to 0.02−0.03 eV for OM*x*/MR-CISD (thresholds 75−90%). In actual applications, it is thus perfectly legitimate to perform such semiempirical excited-state studies at the MR-CISD level; a threshold of 85% seems more than sufficient to ensure close reproduction of the CISDTQ and the FCI results. For the purposes of the present benchmark, we shall, however, rely on CISDTQ.

A final remark in this section concerns the efficiency of semiempirical CI methods relative to TD-DFT. For a direct comparison, we have computed the three low-lying $^1B_2$ excited states of pyridine using the MNDO99 code for OM2/CISDTQ and the TURBOMOLE package (version 5.9.1)[55] for TDDFT-B3LYP/TZVP (starting from a previously converged SCF solution). The ratio of computation times is 1:1578 on a single-processor Intel Pentium 4-EMT64T (3.4 GHz), indicating a difference of about 3 orders of magnitude for a typical example from our benchmark suite. It is clear that the superior speed of the semiempirical CI methods will allow applications that are not feasible with TD-DFT or the even more costly ab initio methods, for example, calculations on larger chromophores or extended excited-state dynamics runs, provided that the accuracy of the semiempirical results is sufficient.

3. General Considerations

Geometries. As already mentioned, the presently used geometries are taken from our previous benchmark.[14] They represent equilibrium ground-state geometries optimized at the MP2/6-31G* level in a suitable point group (assuming the highest symmetry possible). For the sake of consistency, it is obviously advantageous to adopt a common set of geometries during benchmarking, but one may still wonder by how much the results would change upon reoptimizing the geometries at the semiempirical level. One would actually expect rather small changes because semiempirical methods generally yield realistic ground-state geometries for organic molecules (see ref 41 for corresponding statistical data). Test calculations on several of the benchmark molecules confirm this view. Table S7 (Supporting Information) lists the results for a typical example, two low-lying states of pyridine obtained at geometries optimized with MP2/6-31G*, AM1, and OM2. The computed AM1 and OM2 excitation energies vary by up to 0.1 eV for the different geometries, the (small) oscillator strengths appear to be rather insensitive, and the excited-state dipole moments show variations of around 0.1 D. Overall, these changes are small enough to justify the assumption that the qualitative conclusions of the present benchmark study will remain valid also when using geometries optimized at the semiempirical level.

States. Semiempirical methods employ a minimal basis of valence orbitals and can therefore not describe Rydberg states or states with substantial valence/Rydberg mixing properly. The currently used benchmark set was designed to include only valence excited states[14] and should thus be suitable for an assessment of semiempirical methods. One should keep in mind, however, that there is no clear-cut distinction between valence and other excited states in ab initio calculations, and there will be borderline cases especially for higher-lying states whose character may even change at different ab initio levels.[56] Despite these caveats, we present semiempirical results for all valence states considered previously,[14] while acknowledging that it may be easier for semiempirical methods to properly describe the low-lying valence states (say, below 6 eV).

Assignments. For each benchmark molecule, the electronically excited states were first classified according to point-group symmetry. Thereafter, within a given irreducible representation, one has to establish the proper correspondence between the states obtained in the ab initio reference calculations[14] and the present semiempirical calculations. This was accomplished by comparing the computed excited-state wave functions, along with the excitation energies, oscillator strengths, and excited-state dipole moments. Proceeding in this manner, a satisfactory mapping has been achieved in all cases. However, one specific problem should be pointed out. It is well-known[38,39] that the standard NDDO-based methods (MNDO, AM1, PM3) underestimate the gap between bonding and antibonding MOs, with the latter ones being too low in energy because of the symmetric splitting of bonding and antibonding levels. This creates special problems in alternant hydrocarbons and related molecules, where two singly excited configurations strongly contribute to two excited states, which qualitatively correspond to the plus and minus combination of these configurations (for example, HOMO → LUMO + 1 and HOMO − 1 → LUMO generating the L_b and B_b states in the polyenes). In the case of the standard NDDO-based methods, it is easier to populate a higher unoccupied orbital than to vacate the alternancy-related lower occupied orbital (as compared to ab initio methods). Therefore, the relative energies of the two corresponding singly excited configurations will differ, which will translate into different weights in the resulting CI wave functions of the corresponding pair of states. Such differences in the character of states have indeed been observed between standard NDDO-based and ab initio results, with the assignment based primarily on the character of the states in these cases.

4. Vertical Excitation Energies

From our previous study,[14] we have MS-CASPT2 reference data for 152 singlet states and 71 triplet states, and theoretical best estimates (TBE) for the vertical excitation energies of 104 singlet states and 63 triplet states. The full set of the computed semiempirical vertical excitation energies is given in Tables 3 and 4 along with the corresponding MS-CASPT2 and TBE results. INDO/S and INDO/S2 values are listed for all molecules, although they differ only for oxygen-containing compounds (INDO/S2 reparametrization for oxygen). For the OM*x* methods, the vertical excitation energies obtained from FCI, CISDTQ, and MR-CISD (thresholds 75−90%) calculations are documented in more detail in the Supporting Information (Tables S1−S6). A compilation of experimental data is available from one of our previous papers.[51]

In the following discussion of the results, we will focus on comparisons with the TBE values, but our qualitative conclusions remain valid also when considering CASPT2 or CC3 reference data.

Ethene, Butadiene, Hexatriene, and Octatetraene. The energy of the singlet $\pi\pi^*$ state of ethene (TBE 7.80 eV) is well reproduced by the OM*x* methods (errors of less than 0.1 eV), but strongly underestimated by the standard MNDO-type methods (by 1.2−1.6 eV) and somewhat overestimated by INDO/S (by 0.53 eV). The energy of the triplet $\pi\pi^*$ state of ethene (TBE 4.50 eV) is underestimated by all semiempirical methods, but to a varying extent (MNDO/AM1/PM3 by 1.5−1.9 eV, INDO/S by 1.3 eV, OM*x* by 0.34−0.43 eV).

Butadiene is the first member of the $C_{2n}H_{2n+2}$ polyene series. The excitation energy to the bright 1^1B_u $\pi\pi^*$ state (TBE 6.18 eV) is well predicted by the OM*x* methods (errors of less than 0.1 eV) and also by INDO/S (too low by 0.21 eV), while the MNDO/AM1/PM3 values are again much too low (by 1.4−1.9 eV). More interesting is the dark 2^1A_g $\pi\pi^*$ state, which is mainly composed of two single excitations (HOMO → LUMO + 1 and HOMO − 1 → LUMO) and the double excitation HOMO ⇒ LUMO (contributing about 33% to the CISDTQ wave function of any of the NDDO-based methods). While it has been recognized early on that this state becomes the lowest excited singlet for longer polyenes,[57,58] its position in butadiene has remained controversial for a long time. Relative to the currently adopted TBE of 6.55 eV, the excitation energies from MNDO/AM1/PM3 and from OM*x* are too low by 2.0−2.7 and 0.57−0.67 eV, respectively. The standard INDO/S CIS approach does not capture the double-excitation character of the 2^1A_g state and thus overestimates its energy by 0.32 eV; it has been pointed out[59] that inclusion of the doubly excited HOMO⇒LUMO configuration reduces the error significantly (to 0.06 eV). This is partly fortuitous, however, because other doubly excited configurations contribute another 20% to the OM*x*/CISDTQ wave functions such that this state is actually dominated by double excitations. The energies of the two lowest triplet $\pi\pi^*$ states of butadiene (1^1B_u and 2^1A_g) are again underestimated by all semiempirical methods, least so by OM*x* (errors of 0.30−0.59 eV).

In hexatriene and octatetraene, we focus on the state ordering of the two lowest singlet excited states, 2^1A_g and 1^1B_u. The energy gap between these states is minute in hexatriene (TBE 0.01 eV) and still small in octatetraene (TBE 0.19 eV), with 2^1A_g being lower. MNDO/AM1/PM3 and OM*x* overestimate these gaps by 1.0−1.4 and 0.47−0.70 eV, respectively; in the case of OM*x*, the 1^1B_u state is described quite well (too high by 0.04−0.26 eV), while the 2^1A_g state comes out too stable (by 0.23−0.38 eV). As expected, INDO/S CIS calculations give the wrong state ordering for these two small polyenes, with 2^1A_g more than 1.0 eV above 1^1B_u (including the doubly excited HOMO⇒LUMO configuration in the CI calculation does not recover the correct state ordering). The energies for the lowest 1^3B_u triplet state of hexatriene (TBE 2.40 eV) and octatetraene (TBE 2.20 eV) are well reproduced by OM*x* (errors of 0.11 eV or less), while those for the 1^3A_g triplet state (TBE 4.15 and 3.55 eV) are somewhat underestimated (by 0.31−0.42 eV). It should also be noted that all six NDDO-based methods predict realistic singlet−triplet energy differences (which may be important photochemically): the TBE differences $\Delta\Delta E$ ($1^1B_u-1^3B_u$) are reproduced to within 0.17−0.28 eV for hexatriene and 0.19−0.33 eV for octatetraene, and even better in the case of the A_g states (absolute deviations of at most 0.12 and 0.21 eV, respectively).

Cyclopropene, Cyclopentadiene, Norbornadiene, Benzene, and Naphthalene. The energies of the singlet and triplet excited states of cyclopropene are generally underestimated by all NDDO-based methods. For instance, in the case of the 1^1B_2 $\pi\pi^*$ state (TBE 7.06 eV), the deviations are about 1.5 eV for MNDO/AM1/PM3 and 0.47−0.67 eV for OM*x*. This state is rather diffuse at the CASSCF level[60] and may thus be problematic for methods using minimal basis sets.

Table 3. Vertical Singlet Excitation Energies Δ*E* (eV) of All Evaluated Molecules As Compared to MS-CASPT2/TZVP Results and Theoretical Best Estimates (TBE)

molecule	state	CASPT2[a]	TBE[b]	MNDO	AM1	PM3	OM1	OM2	OM3	INDO/S	INDO/S2
ethene	1^1B_{1u} ($\pi \rightarrow \pi^*$)	8.54	7.80	6.18	6.63	6.63	7.82	7.78	7.85	8.33	8.33
E-butadiene	1^1B_u ($\pi \rightarrow \pi^*$)	6.47	6.18	5.18	5.46	5.45	6.24	6.22	6.26	5.97[c]	5.97
	2^1A_g ($\pi \rightarrow \pi^*$)	6.62	6.55	3.90	4.41	4.53	5.88	5.96	5.98	6.87[c]	6.87
hexatriene	1^1B_u ($\pi \rightarrow \pi^*$)	5.31	5.10	4.59	4.78	4.77	5.35	5.33	5.36	4.86[c]	4.86
	2^1A_g ($\pi \rightarrow \pi^*$)	5.42	5.09	3.23	3.63	3.74	4.79	4.86	4.86	5.90[c]	5.90
all-*E*-octatetraene	2^1A_g ($\pi \rightarrow \pi^*$)	4.64	4.47	2.81	3.13	3.24	4.09	4.14	4.13	5.23[c]	5.23
	1^1B_u ($\pi \rightarrow \pi^*$)	4.70	4.66	4.23	4.36	4.35	4.79	4.77	4.79	4.20[c]	4.20
cyclopropene	1^1B_1 ($\sigma \rightarrow \pi^*$)	6.76	6.76	5.22	5.35	5.67	6.33	5.75	5.93	6.92[d]	6.92
	1^1B_2 ($\pi \rightarrow \pi^*$)	7.06	7.06	5.58	5.54	5.58	6.59	6.42	6.39	6.90[d]	6.90
cyclopentadiene	1^1B_2 ($\pi \rightarrow \pi^*$)	5.51	5.55	4.37	4.61	4.68	5.14	5.07	5.09	5.03	5.03
	2^1A_1 ($\pi \rightarrow \pi^*$)	6.31	6.31	3.66	4.11	4.24	5.52	5.60	5.59	6.06	6.06
	3^1A_1 ($\pi \rightarrow \pi^*$)	8.52		6.25	6.51	6.52	7.63	7.47	7.53	7.72	7.72
norbornadiene	1^1A_2 ($\pi \rightarrow \pi^*$)	5.34	5.34	5.04	5.18	5.38	6.10	6.00	6.06	4.50	4.50
	1^1B_2 ($\pi \rightarrow \pi^*$)	6.11	6.11	5.83	5.99	6.11	6.65	6.34	6.46	5.54	5.54
	2^1B_2 ($\pi \rightarrow \pi^*$)	7.32		6.26	6.38	6.49	7.42	7.37	7.42	6.77	6.77
	2^1A_2 ($\pi \rightarrow \pi^*$)	7.45		6.27	6.33	6.45	6.92	6.66	6.74	6.90	6.90
benzene	1^1B_{2u} ($\pi \rightarrow \pi^*$)	5.04	5.08	2.71	3.13	3.19	4.41	4.48	4.51	4.71	4.71
	1^1B_{1u} ($\pi \rightarrow \pi^*$)	6.42	6.54	4.49	4.84	4.90	5.98	5.94	6.03	5.96	5.96
	1^1E_{1u} ($\pi \rightarrow \pi^*$)	7.13	7.13	5.55	5.92	5.88	7.13	7.16	7.20	6.51	6.51
	1^1E_{2g} ($\pi \rightarrow \pi^*$)	8.18	8.41	4.46	5.11	5.20	7.07	7.19	7.22	7.79	7.79
naphthalene	1^1B_{3u} ($\pi \rightarrow \pi^*$)	4.24	4.24	2.35	2.68	2.75	3.76	3.81	3.84	3.92	3.92
	1^1B_{2u} ($\pi \rightarrow \pi^*$)	4.77	4.77	3.84	4.06	4.09	4.85	4.83	4.87	4.50	4.50
	2^1A_g ($\pi \rightarrow \pi^*$)	5.87	5.87	3.20	3.68	3.76	5.16	5.23	5.27	5.52	5.52
	1^1B_{1g} ($\pi \rightarrow \pi^*$)	5.99	5.99	3.80	4.24	4.33	5.67	5.74	5.76	5.65	5.65
	2^1B_{3u} ($\pi \rightarrow \pi^*$)	6.06	6.06	4.87	5.16	5.13	6.12	6.16	6.18	5.53	5.53
	2^1B_{2u} ($\pi \rightarrow \pi^*$)	6.33	6.33	4.70	5.05	5.06	6.22	6.23	6.28	5.96	5.96
	2^1B_{1g} ($\pi \rightarrow \pi^*$)	6.47	6.47	4.80	5.17	5.17	6.29	6.24	6.31	6.39	6.39
	3^1A_g ($\pi \rightarrow \pi^*$)	6.67	6.67	3.85	4.36	4.44	5.94	6.03	6.05	6.79	6.79
	3^1B_{3u} ($\pi \rightarrow \pi^*$)	7.74		4.27	4.84	4.95	6.68	6.80	6.83	7.34	7.34
	3^1B_{2u} ($\pi \rightarrow \pi^*$)	8.17		6.09	6.57	6.57	8.09	8.12	8.18	7.85	7.82
furan	1^1B_2 ($\pi \rightarrow \pi^*$)	6.39	6.32	4.59	4.87	4.87	5.78	5.82	5.88	5.90	5.68
	2^1A_1 ($\pi \rightarrow \pi^*$)	6.50	6.57	3.53	3.96	4.04	5.39	5.43	5.51	5.81	5.74
	3^1A_1 ($\pi \rightarrow \pi^*$)	8.17	8.13	5.72	6.15	6.12	7.44	7.47	7.62	7.88	7.23
pyrrole	2^1A_1 ($\pi \rightarrow \pi^*$)	6.31	6.37	3.37	3.77	3.79	5.21	5.28	5.29	5.38	5.38
	1^1B_2 ($\pi \rightarrow \pi^*$)	6.57	6.57	4.42	4.65	4.55	5.77	5.86	5.94	5.16	5.16
	3^1A_1 ($\pi \rightarrow \pi^*$)	8.17	7.91	5.31	5.65	5.53	7.10	7.18	7.16	6.57	6.57
imidazole	$2^1A'$ ($\pi \rightarrow \pi^*$)	6.19	6.19	4.05	4.32	4.11	5.50	5.59	5.85	5.00	5.00
	$1^1A''$ (n → π^*)	6.81	6.81	5.25	5.24	4.48	5.87	6.00	6.08	5.36	5.36
	$3^1A'$ ($\pi \rightarrow \pi^*$)	6.93	6.93	4.62	4.84	4.70	5.95	6.04	6.16	5.65	5.65
	$2^1A''$ (n → π^*)	7.90		5.82	5.83	5.05	6.81	6.79	6.70	6.19	6.19
	$4^1A'$ ($\pi \rightarrow \pi^*$)	8.16		5.91	6.12	5.71	7.40	7.45	7.69	6.87	6.87
pyridine	1^1B_2 ($\pi \rightarrow \pi^*$)	5.02	4.85	3.01	3.38	3.35	4.56	4.65	4.83	4.76	4.76
	1^1B_1 (n → π^*)	5.17	4.59	4.36	4.29	3.75	4.85	4.85	4.86	4.40	4.40
	1^1A_2 (n → π^*)	5.51	5.11	4.43	4.34	3.96	5.17	5.06	4.84	5.42	5.42
	2^1A_1 ($\pi \rightarrow \pi^*$)	6.39	6.26	4.65	5.01	5.00	6.14	6.11	6.25	6.00	6.00
	2^1B_2 ($\pi \rightarrow \pi^*$)	7.27	7.27	5.85	6.21	6.03	7.39	7.48	7.62	6.75	6.75
	3^1A_1 ($\pi \rightarrow \pi^*$)	7.46	7.18	5.88	6.29	6.11	7.44	7.43	7.63	6.65	6.65
	3^1B_2 ($\pi \rightarrow \pi^*$)	8.60		4.76	5.33	5.32	7.09	7.20	7.44		
	4^1A_1 ($\pi \rightarrow \pi^*$)	8.69		5.31	5.82	5.63	7.53	7.69	8.09		
pyrazine	1^1B_{3u} (n → π^*)	4.21	3.95	3.77	3.56	3.29	3.90	3.81	4.04	3.75	3.75
	1^1A_u (n → π^*)	4.70	4.81	3.75	3.55	3.50	4.38	4.12	3.89	5.08	5.08
	1^1B_{2u} ($\pi \rightarrow \pi^*$)	4.85	4.64	3.37	3.65	3.48	4.61	4.76	5.20	4.61	4.61
	1^1B_{2g} (n → π^*)	5.68	5.56	4.86	4.90	4.13	5.49	5.78	5.86	4.85	4.85
	1^1B_{1g} (n → π^*)	6.41	6.60	5.23	5.36	4.51	6.34	6.54	6.32	6.89	6.89
	1^1B_{1u} ($\pi \rightarrow \pi^*$)	6.89	6.58	4.75	5.10	5.06	6.24	6.22	6.35	6.20	6.20
	2^1B_{2u} ($\pi \rightarrow \pi^*$)	7.66	7.60	6.45	6.66	6.23	7.64	7.67	8.12	7.64	7.64
	2^1B_{1u} ($\pi \rightarrow \pi^*$)	7.79	7.72	6.92	7.15	6.48	7.98	8.06	8.68	7.69	7.69
	1^1B_{3g} ($\pi \rightarrow \pi^*$)	8.47		4.94	5.46	5.39	7.16	7.35	7.68	7.68	7.68
	2^1A_g ($\pi \rightarrow \pi^*$)	8.61		6.04	6.46	6.00	7.89	8.07	8.95	9.31	9.31
pyrimidine	1^1B_1 (n → π^*)	4.44	4.55	3.89	3.78	3.46	4.43	4.34	4.38	4.16	4.16
	1^1A_2 (n → π^*)	4.80	4.91	4.03	3.90	3.66	4.73	4.54	4.40	4.50	4.50
	1^1B_2 ($\pi \rightarrow \pi^*$)	5.24	5.44	3.42	3.71	3.55	4.77	4.86	5.22	5.00	5.00
	2^1A_1 ($\pi \rightarrow \pi^*$)	6.63	6.95	4.92	5.29	5.14	6.37	6.36	6.62	6.39	6.39
	3^1A_1 ($\pi \rightarrow \pi^*$)	7.21		6.43	6.71	6.27	7.76	7.81	8.25	7.34	7.34
	2^1B_2 ($\pi \rightarrow \pi^*$)	7.64		6.37	6.72	6.33	7.63	7.72	8.00	6.92	6.92
	3^1B_2 ($\pi \rightarrow \pi^*$)	8.73		5.93	6.29	5.91	7.92	8.01	8.77	8.27	8.27
	4^1A_1 ($\pi \rightarrow \pi^*$)	9.19		5.33	5.79	5.61	7.37	7.51	8.00	8.13	8.13
pyridazine	1^1B_1 (n → π^*)	3.78	3.78	4.10	4.13	3.15	4.01	4.37	4.14	3.79	3.79
	1^1A_2 (n → π^*)	4.31	4.31	4.29	4.35	3.41	4.39	4.70	4.21	4.66	4.66
	2^1A_1 ($\pi \rightarrow \pi^*$)	5.18	5.18	3.25	3.56	3.41	4.63	4.74	5.08	4.95	4.95
	2^1A_2 (n → π^*)	5.77	5.77	4.89	4.80	4.16	5.40	5.38	5.57	5.68	5.68
	1^1B_2 ($\pi \rightarrow \pi^*$)	6.13		4.97	5.32	5.09	6.34	6.29	6.60	6.30	6.30
	2^1B_1 (n → π^*)	6.52		5.08	5.00	4.48	5.78	5.62	5.51	6.34	6.34
	2^1B_2 ($\pi \rightarrow \pi^*$)	7.29		6.01	6.32	6.03	7.29	7.32	7.76	7.04	7.04
	3^1A_1 ($\pi \rightarrow \pi^*$)	7.62		6.09	6.47	6.21	7.59	7.51	7.81	7.32	7.32
s-triazine	$1^1A_1''$ (n → π^*)	4.60	4.60	4.22	4.00	3.74	4.80	4.51	4.57	4.74	4.74
	$1^1A_2''$ (n → π^*)	4.66	4.66	3.94	3.76	3.53	4.48	4.24	4.28	4.61	4.61
	$1^1E''$ (n → π^*)	4.70	4.70	4.08	3.89	3.61	4.66	4.40	4.45	4.44	4.44
	$1^1A_2'$ ($\pi \rightarrow \pi^*$)	5.79	5.79	3.94	4.13	3.79	5.06	5.12	5.69	5.45	5.45
	$2^1A_1'$ ($\pi \rightarrow \pi^*$)	7.25		5.51	5.80	5.34	6.74	6.78	7.33	6.90	6.90
	$1^1E'$ ($\pi \rightarrow \pi^*$)	7.50		6.93	7.16	6.56	7.72	7.88	8.47	7.57	7.57
	$2^1E''$ (n → π^*)	7.71		6.13	6.00	5.73	7.48	7.01	7.17	7.00	7.00
	$2^1E'$ ($\pi \rightarrow \pi^*$)	8.99		6.26	6.54	6.04	8.26	8.28	9.03	8.95	8.95

Table 3. Continued

molecule	state	CASPT2[a]	TBE[b]	MNDO	AM1	PM3	OM1	OM2	OM3	INDO/S	INDO/S2
s-tetrazine	1^1B_{3u} (n → π^*)	2.29	2.29	2.98	2.91	2.27	2.66	2.83	2.88	2.86	2.86
	1^1A_u (n → π^*)	3.51	3.51	3.80	3.65	2.99	3.53	3.55	3.08	4.43	4.43
	1^1B_{1g} (n → π^*)	4.73	4.73	5.09	5.58	3.69	5.08	6.15	6.49	4.38	4.38
	1^1B_{2u} ($\pi \rightarrow \pi^*$)	4.93	4.93	3.62	3.84	3.50	4.71	4.88	5.74	4.84	4.84
	1^1B_{2g} (n → π^*)	5.20	5.20	4.42	4.41	3.87	5.02	5.33	6.11	4.94	4.94
	2^1A_u (n → π^*)	5.50	5.50	4.99	4.78	3.80	4.89	4.65	5.06	5.59	5.59
	1^1B_{3g} (n → π^*)[e]	5.86		6.54	6.41	5.07	5.93	6.28	5.88		
	2^1B_{2g} (n → π^*)	6.06		5.82	6.02	4.42	5.97	6.73	6.64	6.64	6.64
	2^1B_{1g} (n → π^*)	6.45		5.29	5.12	4.47	5.63	5.65	5.60	6.59	6.59
	3^1B_{1g} (n → π^*)	6.73		5.56	5.39	4.92	6.29	6.39	6.72	7.64	7.64
	2^1B_{3u} (n → π^*)	6.77		5.68	5.38	4.53	5.66	5.22	5.08	7.29	7.29
	1^1B_{1u} ($\pi \rightarrow \pi^*$)	6.94		5.70	5.91	5.29	6.75	6.78	7.23	6.64	6.64
	2^1B_{1u} ($\pi \rightarrow \pi^*$)	7.42		6.68	6.84	6.16	7.67	7.86	9.03	7.27	7.27
	2^1B_{2u} ($\pi \rightarrow \pi^*$)	8.14		7.16	7.32	6.63	8.24	8.16	9.04	8.18	8.18
	2^1B_{3g} ($\pi \rightarrow \pi^*$)	8.34		5.91	6.22	5.55	7.54	7.86	9.62	9.30	9.30
formaldehyde	1^1A_2 (n → π^*)	3.99	3.88	3.21	3.07	2.87	3.71	3.55	3.59	3.62	4.09
	1^1B_1 ($\sigma \rightarrow \pi^*$)	9.14	9.10	8.46	8.68	8.63	9.30	7.93	9.01	10.95	11.36
	2^1A_1 ($\pi \rightarrow \pi^*$)	9.32	9.30	8.56	9.09	8.40	9.61	9.23	9.89	12.09	13.54
acetone	1^1A_2 (n → π^*)	4.44	4.40	3.18	3.46	3.29	3.80	3.98	4.05	3.67	4.13
	1^1B_1 ($\sigma \rightarrow \pi^*$)	9.27	9.10	8.26	8.18	7.46	8.81	7.71	8.34	10.63	11.17
	2^1A_1 ($\pi \rightarrow \pi^*$)	9.31	9.40	7.97	7.86	8.28	8.33	8.08	8.51	10.87	11.79
p-benzoquinone	1^1B_{1g} (n → π^*)	2.76	2.76	2.79	2.88	2.72	2.62	2.64	2.58	2.67	3.00
	1^1A_u (n → π^*)	2.77	2.77	2.95	3.19	2.96	3.17	3.35	3.37	2.64	3.00
	1^1B_{3g} ($\pi \rightarrow \pi^*$)	4.26	4.26	4.27	4.43	4.43	4.82	4.62	4.68	4.75	4.79
	1^1B_{1u} ($\pi \rightarrow \pi^*$)	5.28	5.28	5.26	5.47	5.26	5.64	5.52	5.71	5.60	5.83
	1^1B_{3u} (n → π^*)	5.64	5.64	4.42	4.70	4.68	5.24	5.34	5.31	5.56	5.59
	2^1B_{3g} ($\pi \rightarrow \pi^*$)	6.96	6.96	5.36	5.80	5.65	6.58	6.73	6.93	6.69	6.93
	2^1B_{1u} ($\pi \rightarrow \pi^*$)	7.92		6.28	6.66	6.65	7.81	7.79	7.87	7.41	7.45
formamide	1^1 A″ (n → π^*)	5.63	5.63	3.89	4.11	3.68	4.61	4.56	4.82	4.44	5.09
	2^1A′ ($\pi \rightarrow \pi^*$)	7.39	7.39	5.90	5.93	5.08	6.92	6.71	7.07	7.38	7.64
	3^1A′ ($\pi \rightarrow \pi^*$)	10.54		8.26	8.76	7.95	9.67	9.41	10.11	12.23	13.54
acetamide	1^1 A″ (n → π^*)	5.69	5.69	3.83	4.23	3.79	4.59	4.75	4.98	4.36	5.00
	2^1A′ ($\pi \rightarrow \pi^*$)	7.27	7.27	5.70	5.77	4.92	6.76	6.63	6.95	7.39	7.64
	3^1A′ ($\pi \rightarrow \pi^*$)	10.09		7.82	8.09	7.49	8.93	8.64	9.02	11.31	12.19
propanamide	1^1A″ (n → π^*)	5.72	5.72	3.91	4.35	3.87	4.67	4.85	5.06	4.35	4.99
	2^1A′ ($\pi \rightarrow \pi^*$)	7.20	7.20	5.69	5.77	4.92	6.78	6.64	6.94	7.46	7.71
	3^1A′ ($\pi \rightarrow \pi^*$)	9.94		7.77	7.90	7.41	8.72	8.34	8.50	10.82	11.18
cytosine	2^1A′ ($\pi \rightarrow \pi^*$)	4.67	4.66	3.34	3.47	3.12	4.19	4.21	4.39	4.41	4.50
	1^1A″ (n → π^*)	5.12	4.87	3.63	3.76	3.36	4.19	4.23	4.40	4.10	4.15
	2^1A″ (n → π^*)	5.53	5.26	4.05	4.32	3.83	4.76	4.83	4.95	4.73	5.33
	3^1A′ ($\pi \rightarrow \pi^*$)	5.53	5.62	3.87	4.00	3.43	5.01	5.00	5.05	5.54	5.58
	4^1A′ ($\pi \rightarrow \pi^*$)	6.40		4.66	4.72	4.18	5.88	5.78	5.84	6.14	6.26
	5^1A′ ($\pi \rightarrow \pi^*$)	6.97		4.81	4.94	4.44	6.14	5.91	5.96	6.57	6.61
thymine	1^1A″ (n → π^*)	4.95	4.82	3.77	4.18	3.63	4.34	4.52	4.68	4.05	4.63
	2^1A′ ($\pi \rightarrow \pi^*$)	5.06	5.20	4.02	4.09	3.63	4.97	4.81	4.81	5.07	5.17
	3^1A′ ($\pi \rightarrow \pi^*$)	6.15	6.27	4.75	4.97	4.23	5.78	5.56	5.65	6.04	6.22
	2^1A″ (n → π^*)	6.38	6.16	4.42	4.96	4.21	5.32	5.47	5.69	4.94	5.72
	4^1A′ ($\pi \rightarrow \pi^*$)	6.53	6.53	4.93	4.85	4.37	5.91	5.73	5.90	6.65	6.94
	3^1A″ (n → π^*)	6.85		4.95	5.41	4.91	6.07	6.08	6.08	6.50	6.78
	5^1A′ ($\pi \rightarrow \pi^*$)	7.43		5.67	5.78	5.08	6.71	6.50	6.73	7.35	7.67
	4^1A″ (n → π^*)	7.43		5.62	6.01	5.24	6.48	6.36	6.45	7.05	7.50
uracil	1^1A″ (n → π^*)	4.91	4.80	3.73	4.12	3.61	4.27	4.45	4.64	4.06	4.64
	2^1A′ ($\pi \rightarrow \pi^*$)	5.23	5.35	4.09	4.15	3.66	5.05	4.88	4.90	5.19	5.29
	3^1A′ ($\pi \rightarrow \pi^*$)	6.15	6.26	4.80	4.96	4.24	5.78	5.68	5.86	6.31	6.51
	2^1A″ (n → π^*)	6.28	6.10	4.42	4.95	4.21	5.29	5.43	5.65	4.93	5.70
	4^1A′ ($\pi \rightarrow \pi^*$)	6.74	6.70	4.99	5.02	4.47	6.14	5.85	5.97	6.58	6.87
	3^1A″ (n → π^*)	6.98	6.56	5.00	5.43	4.94	6.09	6.09	6.10	6.47	6.74
	4^1A″ (n → π^*)	7.28		5.65	6.00	5.24	6.47	6.34	6.44	7.01	7.42
	5^1A′ ($\pi \rightarrow \pi^*$)	7.42		5.70	5.78	5.11	6.72	6.50	6.75	7.20	7.42
adenine	1^1A″ (n → π^*)	5.19	5.12	4.08	3.99	3.81	4.81	4.60	4.62	4.55	4.55
	2^1A′ ($\pi \rightarrow \pi^*$)	5.20	5.25	3.08	3.21	2.95	4.19	4.23	4.33	4.33	4.33
	3^1A′ ($\pi \rightarrow \pi^*$)	5.29	5.25	3.73	3.89	3.54	4.83	4.79	4.90	4.61	4.61
	2^1A″ (n → π^*)	5.96	5.75	4.36	4.32	4.09	5.10	5.05	5.16	4.69	4.69
	4^1A′ ($\pi \rightarrow \pi^*$)	6.34		4.21	4.38	4.03	5.61	6.03	5.80	5.87	5.87
	5^1A′ ($\pi \rightarrow \pi^*$)	6.64		4.62	4.75	4.17	6.00	5.91	6.14	5.36	5.36
	6^1A′ ($\pi \rightarrow \pi^*$)	6.87		4.79	4.92	4.45	6.02	6.50	6.27	6.17	6.17
	7^1A′ ($\pi \rightarrow \pi^*$)	7.56		5.14	5.25	4.69	6.49	6.64	6.74		

[a] SA-CASSCF/MS-CASPT2 results using the TZVP basis and MP2/6-31G* ground-state equilibrium geometries.[14] [b] Theoretical best estimates for vertical excitation energies. See ref 14 for details. [c] Including the doubly excited CSF (HOMO⇒LUMO) changes the energy of the $1^1B_u/2^1A_g$ states to 6.37/6.61 eV for butadiene, 5.14/5.29 eV for hexatriene, and 4.41/4.85 eV for octatetraene. [d] Using all orbitals in the active space yields 5.80/6.87 eV for $1^1B_1/1^1B_2$. [e] Double excitation.

Cyclopentadiene has three valence excited singlet states, the lowest one being of B_2 symmetry (HOMO → LUMO transition) followed by two A_1 states (composed of combinations of HOMO − 1 → LUMO and HOMO → LUMO + 1 single excitations as well as double excitations). The semiempirical calculations again underestimate the excitation energies, for example, with OM*x* by 0.41−0.46 eV for 1^1B_2 (TBE 5.55 eV) and by 0.71−0.79 eV for 2^1A_1 (TBE 6.31 eV); similar deviations are found for the corresponding triplet states. The INDO/S results are reasonable for the singlet states, but much too low for the triplet states (by 1.2−1.6 eV).

Norbornadiene with its two nonconjugated double bonds seems to be described reasonably well by all six NDDO-based methods, with the energies of the two lowest singlet excited states (1^1A_2, TBE 5.34 eV; 1^1B_2, TBE 6.11 eV) being

Table 4. Vertical Triplet Excitation Energies ΔE (eV) of All Evaluated Molecules As Compared to MS-CASPT2/TZVP Results and Theoretical Best Estimates (TBE)

molecule	atate	CASPT2[a]	TBE[b]	MNDO	AM1	PM3	OM1	OM2	OM3	INDO/S	INDO/S2
ethene	$1^3B_{1u}\ (\pi \rightarrow \pi^*)$	4.48	4.50	2.58	2.98	3.05	4.07	4.14	4.16	3.23	3.23
E-butadiene	$1^3B_u\ (\pi \rightarrow \pi^*)$	3.34	3.20	1.94	2.21	2.28	2.99	3.04	3.04	2.24	2.24
	$1^3A_g\ (\pi \rightarrow \pi^*)$	5.16	5.08	2.82	3.25	3.32	4.49	4.56	4.59	3.71	3.71
all-*E*-hexatriene	$1^3B_u\ (\pi \rightarrow \pi^*)$	2.71	2.40	1.61	1.81	1.88	2.41	2.46	2.45	2.65	2.65
	$1^3A_g\ (\pi \rightarrow \pi^*)$	4.31	4.15	2.40	2.75	2.81	3.73	3.80	3.81	4.39	4.39
all-*E*-octatetraene	$1^3B_u\ (\pi \rightarrow \pi^*)$	2.33	2.20	1.44	1.59	1.66	2.09	2.12	2.11	2.20	2.20
	$1^3A_g\ (\pi \rightarrow \pi^*)$	3.70	3.55	2.10	2.37	2.43	3.18	3.23	3.24	3.32	3.32
cyclopropene	$1^3B_2\ (\pi \rightarrow \pi^*)$	4.35	4.34	2.53	2.68	2.76	3.77	3.80	3.72	3.18	3.18
	$1^3B_1\ (\sigma \rightarrow \pi^*)$	6.51	6.62	4.89	5.02	5.23	6.01	5.48	5.65	7.25	7.25
cyclopentadiene	$1^3B_2\ (\pi \rightarrow \pi^*)$	3.28	3.25	1.80	2.07	2.18	2.81	2.87	2.86	2.02	2.02
	$1^3A_1\ (\pi \rightarrow \pi^*)$	5.11	5.09	2.67	3.05	3.14	4.23	4.30	4.31	3.46	3.46
norbornadiene	$1^3A_2\ (\pi \rightarrow \pi^*)$	3.75	3.72	2.52	2.88	2.95	4.08	4.27	4.26	2.75	2.75
	$1^3B_2\ (\pi \rightarrow \pi^*)$	4.22	4.16	2.45	2.81	2.84	3.89	4.10	4.07	3.15	3.15
benzene	$1^3B_{1u}\ (\pi \rightarrow \pi^*)$	4.17	4.15	2.13	2.50	2.56	3.66	3.74	3.76	3.72	3.72
	$1^3E_{1u}\ (\pi \rightarrow \pi^*)$	4.90	4.86	2.85	3.25	3.31	4.48	4.54	4.57	4.83	4.83
	$1^3B_{2u}\ (\pi \rightarrow \pi^*)$	5.76	5.88	4.42	4.73	4.72	5.79	5.80	5.85	5.46	5.46
	$1^3E_{2g}\ (\pi \rightarrow \pi^*)$	7.38	7.51	3.85	4.45	4.53	6.20	6.30	6.34	6.97	6.97
naphthalene	$1^3B_{2u}\ (\pi \rightarrow \pi^*)$	3.16	3.11	1.76	2.01	2.07	2.84	2.89	2.90	2.95	2.95
	$1^3B_{3u}\ (\pi \rightarrow \pi^*)$	4.25	4.18	2.50	2.83	2.88	3.87	3.91	3.94	4.15	4.15
	$1^3B_{1g}\ (\pi \rightarrow \pi^*)$	4.51	4.47	2.51	2.86	2.93	4.00	4.07	4.09	4.20	4.20
	$2^3B_{2u}\ (\pi \rightarrow \pi^*)$	4.68	4.64	2.65	3.02	3.07	4.22	4.28	4.31	4.61	4.61
	$2^3B_{3u}\ (\pi \rightarrow \pi^*)$	4.97	5.11	3.91	4.09	4.09	4.95	4.95	4.99	4.61	4.61
	$1^3A_g\ (\pi \rightarrow \pi^*)$	5.53	5.52	3.09	3.51	3.57	4.82	4.90	4.93	5.12	5.12
	$2^3B_{1g}\ (\pi \rightarrow \pi^*)$	6.21	6.48	4.98	5.33	5.32	6.46	6.49	6.52	7.33	7.33
	$2^3A_g\ (\pi \rightarrow \pi^*)$	6.38	6.47	4.84	5.23	5.23	6.50	6.51	6.58	6.39	6.39
	$3^3A_g\ (\pi \rightarrow \pi^*)$	6.59	6.79	3.61	4.11	4.19	5.68	5.76	5.79	7.32	7.32
	$3^3B_{1g}\ (\pi \rightarrow \pi^*)$	6.64	6.76	3.66	4.17	4.23	5.71	5.79	5.83	6.63	6.63
furan	$1^3B_2\ (\pi \rightarrow \pi^*)$	4.18	4.17	2.18	2.47	2.50	3.40	3.50	3.53	2.79	2.95
	$1^3A_1\ (\pi \rightarrow \pi^*)$	5.49	5.48	2.85	3.22	3.29	4.44	4.54	4.59	3.97	4.00
pyrrole	$1^3B_2\ (\pi \rightarrow \pi^*)$	4.51	4.48	2.26	2.58	2.76	3.55	3.76	3.89	2.47	2.47
	$1^3A_1\ (\pi \rightarrow \pi^*)$	5.52	5.51	2.88	3.23	3.23	4.46	4.59	4.64	3.81	3.81
imidazole	$1^3A'\ (\pi \rightarrow \pi^*)$	4.65	4.69	2.46	2.77	2.89	3.70	3.95	4.06	3.04	3.04
	$2^3A'\ (\pi \rightarrow \pi^*)$	5.74	5.79	3.59	3.80	3.60	4.77	4.95	5.27	4.32	4.32
	$1^3A''\ (n \rightarrow \pi^*)$	6.36	6.37	4.89	4.84	4.14	5.43	5.60	5.70	5.42	5.42
	$3^3A'\ (\pi \rightarrow \pi^*)$	6.44	6.55	4.49	4.60	4.12	5.79	5.69	5.99	5.59	5.59
	$4^3A'\ (\pi \rightarrow \pi^*)$	7.44		4.87	4.95	4.48	6.28	6.20	6.37	7.12	7.12
	$2^3A''\ (n \rightarrow \pi^*)$	7.51		5.50	5.48	4.83	6.50	6.47	6.33	7.48	7.48
pyridine	$1^3A_1\ (\pi \rightarrow \pi^*)$	4.27	4.06	2.31	2.65	2.68	3.78	3.86	3.94	3.80	3.80
	$1^3B_1\ (n \rightarrow \pi^*)$	4.57	4.25	3.96	3.86	3.37	4.37	4.48	4.47	4.02	4.02
	$1^3B_2\ (\pi \rightarrow \pi^*)$	4.71	4.64	3.08	3.46	3.44	4.57	4.66	4.83	4.68	4.68
	$2^3A_1\ (\pi \rightarrow \pi^*)$	5.03	4.91	3.22	3.58	3.50	4.67	4.74	4.97	4.93	4.93
	$1^3A_2\ (n \rightarrow \pi^*)$	5.52	5.28	4.36	4.27	3.89	5.11	4.96	4.75	5.84	5.84
	$2^3B_2\ (\pi \rightarrow \pi^*)$	6.03	6.08	4.79	5.08	4.92	6.02	5.97	6.17	6.02	6.02
	$3^3A_1\ (\pi \rightarrow \pi^*)$	7.56		4.61	5.05	4.87	6.56	6.67	7.08		
	$3^3B_2\ (\pi \rightarrow \pi^*)$	7.87		4.25	4.73	4.68	6.42	6.59	6.83		
s-tetrazine	$1^3B_{3u}\ (n \rightarrow \pi^*)$	1.61	1.89	2.43	2.36	1.88	2.08	2.35	2.36	2.30	2.30
	$1^3A_u\ (n \rightarrow \pi^*)$	3.28	3.52	3.55	3.38	2.80	3.31	3.31	2.87	4.32	4.32
	$1^3B_{1g}\ (n \rightarrow \pi^*)$	4.14	4.21	4.23	4.30	3.14	4.12	4.85	4.78	3.72	3.72
	$1^3B_{1u}\ (\pi \rightarrow \pi^*)$	4.37	4.33	2.95	3.11	2.78	3.90	4.07	4.74	3.54	3.54
	$1^3B_{2u}\ (\pi \rightarrow \pi^*)$	4.39	4.54	3.53	3.78	3.49	4.57	4.76	5.62	3.98	3.98
	$1^3B_{2g}\ (n \rightarrow \pi^*)$	4.94	4.93	4.15	4.13	3.62	4.73	5.04	5.66	4.91	4.91
	$2^3A_u\ (n \rightarrow \pi^*)$	5.04	5.03	4.69	4.46	3.67	4.67	4.40	4.79	5.89	5.89
	$2^3B_{1u}\ (\pi \rightarrow \pi^*)$	5.40	5.38	3.92	4.11	3.77	4.96	5.07	6.00	5.27	5.27
	$1^3B_{3g}\ (n \rightarrow \pi^*)$	5.57		5.24	5.53	4.86	5.75	6.00	5.53		
	$2^3B_{2g}\ (n \rightarrow \pi^*)$	5.97		5.65	5.83	4.32	5.84	6.50	6.43	7.05	7.05
	$2^3B_{1g}\ (n \rightarrow \pi^*)$	6.37		5.43	5.38	4.85	6.16	6.29	6.65	7.37	7.37
	$2^3B_{3u}\ (n \rightarrow \pi^*)$	6.54		5.55	5.24	4.35	5.49	5.09	4.95	7.87	7.87
	$2^3B_{2u}\ (\pi \rightarrow \pi^*)$	7.08		5.76	5.87	5.22	6.69	6.73	7.43	7.23	7.23
formaldehyde	$1^3A_2\ (\pi \rightarrow \pi^*)$	3.58	3.50	2.92	2.74	2.57	3.40	3.23	3.24	3.14	3.68
	$1^3A_1\ (\pi \rightarrow \pi^*)$	5.84	5.87	4.90	5.42	5.07	5.67	5.63	6.07	6.41	8.04
acetone	$1^3A_2\ (n \rightarrow \pi^*)$	4.10	4.05	2.87	3.18	3.05	3.53	3.74	3.79	4.02	4.54
	$1^3A_1\ (\pi \rightarrow \pi^*)$	6.04	6.03	4.50	4.97	4.65	5.33	5.45	5.79	8.52	10.33
p-benzoquinone	$1^3B_{1g}\ (n \rightarrow \pi^*)$	2.62	2.51	2.61	2.71	2.57	2.46	2.50	2.42	2.74	3.09
	$1^3A_u\ (n \rightarrow \pi^*)$	2.66	2.62	2.76	3.01	2.82	3.03	3.21	3.23	2.81	3.24
	$1^3B_{1u}\ (\pi \rightarrow \pi^*)$	2.99	2.96	2.18	2.45	2.41	2.75	2.79	2.91	3.48	3.80
	$1^3B_{3g}\ (\pi \rightarrow \pi^*)$	3.32	3.41	2.33	2.63	2.68	3.38	3.32	3.34	3.49	3.64
formamide	$1^3A''\ (n \rightarrow \pi^*)$	5.40	5.36	3.68	3.87	3.49	4.39	4.34	4.58	4.00	4.69
	$1^3A'\ (\pi \rightarrow \pi^*)$	5.58	5.74	4.02	4.16	3.54	4.89	4.77	4.98	5.21	5.93
acetamide	$1^3A''\ (n \rightarrow \pi^*)$	5.41	5.42	3.61	4.00	3.62	4.38	4.54	4.76	4.99	5.69
	$1^3A'\ (\pi \rightarrow \pi^*)$	5.63	5.88	3.86	4.11	3.49	4.86	4.86	5.07	6.61	7.31
propanamide	$1^3A''\ (n \rightarrow \pi^*)$	5.45	5.45	3.68	4.11	3.69	4.44	4.61	4.81	5.00	5.70
	$1^3A'\ (\pi \rightarrow \pi^*)$	5.80	5.90	3.87	4.12	3.50	4.87	4.87	5.07	6.61	7.30

[a] SA-CASSCF/MS-CASPT2 results using the TZVP basis and MP2/6-31G* ground-state equilibrium geometries.[14] [b] Theoretical best estimates for vertical excitation energies. See ref 14 for details.

bracketed by the semiempirical results (with individual deviations of a few tenths of an eV). Concerning the two lowest singlet B_2 states, MNDO/AM1/PM3 predicts a significantly higher oscillator strength for the lower one, while the OM*x* and ab initio calculations give the opposite trend, which is indicative of differences in the corresponding CI wave functions. The energy gap between these two singlet B_2 states is rather low in MNDO/AM1/PM3 (0.38−0.43 eV) as compared to OM*x* (0.77−1.03 eV) and the MS-CASPT2 reference value (1.21 eV).[14]

The electronic spectrum of benzene has often been studied theoretically and has served as a prototypical test case for many computational methods. There are four singlet valence $\pi\pi^*$ states in the 5−9 eV range. MNDO/AM1/PM3 underestimate their energies severely, by 1.2−3.9 eV. The OM*x* methods predict the first two singlet states (1^1B_{2u}, TBE 5.08 eV; 1^1B_{1u}, TBE 6.54 eV) too low by 0.51−0.67 eV, get the correct energy within 0.07 eV for the bright 1^1E_{1u} state (TBE 7.13 eV), and strongly underestimate the energy of the E_{2g} state (TBE 8.41 eV), which has significant double-excitation character, by 1.2−1.3 eV. The standard INDO/S CIS results for these $\pi\pi^*$ states suffer from $\sigma\sigma^*$ contamination, which reduces most of the excitation energies by about 0.5 eV; after excluding the corresponding σ^* MOs from the active space, the results are reasonable, with the excited singlet states in the right order and deviations of 0.37−0.62 eV from the TBE values. For the four triplet $\pi\pi^*$ states, the situation is similar as for the singlets, with an analogous performance of the different methods, which does not warrant further discussion.

In the case of naphthalene, our benchmark set includes 10 excited singlet states and 10 triplet states, which are all of $\pi\pi^*$ type. MNDO/AM1/PM3 again underestimate the excitation energies severely in general and give a partially incorrect state ordering. The OM*x* results are much closer to the TBE reference values, with a tendency to be too low by a few tenths of an eV, and normally produce the correct state ordering; the largest deviations from the TBE values occur for A_g states with significant double-excitation character where the OM*x* excitation energies are typically too low by about 0.6−0.7 eV. The INDO/S results appear to be of overall quality similar to those from OM*x*; the INDO/S energies for the A_g states are closer to the TBE values, which should be considered fortuitous because the INDO/S CIS calculations do not include double excitations.

Furan, Pyrrole, Imidazole, Pyridine, Pyrazine, Pyrimidine, Pyridazine, *s*-Triazine, and *s*-Tetrazine. Furan and pyrrole are isoelectronic five-ring heterocycles, which both have three singlet and two triplet $\pi\pi^*$ valence excited states. All semiempirical methods underestimate the corresponding excitation energies. The deviations from the TBE values are largest for MNDO/AM1/PM3 (1.5−3.0 eV) and still substantial for OM*x* (0.4−1.2 eV) and INDO/S (0.4−2.0 eV). For the first two close-lying excited singlets, the six NDDO-based methods give the reverse order for furan (like INDO/S) and the correct order for pyrrole (unlike INDO/S), while the sequence of the triplet states is always predicted correctly. INDO/S2 differs from INDO/S only in the oxygen parametrization and thus yields slightly different results for furan (surprisingly with slightly larger deviations from the TBE data).

The imidazole spectrum contains nπ^* and $\pi\pi^*$ valence transitions in the range betweeen 6.0 and 8.5 eV.[14] All semiempirical methods underestimate the $\pi\pi^*$ excitation energies in a manner similar to that in furan and pyrrole. The energies of the lowest nπ^* state in the singlet manifold ($1^1A''$, TBE 6.81 eV) and in the triplet manifold ($1^3A''$, TBE 6.37 eV) are underestimated to an extent similar to those of the $\pi\pi^*$ states (MNDO/AM1/PM3 by 1.6−2.3 eV, OM*x* by 0.7−0.9 eV, INDO/S by 1.0−1.5 eV), so that OM*x* and INDO/S give the correct order of the singlet and triplet states.

Pyridine is the first of a series of azabenzenes in our benchmark set. The introduction of one nitrogen atom lowers the symmetry and splits degenerate levels, such that the four valence $\pi\pi^*$ states in benzene correlate with six $\pi\pi^*$ states of symmetry A_1 and B_2 in pyridine. TBE values are available for the lowest four of these states both in the singlet case (4.85, 6.26, 7.18, and 7.27 eV) and in the triplet case (4.06, 4.64, 4.91, and 6.08 eV). The OM*x* results scatter around these reference data (typically within 0.3 eV or less), and the INDO/S results are of similar quality. There are two additional nπ^* states in the singlet manifold (1^1B_1, TBE 4.59 eV; 1^1A_2, TBE 5.11 eV) and in the triplet manifold (1^3B_1, TBE 4.25 eV; 1^3A_2, TBE 5.28 eV) whose energies are again well reproduced by OM*x* (typically within 0.3 eV) and also by INDO/S, even though the splitting of these two states is underestimated by OM*x* and overestimated by INDO/S. Overall, however, both the OM*x* methods and the INDO/S perform quite well for pyridine.

Similar remarks apply to the azabenzenes with two ring nitrogen atoms, that is, pyrazine, pyrimidine, and pyridazine, which are represented in our benchmark set only through their singlet excited states. The available TBE values for the eight lowest singlets in pyrazine and for the four lowest singlets in the other two molecules are generally well reproduced by the OM*x* methods, with typical deviations of about 0.3 eV both for $\pi\pi^*$ and nπ^* transitions and with a state ordering that is generally analogous to the one obtained from the ab initio reference calculations. The INDO/S results are generally in the same ballpark as the OM*x* results. Focusing on problem cases, we note that the deviations from the TBE values are larger than usual for the OM*x* energies of the 1^1A_u nπ^* state in pyrazine (TBE 4.81 eV, OM*x* too low by 0.43−0.92 eV) and for the INDO/S energy of the 1^1B_{2g} nπ^* state in pyrazine (TBE 5.56 eV, INDO/S too low by 0.71 eV); as a consequence, the splitting between these two states is overestimated by OM*x* and underestimated by INDO/S (where their order is actually inverted).

In the case of *s*-triazine, we focus on the four lowest singlet transitions for which TBE values have been derived. The three lowest singlet states are of nπ^* type and almost degenerate, lying within 0.1 eV (TBE 4.60−4.70 eV). The semiempirical calculations also yield three close-lying nπ^* states, typically within 0.3 eV, which appear in a similar energy range with OM*x* (4.3−4.8 eV) and INDO/S (4.4−4.7 eV), although their order is generally different from that found at the ab initio level. The energy of the lowest dark $\pi\pi^*$ singlet state ($1^1A_2'$, TBE 5.79 eV) is underestimated by all semiempirical methods (e.g., OM*x* by 0.10−0.73 eV, INDO/S by 0.34 eV). For the higher transitions above 7 eV, there is fair agreement with the available MS-CASPT2/TZVP data, with deviations roughly as expected in this energy range.

s-Tetrazine has a large number of known nπ^* and $\pi\pi^*$ states, with TBE values being available for the lowest six (eight) excited states in the singlet (triplet) manifold. Without going into detail, we note that the OM*x* and INDO/S methods perform about as well as anticipated from the experience from the other azabenzenes. Given the number of close-lying states, it is inevitable, however, that there are sometimes differences in the state orderings and outliers for some of the computed excitation energies, particularly in the case of OM3.

Formaldehyde, Acetone, *p*-Benzoquinone, Formamide, Acetamide, and Propanamide. The lowest excited states of formaldehyde are nπ* transitions, 1^3A_2 (TBE 3.50 eV) and 1^1A_2 (TBE 3.88 eV). Their energies are underestimated slightly by OM*x* (by 0.10−0.33 eV) and INDO/S (by 0.26−0.36 eV), and more so by MNDO/AM1/PM3 (by 0.58−1.01 eV). For the $\pi\pi$* triplet state (1^3A_1, TBE 5.87 eV), the computed energies are too low for MNDO/AM1/PM3 (by 0.45−0.97 eV), while they scatter around the TBE value for OM*x* (within 0.2 eV) and are too high for INDO/S (by 0.54 eV) and especially for INDO/S2 (by 2.2 eV). The two remaining valence excited singlet states in our benchmark set lie above 9 eV, 1^1B_1 ($\sigma \rightarrow \pi$* type, TBE 9.10 eV) and 2^1A_1 ($\pi \rightarrow \pi$* type, TBE 9.30 eV). Their MS-CASPT2/TZVP wave functions indicate significant contributions from higher excitations in the 1^1B_1 state and some notable Rydberg-valence mixing in the 2^1A_1 state, which should only be partially recovered in semiempirical calculations; given this situation, the errors for the six NDDO-based methods of 0.2−1.2 eV are not excessive, while the INDO/S energies are too high by 1.9−2.8 eV.

The electronic spectrum of acetone is qualitatively similar to that of formaldehyde, and the performance of various semiempirical methods is also similar for both molecules, with a slight increase in the deviations from the TBE values for the lowest three states below 6 eV.

In *p*-benzoquinone, the two lowest singlet excited states are of nπ* type. They are almost degenerate (1^1B_{1g}, TBE 2.76 eV; 1^1A_u, TBE 2.77 eV). All semiempirical methods reproduce the energy of the lowest state quite well (within 0.18 eV or less), but except for INDO/S, they compute the second state too high and thus give a sizable gap between these two dark nπ* states (MNDO/AM1/PM3 0.16−0.31 eV, OM*x* 0.55−0.79 eV). The energy of the third nπ* singlet state (1^1B_{3u}, TBE 5.64 eV) is underestimated to a different extent (MNDO/AM1/PM3 by 1.0−1.2 eV, OM*x* by 0.30−0.40 eV, INDO/S by 0.08 eV). The positions of the four $\pi\pi$* singlet states are given by OM*x* and INDO/S with the expected accuracy of typically 0.4 eV; for example, the first bright B_{1u} transition (TBE 5.28 eV) responsible for the first strong peak in the spectrum is calculated somewhat too high with OM*x* and INDO/S (by 0.24−0.43 eV). The results for the four lowest triplet states are largely analogous and will thus not be discussed in detail, except for noting that the OM*x* methods reproduce the TBE values to within 0.2 eV for all of these states except 1^3A_u (TBE 2.62 eV, OM*x* too large by 0.41−0.61 eV).

Formamide, acetamide, and propanamide have analogous valence excited states. The lowest excited singlet is an nπ* state ($1^1A''$, TBE 5.63−5.72 eV) followed by a $\pi\pi$* state ($2^1A'$, TBE 7.20−7.39 eV) and another high-lying $\pi\pi$* state ($2^1A'$, around 10 eV or higher). The triplet manifold begins with a nπ* state ($1^3A''$, TBE 5.36−5.45 eV) followed by a nearby $\pi\pi$* state ($1^3A'$, TBE 5.74−5.90 eV). The energies of the nπ* singlet and triplet states in these primary amides are generally underestimated by all semiempirical methods (e.g., in OM*x* by 0.64−1.10 eV and in INDO/S by 0.43−1.43 eV). Likewise, the energies of the $\pi\pi$* triplet state are underestimated in OM*x* by similar amounts (0.71−1.03 eV) such that the OM*x* triplet−triplet gaps are realistic, whereas INDO/S gives gaps that are too large. Finally, for the first $\pi\pi$* singlet state, the OM*x* and INDO/S results seem reasonable (OM*x* too low by 0.26−0.51 eV, INDO/S within 0.26 eV).

Cytosine, Thymine, Uracil, and Adenine. The singlet excited states of these nucleobases complete our benchmark set. We focus in the discussion on the OM*x* results because the MNDO/AM1/PM3 excitation energies are generally much too low as usual, whereas the INDO/S results are mostly in the same range as the OM*x* results but appear less regular as compared to the TBE values.

The valence excited states of cytosine consist of four $\pi\pi$* (A′) states and two nπ* (A″) states, which are all dominated by singly excited configurations. The OM*x* energies of these states are consistently lower than the available TBE values, typically by about 0.4−0.5 eV (with individual deviations in the range of 0.27−0.68 eV). Because these deviations are fairly uniform, the same state ordering is obtained from OM*x* as from the ab initio reference calculations. There is one caveat, however: according to the TBE values, the lowest excited state of cytosine is a $\pi\pi$* state at 4.66 eV followed by a nearby nπ* state at 4.87 eV, whereas these two states are essentially degenerate at the OM*x* level (within 0.02 eV).

Thymine and uracil share the same heterocyclic ring structure and differ only in one methyl substituent, and hence their excited states are similar in character and can be discussed together. There are four $\pi\pi$* (A′) and four nπ* (A″) valence excited singlet states. According to the available TBE values, the state ordering is $1^1A'' < 2^1A' < 2^1A'' < 3^1A'$ in both molecules, with an nπ* state being lowest. The same state ordering is found in the OM*x* calculations, which generally underestimate the excitation energies in thymine and uracil, by 0.14−0.53 eV for the lowest two states around 5 eV and by 0.40−0.85 eV for the remaining three or four states with TBE values between 6−7 eV.

In adenine, the three lowest singlet states are close in energy: the first nπ* state ($1^1A''$, TBE 5.12 eV) is followed by two essentially degenerate $\pi\pi$* states ($2^1A'$ and $3^1A'$, TBE 5.25 eV) and a second nπ* state ($2^1A''$, TBE 5.75 eV). The OM*x* energies are generally smaller than these TBE values, as in the case of the other nucleobases, but the deviations are less uniform. The main difference is that the two $\pi\pi$* states are not degenerate, but show a substantial split in OM*x* (by 0.56−0.64 eV). This changes the state ordering such that the lowest state in the OM*x* calculations is a $\pi\pi$* state ($2^1A'$) followed by the close-lying $1^1A''$ and $3^1A'$ states. The energy of the fourth excited state ($2^1A''$) is underestimated in OM*x* to a similar extent as in the other nucleobases (by 0.59−0.70 eV).

5. One-Electron Properties

Most of the benchmarking activities for electronically excited states in the literature address excitation energies, although one-electron properties such as oscillator strengths and excited-state dipole moments could also serve as a sensitive probe for the quality of computational methods. However, as compared to excitation energies, these properties are known to converge more slowly upon basis set extension,[14,15,61] and we have therefore not yet derived corresponding theoretical best estimates for our benchmark set. Available reference data include a range of published ab initio results as well as MS-CASPT2/TZVP and CC2/TZVP values calculated for the benchmark molecules in our previous work.[14] These data can be used to evaluate the performance of our semiempirical results, which were obtained using the standard active spaces and CI treatments described in section 2 (CISDTQ for NDDO-based methods, CIS for INDO/S). Different choices for the active space and the CI treatment would affect the semiempirical results, of course, but these changes are usually rather minor for reasonable alternative choices.

5.1. Oscillator Strengths. Table S8 (Supporting Information) lists the computed oscillator strengths for all optically active states in our benchmark. It contains ab initio values previously collated from the literature,[14] published CASPT2 data from the Roos group, and our own results from MS-CASPT2/TZVP and CC2/TZVP[14] as well as semiempirical (MNDO, AM1, PM3, OM1, OM2, OM3, INDO/S) calculations. Generally speaking, there is broad qualitative agreement between the different sets of results, with a proper distinction between strong, medium, and weak transitions. For the low-lying excited states, the majority of the semiempirical oscillator strengths are comparable in magnitude to the MS-CASPT2 and CC2 results, while larger deviations are sometimes found for high-lying bright states where the ab initio reference data also often show some scatter. As expected, n → π^* transitions are normally weak and thus have low oscillator strengths both at the ab initio and at the semiempirical level. We refrain from detailed individual comparisons at this point and provide a more quantitative statistical evaluation in section 6.2.

5.2. Dipole Moments. Table S9 (Supporting Information) lists ground-state as well as excited-state dipole moments obtained from published CASPT2 work in the Roos group and from our own MS-CASPT2/TZVP,[14] RI-CC2/TZVP, and semiempirical calculations. Coupled-cluster CC2 results were determined using unrelaxed densities with the RICC2[62−65] program of the TURBOMOLE package. As pointed out before,[15] the two sets of CASPT2-based results in Table S9 often differ appreciably, and the CC2 dipole moments correlate only roughly with those from the published CASPT2 and from our own MS-CASPT2/TZVP calculations (with correlation coefficients of 0.8341 and 0.8705, respectively). This scatter in the reference data calls for some caution in the assessment of the semiempirical results.

For all semiempirical methods, the computed ground-state dipole moments agree reasonably well with the ab initio results. The mean absolute deviations are around 0.3 D and thus of similar magnitude as in previous comparisons with experiment.[41] The situation is much less satisfactory for the excited-state dipole moments. The MNDO, AM1, and PM3 results are reasonable for low-lying n → π^* states (e.g., in pyridine with deviations of less than 0.2 D), but they often also lie far away from the range spanned by the three sets of reference data (see above), especially for high-lying states where deviations up to 7 D are encountered (e.g., in uracil and thymine); in these latter cases, the composition of the excited-state wave function is different from the ab initio reference. The OM*x* results also scatter around the range of the reference values, but to a lesser extent. Again, there are low-lying states that are well described (e.g., in cyclopentadiene and pyrrole), while there are outliers for high-lying states especially of the nucleobases (e.g., in thymine with deviations reaching 4 D). Finally, for technical reasons, INDO/S and INDO/S2 dipole moments have been computed only for singlet states; they differ because of the use of different C_{kk} and D_{kk} parameters in the CM2 approach.[35] Again the INDO/S and INDO/S2 results deteriorate for the high-lying states (e.g., in thymine with deviations of 8−9 D from MS-CASPT2/TZVP). A more quantitative assessment of the semiempirical excited-state dipole moments is given in section 6.2.

6. Statistics

In the following, we present a detailed statistical evaluation of the computed vertical excitation energies and one-electron properties for all semiempirical methods studied. As reference data, we mainly use theoretical best estimates of vertical excitation energies and MS-CASPT2/TZVP oscillator strengths and dipole moments.

6.1. Vertical Excitation Energies. Table 5 summarizes statistical results for singlet excited states, both for the full set of 104 states and for three subsets (hydrocarbons, CHN compounds, CHO and CHNO compounds). The standard NDDO-based methods (MNDO, AM1, PM3) obviously underestimate the vertical excitation energies systematically and by a large margin, with mean absolute deviations (MAD) above 1.0 eV in each category. The overall MAD values are 1.35, 1.19, and 1.41 eV, respectively, with the worst performance being found in MNDO for hydrocarbons (MAD 1.68 eV) and in PM3 for CHN and oxygen-containing compounds (MAD 1.43−1.48 eV). The OM*x* methods (OM1, OM2, OM3) yield much better results: the vertical excitation energies are still mostly underestimated (overall on average by 0.34, 0.36, and 0.22 eV), with slightly higher mean absolute deviations (0.45, 0.50, and 0.45 eV). Looking at the three subsets separately, the OM*x* methods perform rather uniformly for hydrocarbons (MAD 0.40−0.42 eV) and also rather well for CHN compounds (MAD 0.38−0.46 eV), but somewhat larger errors occur for oxygen-containing compounds especially in OM1 (MAD 0.57 eV) and in OM2 (MAD 0.62 eV). Among the OM*x* methods, the OM3 errors appear to be most uniform, and the systematic underestimation of the energies seems least pronounced (with an overall mean error of −0.22 eV). However, one should not overemphasize this distinction from OM1 and OM2 because the three OM*x* methods perform quite similarly in general. Both INDO/S and its variant INDO/S2 with modified oxygen parameters yield an overall MAD value of 0.51 eV, comparable to OM2 (see above). They also tend to under-

Table 5. Deviations of Semiempirical Vertical Excitation Energies (in eV) for Singlet States with Respect to Theoretical Best Estimates

	MNDO	AM1	PM3	OM1	OM2	OM3	INDO/S	INDO/S2
CH-Containing Molecules								
count[a]	25	25	25	25	25	25	25	25
mean	−1.68	−1.36	−1.29	−0.27	−0.29	−0.25	−0.20	−0.20
abs. mean	1.68	1.36	1.29	0.42	0.41	0.40	0.42	0.42
std. dev.	1.89	1.55	1.48	0.52	0.51	0.49	0.47	0.47
max. (+) dev.			0.04	0.76	0.66	0.72	0.81	0.81
max. (−) dev.	3.95	3.30	3.21	1.34	1.22	1.19	0.84	0.84
CHN-Containing Molecules								
count[a]	43	43	43	43	43	43	42	42
mean	−1.11	−1.01	−1.43	−0.27	−0.21	−0.09	−0.34	−0.34
abs. mean	1.22	1.13	1.43	0.38	0.45	0.46	0.48	0.48
std. dev.	1.41	1.27	1.53	0.49	0.54	0.57	0.63	0.63
max. (+) dev.	0.69	0.85		0.37	1.42	1.76	0.92	0.92
max. (−) dev.	3.00	2.60	2.58	1.16	1.09	1.08	1.45	1.45
CHO and CHNO-Containing Molecules								
count[a]	36	36	36	36	36	36	36	36
mean	−1.27	−1.09	−1.46	−0.47	−0.55	−0.35	−0.12	0.22
abs. mean	1.29	1.14	1.48	0.57	0.62	0.47	0.62	0.61
std. dev.	1.44	1.25	1.63	0.64	0.69	0.52	0.88	1.04
max. (+) dev.	0.18	0.42	0.19	0.56	0.58	0.60	2.79	4.24
max. (−) dev.	3.04	2.61	2.53	1.19	1.39	1.06	1.37	0.90
All Molecules								
count[a]	104	104	104	104	104	104	103	103
mean	−1.30	−1.12	−1.40	−0.34	−0.36	−0.22	−0.23	−0.11
abs. mean	1.35	1.19	1.41	0.45	0.50	0.45	0.51	0.51
std. dev.	1.55	1.34	1.55	0.55	0.59	0.54	0.70	0.77
max. (+) dev.	0.69	0.85	0.19	0.76	1.42	1.76	2.79	4.24
max. (−) dev.	3.95	3.30	3.21	1.34	1.39	1.19	1.45	1.45

[a] Total number of states considered.

Table 6. Deviations of Semiempirical Vertical Excitation Energies (in eV) for nπ* and $\pi\pi$* Singlet States with Respect to Theoretical Best Estimates up to 6 eV

	MNDO	AM1	PM3	OM1	OM2	OM3	INDO/S	INDO/S2
nπ*								
count[a]	33	33	33	33	33	33	32	32
mean	−0.62	−0.61	−1.08	−0.26	−0.18	−0.14	−0.31	−0.14
abs. mean	0.77	0.80	1.10	0.38	0.46	0.43	0.48	0.35
std. dev.	0.92	0.89	1.20	0.50	0.56	0.54	0.61	0.44
max. (+) dev.	0.69	0.85	0.19	0.40	1.42	1.76	0.92	0.92
max. (−) dev.	1.86	1.52	1.95	1.19	1.07	0.92	1.37	1.06
$\pi\pi$*								
count[a]	25	25	25	25	25	25	25	25
mean	−1.42	−1.18	−1.30	−0.27	−0.26	−0.11	−0.19	−0.17
abs. mean	1.42	1.20	1.32	0.44	0.40	0.39	0.38	0.38
std. dev.	1.59	1.34	1.48	0.50	0.46	0.45	0.45	0.46
max. (+) dev.	0.01	0.19	0.17	0.76	0.66	0.81	0.81	0.81
max. (−) dev.	2.67	2.19	2.30	1.06	1.02	0.92	0.92	0.92

[a] Total number of states considered.

estimate the vertical excitation energies (overall mean errors of −0.23 and −0.11 eV, respectively), but their results scatter more strongly than the OM*x* results (overall standard deviations of 0.70−0.77 eV as compared to 0.54−0.59 eV for OM*x*). Concerning the subsets, our evaluation confirms[66] that INDO/S tends to give somewhat higher errors for oxygen-containing compounds (MAD 0.62 eV), but the claimed improvement by the reparametrized INDO/S2 variant[35] is hardly seen in our data (MAD 0.61 eV).

The statistics in Table 5 include several singlet states above 6 eV for which TBE values are available. Because we expect in general that minimal-basis-set semiempirical calculations on valence excited states will become less appropriate the higher the energy, we have performed a second evaluation restricted to singlet states up to 6 eV, considering $\pi\pi$* and nπ* states separately. The results are shown in Table 6. It is obvious that the restriction to energies below 6 eV improves the statistics especially for MNDO, AM1, and PM3, where the MAD values for $\pi\pi$* states remain in the 1.2−1.4 eV range while those for the nπ* states amount to 0.8−1.1 eV. The improvement is less pronounced for the OM*x* and INDO/S methods, which treat the two types of excitation in a fairly balanced manner: for example, the MAD values in OM*x* range from 0.39−0.44 eV for $\pi\pi$* states and from 0.38−0.46 eV for nπ* states; the corresponding values for INDO/S are 0.38 and 0.48 eV. Comparison of the INDO/S and INDO/S2 statistics indicates that the reparametrization for oxygen

Table 7. Deviations of Semiempirical Vertical Excitation Energies (in eV) for Triplet States with Respect to Theoretical Best Estimates

	MNDO	AM1	PM3	OM1	OM2	OM3	INDO/S	INDO/S2
CH-Containing Molecules								
count[a]	27	27	27	27	27	27	27	27
mean	−1.84	−1.50	−1.44	−0.42	−0.37	−0.35	−0.38	−0.38
abs. mean	1.84	1.50	1.44	0.45	0.42	0.41	0.57	0.57
std. dev.	1.96	1.61	1.55	0.55	0.54	0.52	0.74	0.74
max. (+) dev.				0.36	0.55	0.54	0.85	0.85
max. (−) dev.	3.66	3.06	2.98	1.31	1.21	1.17	1.63	1.63
CHN-Containing Molecules								
count[a]	20	20	20	20	20	20	20	20
mean	−1.24	−1.10	−1.44	−0.39	−0.27	−0.05	−0.43	−0.43
abs. mean	1.29	1.16	1.44	0.43	0.44	0.49	0.70	0.70
std. dev.	1.50	1.32	1.55	0.56	0.52	0.56	0.92	0.92
max. (+) dev.	0.54	0.47		0.19	0.64	1.08	0.86	0.86
max. (−) dev.	2.63	2.28	2.43	1.05	0.92	0.87	2.01	2.01
CHO and CHNO-Containing Molecules								
count[a]	16	16	16	16	16	16	16	16
mean	−1.35	−1.07	−1.34	−0.57	−0.53	−0.39	−0.04	0.60
abs. mean	1.38	1.15	1.37	0.62	0.60	0.49	0.72	1.02
std. dev.	1.54	1.29	1.56	0.73	0.70	0.57	0.96	1.44
max. (+) dev.	0.14	0.39	0.20	0.41	0.59	0.61	2.49	4.30
max. (−) dev.	2.63	2.26	2.40	1.04	1.03	0.89	1.54	1.48
All Molecules								
count[a]	63	63	63	63	63	63	63	63
mean	−1.52	−1.27	−1.41	−0.45	−0.38	−0.26	−0.31	−0.15
abs. mean	1.55	1.30	1.42	0.49	0.47	0.45	0.65	0.72
std. dev.	1.72	1.44	1.55	0.61	0.58	0.54	0.86	1.01
max. (+) dev.	0.54	0.47	0.20	0.41	0.64	1.08	2.49	4.30
max. (−) dev.	3.66	3.06	2.98	1.31	1.21	1.17	2.01	2.01

[a] Total number of states considered.

has not affected the $\pi\pi^*$ states much, but has improved the energies of the $n\pi^*$ states below 6 eV (MAD 0.35 eV).

Table 7 presents the statistical results for triplet excited states, again for all states available and for three subsets (hydrocarbons, CHN compounds, CHO and CHNO compounds). The total number of states with TBE values is smaller than in the singlet case (63 vs 104), but still large enough for meaningful evaluations. As compared to the singlets, the overall performance for triplets is essentially the same for OM*x* (MAD 0.45−0.49 eV vs 0.45−0.50 eV), but somewhat worse for MNDO/AM1/PM3 (MAD 1.30−1.55 eV vs 1.19−1.41 eV) and also for INDO/S and INDO/S2 (MAD 0.65−0.72 eV vs 0.51 eV), but the general trends remain the same. The semiempirical excitation energies again tend to be too low, even slightly more so than in the case of the singlets (see the overall mean errors). Considering the subsets, the largest mean absolute deviations occur for the hydrocarbons in MNDO/AM1/PM3 and for the oxygen-containing compounds for OM1, OM2, INDO/S, and INDO/S2 (again in analogy to the singlets). Somewhat surprisingly, we find for the oxygen-containing compounds that INDO/S2 performs worse than INDO/S for the triplet states (MAD 1.02 eV vs 0.72 eV) despite the reparametrization, while OM3 again appears to be most balanced and performs best (MAD 0.47 eV vs 0.57−0.62 eV for OM1 and OM2).

For many photophysical and photochemical processes, the energy difference between excited states is of crucial importance. Table 8 provides a statistical evaluation for three such differences that are particularly relevant, between the two lowest singlets, between the two lowest triplets, and between the lowest singlet and triplet states. It is obvious that the OM*x* methods show by far the best performance, with mean absolute deviations from the TBE values of 0.40−0.43, 0.23−0.32, and 0.20−0.21 eV for the S_2−S_1, T_2−T_1, and S_1−T_1 energy gaps, respectively. The corresponding MAD values for the other semiempirical methods (MNDO, AM1, PM3, INDO/S, INDO/S2) are significantly higher and generally lie in the range between 0.4−0.6 eV for all three energy gaps considered. The associated mean errors indicate that these energy differences are normally underestimated in MNDO, AM1, and PM3 (on average by 0.3−0.4 eV for T_2−T_1 and by 0.1−0.2 eV otherwise), while they are generally overestimated in INDO/S and INDO/S2 (on average by 0.4−0.6 eV for T_2−T_1 and by 0.2−0.3 eV otherwise); by contrast, the OM*x* energy gaps are too low on average by only 0.1−0.2 eV for T_2−T_1 and scatter around the TBE values for S_2−S_1 and S_1−T_1 (on average within 0.1 eV). These data show that the OM*x* methods predict the energy sequence within the excited states much better than the other semiempirical methods studied.

The overall performance for all singlet and triplet states considered can be judged from the correlation plots shown in Figure 1. The OM*x* results are rather close to the ideal correlation line with unit slope and yield reasonably high correlation coefficients ($r = 0.9391−0.9520$). The MNDO energies are generally too low and scatter considerably ($r = 0.8107$), while the AM1 and PM3 energies are also too low but more regular ($r = 0.8574−0.8838$). The INDO/S and INDO/S2 data also show some scatter around the TBE values

Table 8. Deviations of Semiempirical S_2-S_1, T_2-T_1, and S_1-T_1 Energy Differences (in eV) with Respect to Theoretical Best Estimates

	MNDO	AM1	PM3	OM1	OM2	OM3	INDO/S	INDO/S2
S_2-S_1[a]								
count[b]	27	27	27	27	27	27	27	27
mean	−0.13	−0.14	−0.24	0.04	−0.11	−0.10	0.27	0.20
abs. mean	0.64	0.52	0.57	0.42	0.40	0.43	0.64	0.55
std. dev.	0.83	0.68	0.67	0.47	0.46	0.49	0.88	0.80
max. (+) dev.	1.35	1.14	1.02	0.89	0.70	0.79	2.26	2.34
max. (−) dev.	1.65	1.42	1.29	0.75	0.97	1.02	1.22	1.22
T_2-T_1[c]								
count[b]	19	19	19	19	19	19	19	19
mean	−0.30	−0.26	−0.37	−0.12	−0.17	−0.14	0.45	0.57
abs. mean	0.47	0.43	0.48	0.23	0.30	0.32	0.58	0.72
std. dev.	0.61	0.52	0.53	0.30	0.36	0.41	0.88	1.17
max. (+) dev.	1.46	1.02	0.50	0.46	0.60	0.70	2.52	3.81
max. (−) dev.	1.00	0.86	0.88	0.63	0.67	1.12	0.41	0.41
S_1-T_1[d]								
count[b]	20	20	20	20	20	20	20	20
mean	−0.05	−0.06	−0.10	0.04	−0.07	−0.04	0.28	0.25
abs. mean	0.44	0.38	0.38	0.20	0.21	0.21	0.61	0.59
std. dev.	0.61	0.52	0.47	0.25	0.26	0.25	0.76	0.75
max. (+) dev.	1.52	1.11	0.81	0.54	0.46	0.39	1.80	1.80
max. (−) dev.	1.07	1.07	0.86	0.31	0.55	0.50	0.92	0.98

[a] Energy difference between the two lowest singlet excited states. [b] Total number of molecules considered. [c] Energy difference between the two lowest triplet excited states. [d] Energy difference between the lowest singlet and triplet excited states.

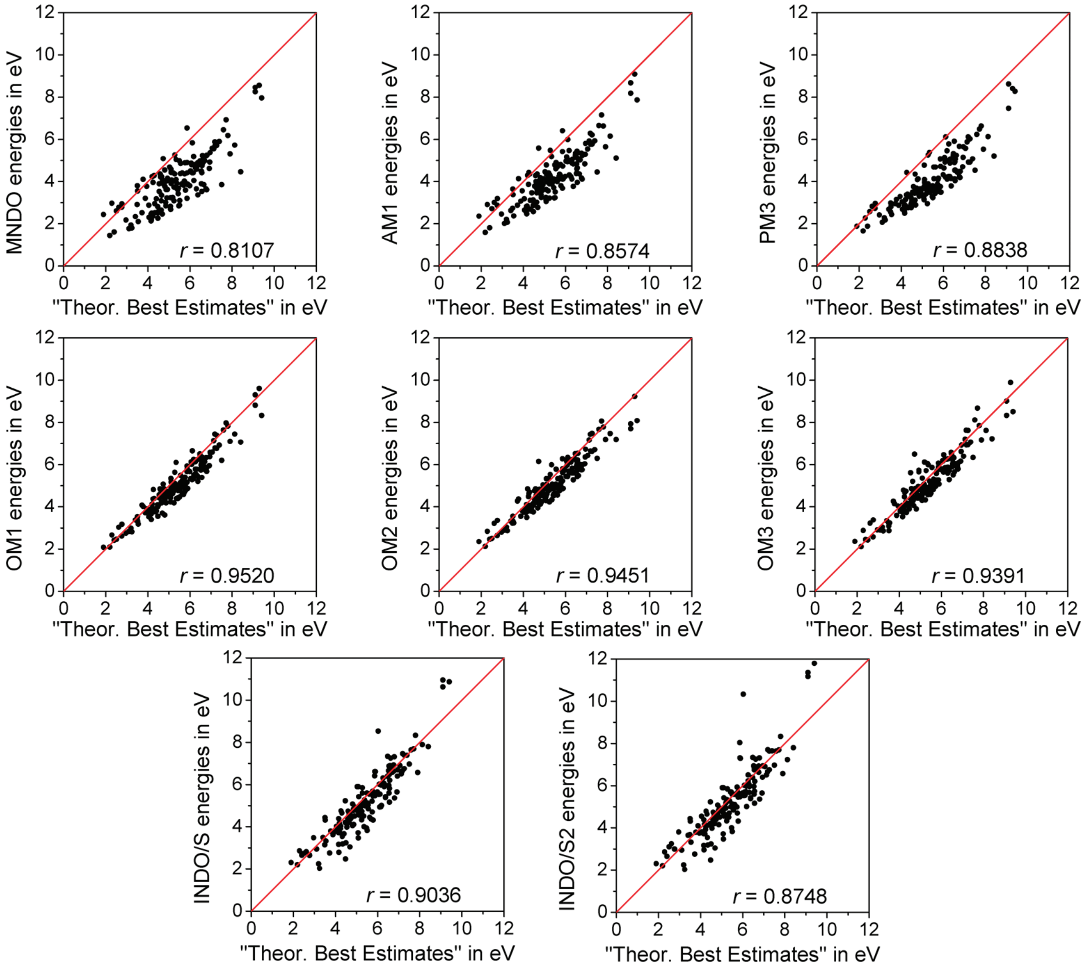


Figure 1. Correlation plots of vertical excitation energies for all states considered in this study using theoretical best estimates as reference data.

($r = 0.8748-0.9036$). Similar separate correlation plots for the singlet and triplet states are presented in the Supporting Information (Figures S1 and S2).

Figure 2 provides a visual summary in the form of a histogram plot for the deviations of the computed vertical excitation energies from the TBE values of all states below

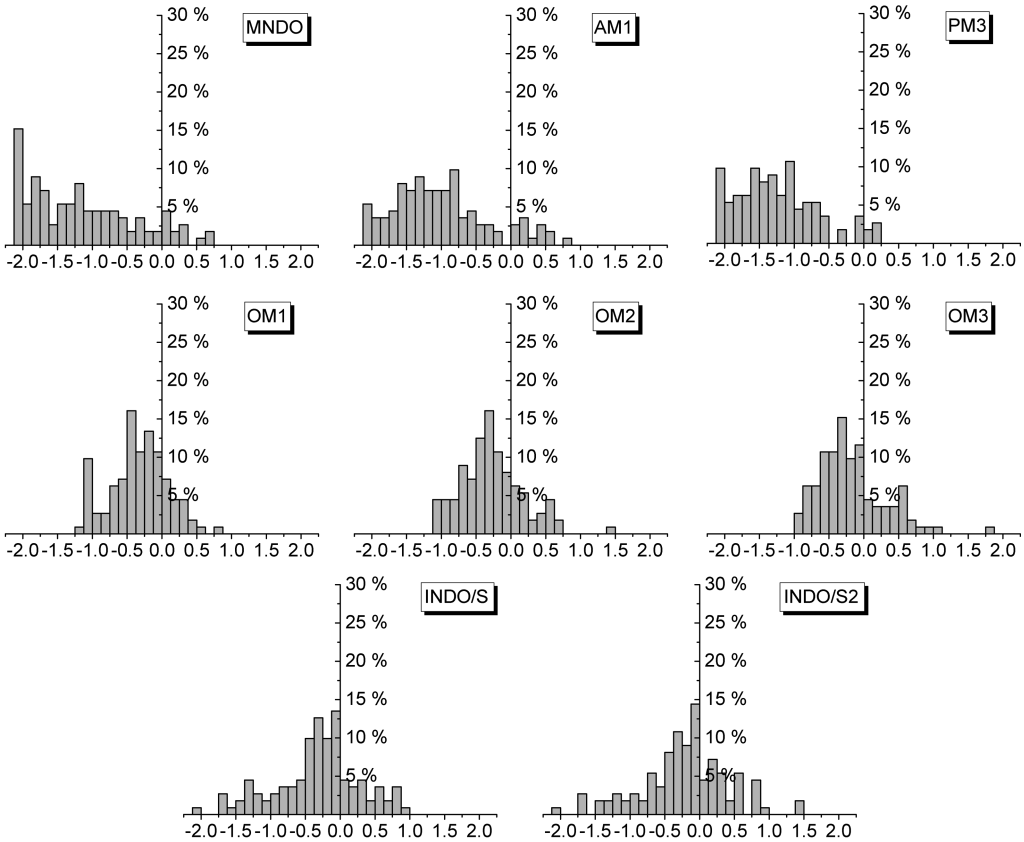


Figure 2. Histogram plots for the deviations of the calculated vertical excitation energies from the theoretical best estimates for all states below 6 eV considered in this study. In the case of MNDO, AM1, and PM3, the left-hand column collects all states whose energies are underestimated by more than 2 eV.

6 eV. Analogous histograms are given in the Supporting Information (Figures S3 and S4) separately for the nπ* and $\pi\pi$* states. On the basis of these histograms and the statistical data presented in this section, it is possible to rank the semiempirical methods investigated here according to their overall performance to describe excitation energies. The OM*x* methods are clearly the best choice, with minor differences in the performance of the three variants, with a slight edge of OM2 and OM3 over OM1. INDO/S and INDO/S2 also perform reasonably well on average, specially for low-lying states, but show a considerably wider error distribution. Finally, the MNDO, AM1, and PM3 methods that are well established and widely used for ground states are least suitable for electronically excited states in their standard ground-state parametrization.

Among the previously studied DFT-based methods,[15] TD-BP86 and TD-BHLYP show mean absolute deviations from the theoretical best estimates similar to those of the OM*x* methods, while TD-B3LYP and especially DFT/MRCI perform better (MAD for singlets, TD-B3LYP 0.27 eV, DFT/MRCI 0.22 eV, OM*x* 0.45−0.50 eV; triplets, TD-B3LYP 0.45 eV, DFT/MRCI 0.25 eV, OM*x* 0.45−0.49 eV). This is not unexpected in view of the computational costs, which are higher by about 3 orders of magnitude for TD-B3LYP (see section 2) and by even more in the case of DFT/MRCI. The DFT-based results have also been evaluated against other ab initio reference data.[15] The corresponding statistical evaluations for the semiempirical methods are documented in the Supporting Information (Tables S10−S19) but will not be discussed here.

6.2. One-Electron Properties. As mentioned before, we have not yet established theoretical best estimates for the oscillator strengths and state dipole moments in our benchmark set. Therefore, we use three sets of ab initio reference data: the comparisons with our own MS-CASPT2/TZVP results[14] are presented here, while those with the published CASPT2 data mainly from the Roos group and with our own CC2/TZVP data[15] are given in the Supporting Information (Figures S5−S8, Tables S20 and S21). The qualitative conclusions from these comparisons are the same for each set of reference data.

Table 9 summarizes the statistical evaluation of the calculated oscillator strengths for all dipole-allowed transitions, and Figure 3 shows the corresponding correlation plots against the MS-CASPT2/TZVP reference data. Visual inspection of these plots indicates that the OM*x* oscillator strengths scatter around the ideal correlation line with unit slope and correlate reasonably well with the reference data (r = 0.9032−0.9425). The MNDO/AM1/PM3 oscillator strengths tend to be too low and correlate less well (r = 0.8690−0.9235), while INDO/S and INDOS/2 tend to overestimate the reference values and have low correlation coefficients (r = 0.8781−0.8882). These visual impressions are corroborated by the statistical results. For the whole set of dipole-allowed transitions considered, the mean absolute deviations from the MS-CASPT2/TZVP oscillator strengths

Table 9. Deviations of Semiempirical Oscillator Strengths (in Length Representation) of Dipole-Allowed States with Respect to ab Initio Reference Data

	MNDO	AM1	PM3	OM1	OM2	OM3	INDO/S	INDO/S2
Published CASPT2 Results[a]								
count[b]	100	100	100	100	100	100	98	98
mean	−0.085	−0.058	−0.066	0.014	0.006	0.027	0.075	0.070
abs. mean	0.114	0.100	0.102	0.081	0.075	0.102	0.148	0.150
std. dev.	0.203	0.184	0.182	0.146	0.119	0.181	0.222	0.226
max. (+) dev.	0.511	0.564	0.406	0.309	0.307	0.831	0.707	0.707
max. (−) dev.	0.870	0.844	0.831	0.830	0.432	0.822	0.627	0.750
MS-CASPT2/TZVP Results[a]								
count[b]	100	100	100	100	100	100	98	98
mean	−0.100	−0.073	−0.081	−0.001	−0.009	0.011	0.058	0.053
abs. mean	0.131	0.113	0.105	0.086	0.077	0.094	0.134	0.137
std. dev.	0.197	0.166	0.155	0.134	0.112	0.152	0.203	0.208
max. (+) dev.	0.346	0.342	0.314	0.421	0.393	0.609	0.659	0.659
max. (−) dev.	0.711	0.529	0.516	0.515	0.320	0.507	0.671	0.794
CC2/TZVP Results[a]								
count[b]	100	100	100	100	100	100	98	98
mean	−0.035	−0.008	−0.015	0.064	0.057	0.077	0.124	0.120
abs. mean	0.092	0.087	0.084	0.102	0.082	0.118	0.151	0.150
std. dev.	0.171	0.153	0.146	0.158	0.130	0.189	0.229	0.230
max. (+) dev.	0.490	0.543	0.385	0.440	0.387	0.810	0.845	0.845
max. (−) dev.	0.795	0.670	0.670	0.475	0.312	0.467	0.414	0.537

[a] See ref 14. [b] Total number of states considered.

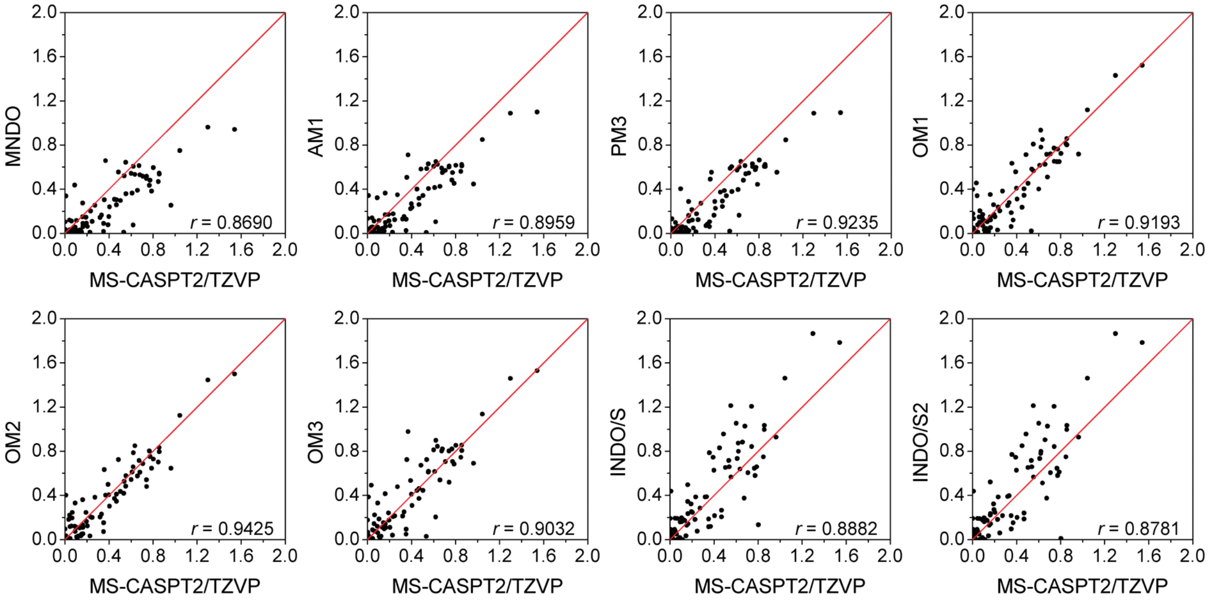


Figure 3. Correlation plots of oscillator strengths for all dipole-allowed transitions using MS-CASPT2/TZVP results as reference data.

are 0.077−0.094 for OMx, 0.105−0.131 for MNDO/AM1/PM3, and 0.134−0.137 for INDO/S and INDO/S2. The overall mean deviations confirm that the OMx oscillator strengths are on average close to the MS-CASPT2/TZVP reference data (within 0.01), while the results from MNDO/AM1/PM3 are mostly too low (on average by 0.07−0.10) and those from INDO/S and INDO/S2 mostly too high (on average by 0.05−0.06).

Table 10 presents the statistical evaluation of all nonzero state dipole moments in our benchmark set, and Figure 4 shows the corresponding correlation plots against the MS-CASPT2/TZVP reference data. It is evident that the semiempirical results often differ rather strongly from the ab initio data. The performance of OMx (MAD 0.68−0.71 D) and MNDO/AM1/PM3 (MAD 0.66−0.76 D) is similar in this case and better than that of INDO/S and INDO/S2 (MAD 1.08−1.19 D). The overall mean errors suggest that state dipole moments tend to be underestimated by OMx (on average by 0.10−0.23 D) and somewhat more by MNDO/AM1/PM3 (on average by 0.27−0.37 D), while INDO/S and INDO/S2 tend to overestimate them (on average by 0.40−0.49 D). These trends are also apparent from the correlation plots.

Summarizing the statistical evaluations for the oscillator strengths and state dipole moments, the OMx methods show the best overall performance for these one-electron properties, without any pronounced preference for one of the three variants. To put the OMx results into perspective, we note that rather similar deviations from the ab initio reference data have also been found for DFT-based methods, both for oscillator strengths (MAD: TD-B3LYP 0.113, DFT/MRCI

Table 10. Deviations of Semiempirical State Dipole Moments (in D) with Respect to ab Initio Reference Data[a]

	MNDO	AM1	PM3	OM1	OM2	OM3	INDO/S	INDO/S2
Published CASPT2 Results[b]								
count[c]	142	142	142	142	142	142	104	104
mean	−0.61	−0.58	−0.51	−0.47	−0.11	−0.34	0.11	0.20
abs. mean	0.93	1.03	1.02	0.94	0.86	0.92	1.25	1.35
std. dev.	1.44	1.53	1.54	1.44	1.38	1.43	2.16	2.26
max. (+) dev.	2.37	3.81	2.98	5.22	5.75	5.49	9.73	9.92
max. (−) dev.	7.17	7.28	7.21	6.58	5.82	6.10	4.92	5.08
MS-CASPT2/TZVP Results[b]								
count[c]	142	142	142	142	142	142	104	104
mean	−0.37	−0.34	−0.27	−0.23	−0.13	−0.10	0.40	0.49
abs. mean	0.66	0.76	0.72	0.68	0.68	0.71	1.08	1.19
std. dev.	0.91	1.05	1.00	1.00	1.07	1.07	2.03	2.11
max. (+) dev.	3.52	3.73	3.98	5.97	5.64	5.38	10.01	9.81
max. (−) dev.	4.91	4.60	4.91	4.25	4.68	4.37	4.90	4.30
RI-CC2/TZVP Results[b]								
count[c]	140	140	140	140	140	140	103	103
mean	−0.16	−0.13	−0.07	−0.04	0.33	0.09	0.66	0.75
abs. mean	0.83	0.88	0.80	0.74	0.80	0.81	1.16	1.26
std. dev.	1.11	1.19	1.12	1.15	1.25	1.22	2.18	2.32
max. (+) dev.	4.25	4.53	4.57	6.81	5.35	5.01	10.62	9.75
max. (−) dev.	2.93	3.44	3.78	1.85	4.90	4.59	1.98	2.13

[a] Ground and excited states. Only singlet states in the case of INDO/S and INDO/S2. [b] See ref 14. [c] Total number of states considered.

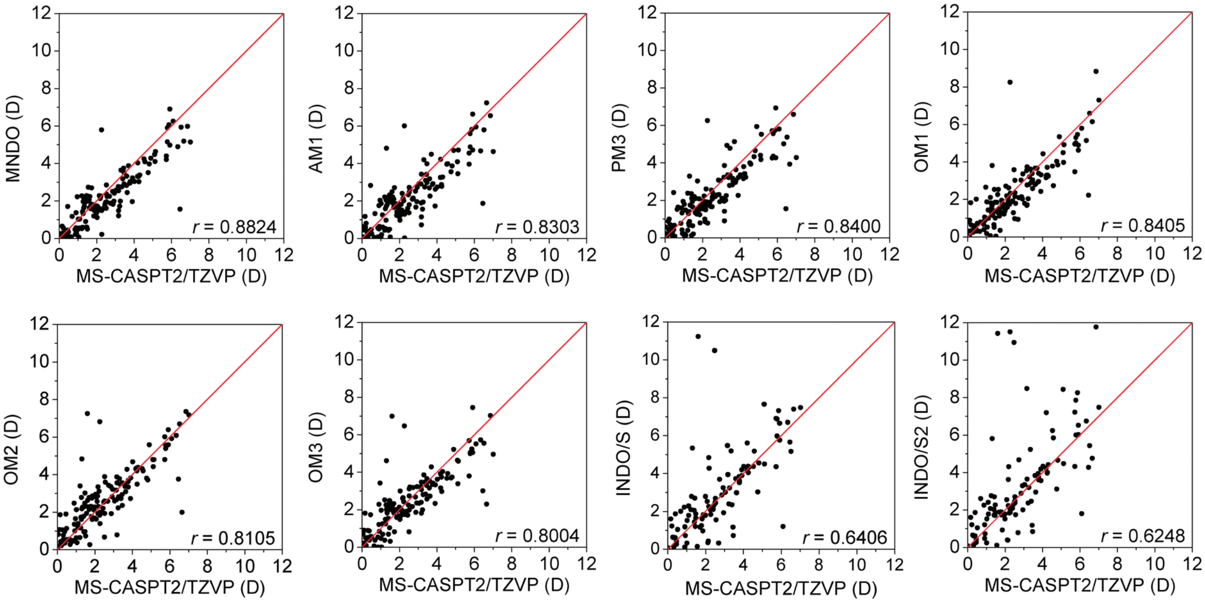


Figure 4. Correlation plots of nonzero state dipole moments (in D) for all states considered, using MS-CASPT2/TZVP results as reference data.

0.069, OM*x* 0.077−0.094) and for state dipole moments (MAD: TD-B3LYP 0.59 D, DFT/MRCI 0.58 D, OM*x* 0.68−0.71 D).[15]

7. Conclusions

We have performed a comprehensive validation study for eight semiempirical methods to assess their performance in describing vertical excitation energies and one-electron properties of electronically excited states. The semiempirical results were compared against ab initio reference data for a recently developed benchmark set of 28 medium-sized organic molecules, focusing on the published theoretical best estimates for excitation energies and MS-CASPT2/TZVP oscillator strengths and excited-state dipole moments.[14] The benchmark set includes only valence excited states because Rydberg states cannot be described by semiempirical methods due to the use of a minimal basis set.

The standard ground-state methods MNDO, AM1, and PM3 strongly underestimate the reference excitation energies, typically by more than 1 eV, and they often give a wrong order of the excited states because the errors are not uniform. These methods thus seem unsuitable for excited-state work unless one is prepared to undertake a system-specific reparametrization.

The INDO/S method has been parametrized for spectroscopic purposes and has often been applied to compute electronic spectra at the semiempirical level. In the current benchmark, it performs reasonably well, with typical errors of 0.5 eV for excited singlets and 0.6−0.7 eV for triplet states. By design, INDO/S does not properly account for states with significant double-excitation character, and it is thus not surprising that the INDO/S errors are smaller for low-lying than for high-lying singlet states. The INDO/S2 variant has been reparametrized for oxygen,[35] which indeed improves the results for low-lying singlet states of oxygen-containing compounds, but turns out to be detrimental for the high-lying singlets and the triplet states of our benchmark molecules. Hence, INDO/S2 cannot be recommended as a general-purpose alternative to INDO/S.

The NDDO-based orthogonalization-corrected methods OM1, OM2, and OM3 show the best overall performance in the present benchmark. They give vertical excitation energies with typical errors of 0.4−0.5 eV both for valence excited singlet and for triplet states, and they predict the order of the excited states and the gaps between the low-lying states more reliably than the other semiempirical methods. The OM*x*/CISDTQ wave functions for valence excited states normally show a qualitatively similar composition to the corresponding ab initio CASSCF and CASPT2 wave functions (with regard to the leading configurations). The OM*x* methods provide reasonable one-electron properties (oscillator strengths and excited-state dipole moments), which are overall somewhat closer to the ab initio reference data than those obtained from the other semiempirical methods. The performance of the three OM*x* variants is generally quite similar, although OM2 and OM3 would seem to have a slight edge over OM1 in an overall assessment.

Conceptually, the OM*x* methods are superior to INDO/S because of the use of the more refined NDDO integral approximation, and to all other semiempirical methods considered here because of the explicit inclusion of corrections that mimic Pauli exchange repulsion and thus cause an unsymmetric splitting of bonding and antibonding levels. These advances are believed to contribute to the good OM*x* performance for electronically excited states, which is achieved despite the fact that the adjustable OM*x* parameters have been determined by calibrating purely against ground-state reference data. One may anticipate that further improvements are possible through a general-purpose reparametrization of the OM*x* methods that includes ground-state and excited-state reference data in a balanced manner. In such reparametrization work as well as in other applications, it seems justified to replace the CISDTQ treatment employed here by a more cost-efficient MR-CISD approach, which yields essentially the same results provided that the reference configurations dominate the CI wave functions (e.g., by requiring their combined weight to exceed 85%).

In summary, at the semiempirical level, the OM*x* methods appear to be the best choice for studying excited-state phenomena in large organic chromophores. While the results from the current benchmark support such applications using the existing ground-state parametrization, further improvements in the OM*x* description of electronically excited states may be expected from a balanced reparametrization.

Acknowledgment. This work was supported by the Deutsche Forschungsgemeinschaft (SFB663, project C4). M.R.S.-J. gladly acknowledges support by the Deutscher Akademischer Austauschdienst (DAAD). We thank Prof. Christoph J. Cramer for providing a modified version of the ZINDO program and Dmitro Khoroshun as well as Axel Koslowski for their early contributions to this work.

Supporting Information Available: Additional numerical results for excitation energies and one-electron properties (Tables S1−S9), additional correlation and histogram plots (Figures S1−S8), and additional statistical evaluations (Tables S10−S22). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. *J. Phys. Chem.* **1990**, *94*, 5483–5488.

(2) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218–1226.

(3) Roos, B. O.; Andersson, K.; Fülscher, M. P.; Malmqvist, P.-Å.; Serrano-Andrés, L.; Pierloot, K.; Merchán, M. Multiconfigurational Perturbation Theory: Applications in Electronic Spectroscopy. In *Advances in Chemical Physics: New Methods in Computational Quantum Mechanics*; Prigogine, I., Rice, S. A., Eds.; John Wiley & Sons: New York, 1996; Vol. 93.

(4) Finley, J.; Malmqvist, P.-Å.; Roos, B. O.; Serrano-Andrés, L. *Chem. Phys. Lett.* **1998**, *288*, 299–306.

(5) Christiansen, O.; Koch, H.; Jørgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409–418.

(6) Christiansen, O.; Koch, H.; Jørgensen, P. *J. Chem. Phys.* **1995**, *103*, 7429–7441.

(7) Koch, H.; Christiansen, O.; Jørgensen, P.; de Meras, A. M. S.; Helgaker, T. *J. Chem. Phys.* **1997**, *106*, 1808–1818.

(8) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997–1000.

(9) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009–4037.

(10) Casida, M. E. *J. Mol. Struct. (THEOCHEM)* **2009**, *914*, 3–18.

(11) Grimme, S.; Waletzke, M. *J. Chem. Phys.* **1999**, *111*, 5645–5655.

(12) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

(13) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374–7383.

(14) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110/1–25.

(15) Silva-Junior, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *129*, 104103/1–14.

(16) Sauer, S. P. A.; Schreiber, M.; Silva-Junior, M. R.; Thiel, W. *J. Chem. Theory Comput.* **2009**, *5*, 555–564.

(17) Jacquemin, D.; Wathelet, V.; Perpete, E. A.; Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420–2435.

(18) Goerigk, L.; Moellmann, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4611–4620.

(19) Dreuw, A.; Head-Gordon, M. *J. Am. Chem. Soc.* **2004**, *126*, 4007–4016.

(20) Cordova, F.; Doriol, L. J.; Ipatov, A.; Casida, M. E.; Filippi, C.; Vela, A. *J. Chem. Phys.* **2007**, *127*, 164111/1–18.

(21) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.

(22) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(23) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209–220.

(24) Thiel, W. Semiempirical Methods. In *Handbook of Molecular Physics and Quantum Chemistry*; Wilson, S., Ed.; Wiley: Chicester, UK, 2003; Vol. 2.

(25) Thiel, W. Semiempirical Quantum-Chemical Methods in Computational Chemistry. In *Theory and Applications of Computational Chemistry: The First 40 Years*; Dykstra, C. E., Kim, K., Frenking, G. E., Eds.; Elsevier: Amsterdam, 2005.

(26) Bredow, T.; Jug, K. *Theor. Chim. Acta* **2005**, *113*, 1–14.

(27) Thiel, W. *J. Am. Chem. Soc.* **1981**, *103*, 1425–1431.

(28) Dewar, M. J. S.; Fox, M. A.; Campbell, K. A.; Chen, C. C.; Friedheim, J. E.; Holloway, M. K.; Kim, S. C.; Liescheski, P. B.; Pakiari, A. M.; Tien, T. P.; Zoebisch, E. G. *J. Comput. Chem.* **1984**, *5*, 480–485.

(29) Gonzalez-Lafont, A.; Truong, T. N.; Truhlar, D. G. *J. Phys. Chem.* **1991**, *95*, 4618–4627.

(30) Creatini, L.; Cusati, G. G.; Persico, M. *Chem. Phys.* **2008**, *347*, 492–502.

(31) Ridley, J.; Zerner, M. *Theor. Chim. Acta* **1973**, *32*, 111–134.

(32) Ridley, J. E.; Zerner, M. C. *Theor. Chim. Acta* **1976**, *42*, 223–236.

(33) Zerner, M. C.; Loew, G. H.; Kirchner, R. F.; Muellerwesterhoff, U. T. *J. Am. Chem. Soc.* **1980**, *102*, 589–599.

(34) Kotzian, M.; Rösch, N.; Zerner, M. C. *Theor. Chim. Acta* **1992**, *81*, 201–222.

(35) Li, J.; Williams, B.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *110*, 724–733.

(36) Kolb, M. Thesis, Universität Wuppertal, 1991.

(37) Kolb, M.; Thiel, W. *J. Comput. Chem.* **1993**, *14*, 775–789.

(38) Weber, W. Thesis, Universität Zürich, 1996.

(39) Weber, W.; Thiel, W. *Theor. Chim. Acta* **2000**, *103*, 495–506.

(40) Scholten, M. Thesis, Universität Düsseldorf, 2003.

(41) Otte, N.; Scholten, M.; Thiel, W. *J. Phys. Chem. A* **2007**, *111*, 5751–5755.

(42) Strodel, P.; Tavan, P. *J. Chem. Phys.* **2002**, *117*, 4677–4683.

(43) Wanko, M.; Hoffmann, M.; Strodel, P.; Koslowski, A.; Thiel, W.; Neese, F.; Frauenheim, T.; Elstner, M. *J. Phys. Chem. B* **2005**, *109*, 3606–3615.

(44) Hoffmann, M.; Wanko, M.; Strodel, P.; König, P. H.; Frauenheim, T.; Schulten, K.; Thiel, W.; Tajkhorshid, E.; Elstner, M. *J. Am. Chem. Soc.* **2006**, *128*, 10808–10818.

(45) Keal, T. W.; Koslowski, A.; Thiel, W. *Theor. Chim. Acta* **2007**, *118*, 837–844.

(46) Fabiano, E.; Keal, T. W.; Thiel, W. *Chem. Phys.* **2008**, *349*, 334–347.

(47) Fabiano, E.; Thiel, W. *J. Phys. Chem. A* **2008**, *112*, 6859–6863.

(48) Lan, Z.; Fabiano, E.; Thiel, W. *J. Phys. Chem. B* **2009**, *113*, 3548–3555.

(49) Lan, Z.; Fabiano, E.; Thiel, W. *ChemPhysChem* **2009**, *10*, 1225–1229.

(50) Keal, T. W.; Wanko, M.; Thiel, W. *Theor. Chim. Acta* **2009**, *123*, 145–156.

(51) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110/1–25; Supporting Information deposited as EPAPS Document No. E-JCPSA6-128-032811. For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html.

(52) Thiel, W. *MNDO99, version 6.1*; Mülheim an der Ruhr, Max-Planck-Institut für Kohlenforschung: Germany, 2007.

(53) Zerner, M. C.; Ridley, J. E.; Bacon, A. D.; Head, J. D.; Edwards, W. D.; McKelvey, J.; Cuberson, J. C.; Knappe, P.; Cory, M. G.; Weiner, B.; Baker, J. D.; Parkinson, W. A.; Kannis, D.; Yu, J.; Roesch, N.; Kotzian, M.; Karelson, T. T. M. M.; Zheng, X.; Pearl, G.; Broo, A.; Cullen, K. A. J. M.; Li, J.; Hawkins, G. D.; Thompson, J. D.; Kelly, C. P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *ZINDO-MN, version 1.2*; Quantum Theory Project, University of Florida, Gainesville, and Department of Chemistry, University of Minnesota, Minneapolis: Florida/Minnesota, 2005.

(54) Koslowski, A.; Beck, M. E.; Thiel, W. *J. Comput. Chem.* **2003**, *24*, 714–726.

(55) *TURBOMOLE V5.9.1 2007*, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989−2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com.

(56) Silva-Junior, M. R.; Sauer, S. P. A.; Schreiber, M.; Thiel, W. *Mol. Phys.* **2010**, *108*, 453–465.

(57) Schulten, K.; Karplus, M. *Chem. Phys. Lett.* **1972**, *14*, 305–309.

(58) Hudson, B. S.; Kohler, B. E. *Chem. Phys. Lett.* **1972**, *14*, 299–304.

(59) Head, J. D. *Int. J. Quantum Chem.* **2003**, *95*, 580–592.

(60) González-Luque, R.; Merchán, M.; Roos, B. O. *Z. Phys. D: At., Mol. Clusters* **1996**, *36*, 311–316.

(61) Miura, M.; Aoki, Y.; Champagne, B. *J. Chem. Phys.* **2007**, *127*, 084103/1–16.

(62) Hättig, C.; Weigend, F. *J. Chem. Phys.* **2000**, *113*, 5154–5161.

(63) Hättig, C.; Hald, K. *Phys. Chem. Chem. Phys.* **2002**, *4*, 2111–2118.

(64) Hättig, C.; Köhn, A.; Hald, K. *J. Chem. Phys.* **2002**, *116*, 5401–5410.

(65) Hättig, C.; Köhn, A. *J. Chem. Phys.* **2002**, *117*, 6939–6951.

(66) Voityuk, A. A.; Zerner, M. C.; Rösch, N. *J. Phys. Chem. A* **1999**, *103*, 4553–4559.

CT100030J

JCTC Journal of Chemical Theory and Computation

New Variational Method for the Ab Initio Study in Valence Coordinates of the Renner−Teller Effect in Tetra-Atomic Systems

Laurent Jutier* and Céline Léonard

Université Paris-Est, Laboratoire Modélisation et Simulation Multi Echelle, MSME UMR 8208 CNRS, 5 bd Descartes, 77454 Marne-la-Vallée, France

Received February 5, 2010

Abstract: A new variational methodology for the treatment of the Renner−Teller effect in tetra-atomic molecules has been developed in valence coordinates. The kinetic-energy operator of Bramley et al. [*Mol. Phys.* **1991**, *73*, 1183] for any sequentially bonded four-atom molecule, A−B−C−D, in the singlet nondegenerate electronic state has been adapted to the Renner−Teller and spin couplings by modifying the expression of the nuclear angular momentum. The total Schrödinger equation is solved by diagonalizing the Hamiltonian matrix in a three-step contraction scheme. The main advantage of this new theoretical development is the possibility of studying different isotopomers using the same potential-energy surfaces. This procedure has been tested on $HCCH^+$ and its deuterated derivatives $DCCD^+$ and $DCCH^+$. The calculated rovibronic band origins were compared with previous data deduced from the Jacobi coordinates methodology, dimensionality reduced variational treatment, and photoelectron spectra with a good global agreement. Rotational structures for these systems are also tackled.

1. Introduction

The Renner−Teller effect is a key issue for the study of many linear radicals and ions. Indeed, only the infrared and microwave spectra obtained from Σ electronic states can be understood without taking into account the couplings between the rovibrational degrees of freedom and the electronic orbital momentum.

In a previous paper,[1] we described a new variational method for the treatment of the Renner−Teller effect in tetra-atomic systems using Jacobi coordinates. We included couplings between all degrees of freedom intervening in the molecular Hamiltonian: rotation, vibration, electronic orbital, and electronic spin. In this coordinate set, the central stretch links the centers of mass of both diatomic fragments. It was successfully applied on the low-energy rovibronic states of the acetylene cation.[2] Nevertheless, three points motivated the development of a new numerical method using valence coordinates. (i) For well-bounded systems, the valence coordinates often reduce the crossing terms in the potential-energy surfaces (PESs) compared to Jacobi coordinates. In the case of the acetylene cation, for which both external atoms are hydrogens, the difference between both coordinate sets is not crucial. On the other hand, for a system such as thioketenyl (HCCS), the center of mass of the CS fragment is closer to S than to C. This simple fact involves very high crossing terms between both central and CS stretches as well as with the angle between them. The PESs are then much more difficult to fit with an analytical function, and the basis set should be very flexible in the subspaces associated with highly coupled coordinates. (ii) In Jacobi coordinates, the definition of the central stretch is different from an isotopomer to another. It is then necessary to define new PESs for each one. On the other hand, in valence coordinates, it is possible to use exactly the same analytical form of the PESs for different isotopomers, such as $HCCH^+$, $DCCH^+$, and $DCCD^+$. (iii) In the previous contraction scheme, only the stretching part of the Hamiltonian was diagonalized in a first step while both bending modes, the rotation, and the electronic orbital and spin angular momenta were treated together directly from the primitive basis set. The high number of required basis functions constrained us to

* To whom correspondence should be addressed: E-mail: jutier@univ-mlv.fr.

10.1021/ct100071d © 2010 American Chemical Society
Published on Web 04/20/2010

converge the eigenstates without the spin angular momentum and to study its effects with less basis functions for the bending degrees of freedom. This procedure is closer to a perturbative scheme than to a variational one and was only valid because of the low value of the spin−orbit constant (about −30 cm^{-1}).

To achieve our goal, we start from the work of Bramley et al.,[3,4] who derived an exact form of the nuclear Hamiltonian for tetra-atomic systems. This form can be used for the study of the rovibrational spectra from Σ electronic states, for which its dependence with the total angular momentum and its projections on the body-fixed (BF) axes are not differentiated from the nuclear contributions. For Renner−Teller systems, it must then be modified by subtracting the effects of the electronic orbital and spin angular momenta, following the same principle used for triatomic systems.[5,6] Moreover, in the case of a nonzero value of the electronic orbital momentum, the electronic part of the Hamiltonian depends on the choice of the BF axes because of the form of nonadiabatic couplings terms (NACTs) due to the torsion.

The present methodology has been tested on the $X^2\Pi$ electronic ground states of $HCCH^+$ and its deuterated isotopomers $DCCH^+$ and $DCCD^+$. $HCCH^+$ has the advantage of being well-studied experimentally and theoretically. It represents hence a benchmark system to validate our previous theoretical treatment based on the use of Jacobi coordinates.[1,2] Indeed, recent photoelectron spectra have been recorded by Tang et al.[7] and Yang and Mo[8] with a high resolution of the rovibronic bands of $HCCH^+$. The $DCCH^+$ and $DCCD^+$ photoelectron spectra of Reutt et al. allowed us to obtain some information about the vibrational modes of the deuterated species.[9]

Using their reduced degrees of freedom variational treatment, Perić et al. studied also the Renner−Teller effect and the spin−orbit coupling in the electronic ground state $^2\Pi$ of $DCCH^+$[10] and $^2\Pi_u$ of the symmetric species $HCCH^+$, $DCCD^+$.[11,12] Their approach was based on a harmonic representation of the potential-energy surfaces using the Renner−Teller parameters. The nuclear kinetic-energy operator considers derivatives with respect to four polar coordinates which describe the trans and cis bending vibrations, torsion and rotation around the axis associated with the smallest moment of inertia. These works were the first theoretical treatment of the Renner−Teller effect in symmetric and asymmetric linear tetra-atomic molecules.

The importance of modeling the spectroscopy of isotopomer species with a minimum computational effort must be emphasized, since deuterated istopomers of $HCCH^+$ are present in interstellar medium in which the chemistry is dominated by reactions between neutral and ionic molecules.[13] Moreover, the fraction H/D in comets water indicates that this water could be synthesized in interstellar medium.[14]

This paper is structured as follows. The complete definition of the Bramley et al. molecular Hamiltonian is explained in section 2. The modifications of the nuclear kinetic-energy operator and the total Hamiltonian related to the treatment of the Renner−Teller and spin−orbit couplings are given in section 3 as well as the contraction scheme and the primitive basis sets for each degree of freedom. The present methodology was checked by comparing rovibrational energies of the $X^3\Sigma_g^-$ electronic ground state of $HCCH^{2+}$ obtained from the present code, without taking into account the electronic angular momenta (i.e., $\Lambda = 0$, $S = 0$), and a variational code based on work by Bramley et al. In the last section, we validate the code and the associated methodology by comparison of our final rovibronic states of $HCCH^+$ and its isotopomers with already existent results, given by theory[2,11,12] and experiments.[7−9] Information concerning the rotational structures of both deuterated isotopomers is predictive.

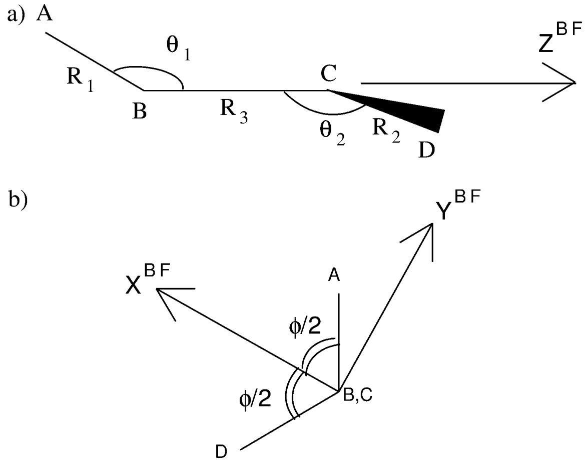


Figure 1. Definition of the valence coordinates and of the body-fixed (BF) frame.

2. Rovibrational Energies of a Nondegenerate Electronic State

This section summarizes the work done by Bramley et al.[3,4] that we used as the starting point for our new methodology. These authors determined an efficient variational method for calculating the rovibrational eigenstates of any sequentially bonded four-atom molecule, A−B−C−D. This method is suitable only for a singlet nondegenerate electronic state, since the nuclear kinetic-energy operator is a function of $\hat{\mathbf{J}}$, which refers to the total rovibrational angular momentum, i.e., neither orbital ($\hat{\mathbf{L}}$) nor spin ($\hat{\mathbf{S}}$) electronic angular momenta are considered, and the spin−orbit effects have not been introduced.

2.1. Rovibrational Hamiltonian. In the Born−Oppenheimer approximation, the molecular Hamiltonian can be decomposed as:

$$\hat{H}_{VR} = \hat{T}_{VR} + \hat{V}_N \tag{1}$$

The exact kinetic-energy operator $\hat{T}_{VR}$ is expressed in internal valence coordinates defined as follows. R_1, R_2, and R_3 are, respectively, the A−B, C−D, and B−C internuclear distances (Figure 1a). θ_1 and θ_2 are, respectively, the $\widehat{ABC}$ and $\widehat{BCD}$ angles (Figure 1a). Both vary in the interval $[0:\pi]$. ϕ is the dihedral angle between the planes defined by A, B, C and B, C, D (Figure 1b). This angle is in the $[0:2\pi]$ range. The origin O of the body-fixed frame is the center of mass of the whole system. The Z^{BF} axis is defined by the direction of the $\overrightarrow{BC}$ vector (Figure 1a). The B−C bond does not necessarily coincide with the Z^{BF} axis because of the center of mass position. The X^{BF} axis is perpendicular to Z^{BF} and a bisector of the torsional dihedral angle ϕ (Figure 1b). The Y^{BF} axis is such that $(O, X^{\mathrm{BF}}, Y^{\mathrm{BF}}, Z^{\mathrm{BF}})$ is a direct orthogonal frame (Figure 1b). The Euler angles (α, β, γ) link the space-fixed frame to the body-fixed frame defined above and are associated with the whole-molecule rotation angles. Derivatives with respect to these angles in $\hat{T}_{\mathrm{VR}}$ are replaced by $\hat{J}_x, \hat{J}_y, \hat{J}_z$, where $\hat{\mathbf{J}}$ represents the rovibrational angular momentum about body-fixed axes through the nuclear center of mass, obeying anomalous commutation relations.[15,16]

The translation motion, which can be exactly decoupled from all other degrees of freedom, was removed from $\hat{T}_{\mathrm{VR}}$ by placing the origin of the body-fixed frame at the molecular center of mass.

The use of valence internal coordinates is often very appropriate for a polynomial expansion form of $\hat{V}_N$ in the case of bonded systems, except for the torsion which must verify a modulo 2π periodicity. V_N is then modeled by a compact and factorizable expansion:

$$V_N = \sum_i C_i(R_1 - R_{1_{eq}})^{n_{1i}}(R_2 - R_{2_{eq}})^{n_{2i}}(R_3 - R_{3_{eq}})^{n_{3i}} \times (\theta_1 - \pi)^{n_{4i}}(\theta_2 - \pi)^{n_{5i}} \cos(n_{6_i}\phi) \quad (2)$$

In these coordinates $\hat{T}_{VR}$ is also separable and factorizable. The complete form can be found in the annexes of refs 3 and 4.

2.2. Primitive Basis Functions. The complexity of $\hat{T}_{VR}$ is compensated by the cheap evaluation of the Hamiltonian matrix elements calculated from products of one-dimensional integrals.

Before any contraction achieved by the resolution of the Schrödinger equation in a given subspace, the primitive basis functions can be expressed in the general form

$$\Phi_{\mathrm{rovib}} = \Phi_i^{str3D}(R_1, R_2, R_3) \cdot P_{l_1}^{m_1}(\theta_1) \cdot P_{l_2}^{m_2}(\theta_2) \cdot \mathrm{e}^{i\omega\phi} \cdot |J, K, M\rangle \quad (3)$$

where $|J,K,M\rangle$ is a rotational symmetric top function in which K is the eigenvalue of $\hat{J}_z$ in the body-fixed frame (BF) and M in the space-fixed frame (SF). In the absence of external electric fields, the quantum number M can be dropped.

The rules

$$\left.\begin{array}{lll} K(\mathrm{odd}) & \Rightarrow & \omega = (2m + 1)/2 \\ K(\mathrm{even}) & \Rightarrow & \omega = m \end{array}\right\} m = 0, 1, 2, ... \quad (4)$$

impose that Φ_{rovib} is single valued for ϕ and γ moving in the range $0 \rightarrow 2\pi$.

$P_{l_i}^{m_i}$ functions remove singularities of $\hat{T}_{VR}$ if

$$\begin{array}{ll} \text{for } m_1 = 0, & \omega = \dfrac{K}{2} \\ \text{for } m_2 = 0, & \omega = -\dfrac{K}{2} \end{array} \quad (5)$$

Basis functions that violate these rules have infinite expectation values across $\hat{T}_{VR}$.

The stretch basis set, $\Phi^{\mathrm{str3D}}(R_1, R_2, R_3)$, is made by the 1-dimensional products of harmonic or Morse oscillators of q_1, q_2, R_3 coordinates, $\Phi_{v_1}(q_1)\Phi_{v_2}(q_2)\Phi_{v_3}(R_3)$. q_1, q_2 are the symmetrized combinations of R_1, R_2 in the case of the $D_{\infty h}(M)$ system or equal to R_1, R_2 otherwise.

3. Rovibronic Energies of a Degenerate Electronic State

For linear degenerate electronic states, the couplings between orbital, spin electronic, and rovibrational angular momenta should be taken into account.

3.1. Rovibronic Hamiltonian. The Hamiltonian that considers the Renner−Teller and spin−orbit effects is

$$\hat{H} = \hat{T}_N + \hat{H}_{\mathrm{e}} + \hat{H}_{\mathrm{SO}} \quad (6)$$

where $\hat{H}_{\mathrm{e}}$ is the electronic Hamiltonian and $\hat{H}_{\mathrm{SO}}$ is the perturbative spin−orbit contribution. $\hat{T}_N$ comes from the adapted kinetic-energy operator $\hat{T}_{VR}$ in the valence coordinates of Bramley et al.[3,4] described in the previous section to the Renner−Teller and spin−orbit treatment.

In molecules having a nonzero electronic angular momentum, either orbital ($\hat{\mathbf{L}}$) or spin ($\hat{\mathbf{S}}$) or both, $\hat{\mathbf{J}}$ refers to the total rovibronic angular momentum and the $\hat{\mathbf{J}}$ introduced in section 2.1 must be replaced by $\hat{\mathbf{J}}-\hat{\mathbf{L}}-\hat{\mathbf{S}}$ in $\hat{T}_N$. Indeed, the kinetic-energy operator depends only on nuclear coordinates. The projections of the total angular momentum on the BF axes $\hat{J}_{x,y,z}$ are replaced by $\hat{J}_{x,y,z}-\hat{S}_{x,y,z}-\hat{L}_{x,y,z}$, $\hat{S}_{x,y,z}$ and $\hat{L}_{x,y,z}$ being the projections of the electron spin and orbital angular momenta on the BF axes, respectively. Focusing on a single degenerate electronic state well isolated from all other ones, it is possible to neglect the effects of both $\hat{L}_x$ and $\hat{L}_y$ projections of the electronic orbital momentum on the X^{BF} and Y^{BF} axes. $\hat{J}_{x,y,z}$ obeys anomalous commutation rules,[16−18] whereas $\hat{L}_{x,y,z}$ and $\hat{S}_{x,y,z}$ obey normal ones:

normal commutation rules:

$$\begin{aligned} \hat{S}_x|S,\Sigma\rangle &= \frac{1}{2}\sqrt{S(S+1) - \Sigma(\Sigma+1)}|S,\Sigma+1\rangle \\ &\quad + \frac{1}{2}\sqrt{S(S+1) - \Sigma(\Sigma-1)}|S,\Sigma-1\rangle \\ \hat{S}_y|S,\Sigma\rangle &= \frac{-i}{2}\sqrt{S(S+1) - \Sigma(\Sigma+1)}|S,\Sigma+1\rangle \\ &\quad + \frac{i}{2}\sqrt{S(S+1) - \Sigma(\Sigma-1)}|S,\Sigma-1\rangle \end{aligned} \quad (7a)$$

anomalous commutation rules:

$$\begin{aligned} \hat{J}_x|J,P\rangle &= \frac{-i}{2}\sqrt{J(J+1) - P(P+1)}|J,P+1\rangle \\ &\quad + \frac{i}{2}\sqrt{J(J+1) - P(P-1)}|J,P-1\rangle \\ \hat{J}_y|J,P\rangle &= \frac{1}{2}\sqrt{J(J+1) - P(P+1)}|J,P+1\rangle \\ &\quad + \frac{1}{2}\sqrt{J(J+1) - P(P-1)}|J,P-1\rangle \end{aligned} \quad (7b)$$

where $|S,\Sigma\rangle$ and $|J,P\rangle$ are the eigenstates of $\hat{S}_z$, $\hat{S}^2$ and $\hat{J}_z$, $\hat{J}^2$, respectively, such as the quantum numbers Σ and P are defined by $\hat{S}_z|S,\Sigma\rangle = \Sigma|S,\Sigma\rangle$ and $\hat{J}_z|J,P\rangle = P|J,P\rangle$ with

$$P = K + \Lambda + \Sigma \quad (8)$$

where K and Λ correspond to the quantum numbers associated with the projections on Z^{BF} of the rovibrational and electronic angular momenta, $\hat{\mathbf{J}}_{\mathbf{N}} = \hat{\mathbf{J}}-\hat{\mathbf{L}}-\hat{\mathbf{S}}$ and $\hat{\mathbf{L}}$, respectively.

3.2. Electronic States. If, for bent geometries, each component of a degenerate electronic state correlates to a Π or Δ, Φ... electronic state at linearity, the eigenvectors $|\pm\Lambda\rangle$ of the electronic orbital angular momentum, $\hat{L}_z$, can still be used as a basis set for the electronic orbital part of the total rovibronic wave function. Then $\hat{L}_z|+\Lambda\rangle = +\Lambda|+\Lambda\rangle$ and $\hat{L}_z|-\Lambda\rangle = -\Lambda|-\Lambda\rangle$, where Λ is the absolute value of the projection of the electronic orbital momentum on Z^{BF}.

On the other hand, operators $\hat{L}_{x,y}$ only couple electronic states associated with values of Λ differing of ± 1, for instance, a Σ electronic state close in energy to a Π electronic state. The action of these operators has been neglected.

The basis functions for the electronic orbital degree of freedom are then

$$\Phi_{\mathrm{e}}^{\pm} = |\pm\Lambda\rangle \approx \mathrm{e}^{\pm i\Lambda(\theta_{\mathrm{e}}^{\mathrm{BF}})} = \frac{X \pm iY}{\sqrt{2}} \tag{9}$$

X and Y correspond to the real electronic components of the considered degenerate electronic state.

$\hat{L}_z$ is defined as $-i(\partial)/(\partial\theta_{\mathrm{e}}^{\mathrm{BF}})$, where $\theta_{\mathrm{e}}^{\mathrm{BF}}$ is the collective electronic orbital angle. If the electronic state results from a configuration with only one electron or one vacuum in a degenerate molecular orbital, $\theta_{\mathrm{e}}^{\mathrm{BF}}$ is associated with this only unpaired electron or vacuum in the one-electron approximation. In the body-fixed frame defined in section 2.1, the electronic Hamiltonian, $\hat{H}_{\mathrm{e}}$, expressed in the basis set $\{\Phi_{\mathrm{e}}^{+}, \Phi_{\mathrm{e}}^{-}\}$ is

$$\tilde{H}_e = \begin{pmatrix} \dfrac{V^X + V^Y}{2} & \dfrac{V^X - V^Y}{2} \\ \dfrac{V^X - V^Y}{2} & \dfrac{V^X + V^Y}{2} \end{pmatrix} \tag{10}$$

V^X and V^Y are the potential-energy surfaces associated with X and Y electronic states. It must be noted that the electronic Hamiltonian matrix $\tilde{H}_{\mathrm{e}}$ differs from the one described in ref 1 due to a theoretical development in a body-fixed frame defined such that reference plane ($X^{\mathrm{BF}}OZ^{\mathrm{BF}}$) is the bisector of the dihedral angle ϕ (see Figure 1) instead of E_2 in ref 1. For the same reason, the nonadiabatic electronic coupling terms characteristic of the Renner−Teller effect are simply

$$\left\langle Y\middle|\frac{\partial}{\partial\gamma}\middle|X\right\rangle \simeq \Lambda; \quad \left\langle Y\middle|\frac{\partial}{\partial\phi}\middle|X\right\rangle \simeq 0 \tag{11}$$

Both frameworks are equivalent.

When the molecule is linear, both electronic components are degenerate and the potential-energy surface can be described by the following analytical function

$$V_{\mathrm{stretch}} = \sum_{ijk} A_{ijk}(R_1 - R_{1,eq})^i(R_2 - R_{2,eq})^j(R_3 - R_{3,eq})^k \tag{12}$$

where $R_{i,\,eq}$ is the reference for the R_i stretch and $A_{ijk} = A_{jik}$ for symmetry reasons.

When the molecule is no longer linear, it is necessary to express two PESs. We directly fitted $(V^X + V^Y)/2 = V^{\mathrm{average}}$ and $(V^X - V^Y)/2 = V^{\mathrm{diff}}$ rather than V^X and V^Y. For pure bending displacements:

$$\begin{aligned} V_{\mathrm{bend}}^{\mathrm{average}} &= \sum_{lmn} B_{lmn}\cdot\theta_1^l\cdot\theta_2^m\cdot\cos(n\phi^{\mathrm{BF}}) \\ V_{\mathrm{bend}}^{\mathrm{diff}} &= \sum_{lmn} C_{lmn}\cdot\theta_1^l\cdot\theta_2^m\cdot\cos(n\phi^{\mathrm{BF}}) \end{aligned} \tag{13}$$

For couplings between bending and stretching displacements:

$$\begin{aligned} V_{\mathrm{sb}}^{\mathrm{average}} &= \sum_{ijklmn} D_{ijklmn}(R_1 - R_{1,eq})^i(R_2 - R_{2,eq})^j \times \\ &\quad (R_3 - R_{3,eq})^k\cdot\theta_1^l\cdot\theta_2^m\cdot\cos(n\phi^{BF}) \\ V_{\mathrm{sb}}^{\mathrm{diff}} &= \sum_{ijklmn} E_{ijklmn}(R_1 - R_{1,eq})^i(R_2 - R_{2,eq})^j \times \\ &\quad (R_3 - R_{3,eq})^k\theta_1^l\theta_2^m\cdot\cos(n\phi^{BF}) \end{aligned} \tag{14}$$

These expressions of the analytical PESs are similar to the ones used in our previous study.[1]

The potential energies were evaluated for 204 independent geometries in the C_S symmetry point group, i.e., $\phi^{\mathrm{BF}} = 0$ (cis conformation) or $\phi^{\mathrm{BF}} = \pi$ (trans conformation). X, Y are correlated to the A, B irreducible representations in the C_2 point group, respectively. The RCCSD(T) method[19−21] and the cc-pV5Z basis set[22] were used within the MOLPRO package.[23] The global rms is less than 0.5 cm^{-1}, with 35 independent coefficients for the average surface, V^{average} and 18 for V^{diff} (which does not contain pure stretching coefficients). All coefficients of the analytical representations of the PESs are given.

Table 1 gives the coefficients for the pure stretching part of the average surface A_{ijk}.

Table 2 gives the coefficients for the pure bending part of the average surface B_{ijk}.

In Table 3 the coefficients for the couplings between bending and stretching modes of the average surface D_{ijklmn} are listed.

Table 1. Coefficients of the Analytic Representation of the Degenerate PES at Linearity, V_{stretch} of Eq 12 (in au)[a]

A_i	n_1	n_2	n_3
−76.797258052	0	0	0
0.18775932871	2/0	0/2	0/0
0.00049219849978	1	1	0
−0.010446131479	1/0	0/1	1/1
0.44482987202	0	0	2
−0.18875621850	3/0	0/3	0/0
0.0034470302368	2/0	0/2	1/1
−0.00046558758339	1/0	0/1	2/2
−0.45136275104	0	0	3
0.12843527244	4/0	0/4	0/0
−0.0024256691021	3/0	0/3	1/1
−0.0042693423415	2/0	0/2	2/2
0.29119003380	0	0	4
−0.15477459177	0	0	5
0.058833359015	0	0	6
−0.080493877749	5/0	0/5	0/0
−0.0022746306025	3/2	2/3	0/0
0.041225800651	6/0	0/6	0/0

[a] When $n_1 \neq n_2$, the A_i coefficients are common to triplets in which n_1 and n_2 are inverted. The zeroth-order coefficient, involving only a global shift for the wavefunction energies, is not considered in variational calculations. $R_{1,\mathrm{eq}}$, $R_{2,\mathrm{eq}}$, and $R_{3,\mathrm{eq}}$ are defined such that the first-order coefficients cancel: $R_{1,\mathrm{eq}} = R_{2,\mathrm{eq}} = 2.0382$ bohr and $R_{3,\mathrm{eq}} = 2.3627$ bohr.

Table 2. Coefficients of the Analytic Representation of the Average PES, $V_{\mathrm{bend}}^{\mathrm{average}}$ of Eq 13 (in au)[a]

B_i	n_1	n_2	n_3
0.031380342315	2/0	0/2	0/0
−0.00010589082066	4/0	0/4	0/0
0.0075651209859	2	2	0
−0.00032900039217	6/0	0/6	0/0
0.015301967243	1	1	1
−0.0069697677988	3/1	1/3	1/1
0.0052863195870	3	3	1

[a] When $n_4 \neq n_5$, the B_i coefficients are common to triplets in which n_4 and n_5 are inverted.

Table 3. Coefficients of the Analytic Representation of the Average PES, $V_{\mathrm{sb}}^{\mathrm{average}}$ of Eq 14 (in au)[a]

D_i	n_1	n_2	n_3	n_4	n_5	n_6
−0.0085356746578	1/0	0/1	0/0	2/0	0/2	0/0
−0.0016760742892	1/0	0/1	0/0	1/1	1/1	1/1
−0.0017279408242	0/1	1/0	0/0	2/0	0/2	0/0
−0.036304901251	0/0	0/0	1/1	2/0	0/2	0/0
0.028175216677	0	0	1	1	1	1
−0.0015969509525	2/0	0/2	0/0	2/0	0/2	0/0
−0.0014800493600	2/0	0/2	0/0	1/1	1/1	1/1
−0.0082527829467	0/0	0/0	2/2	2/0	0/2	0/0
0.0050003688608	0	0	2	1	1	1
0.011960852836	1/0	0/1	1/1	2/0	0/2	0/0

[a] When $(n_1, n_4) \neq (n_2, n_5)$, the D_i coefficients are common to sextuplets in which (n_1, n_4) and (n_2, n_5) are inverted.

Table 4. Coefficients of the Analytic Representation of the Difference PES, $V_{\mathrm{bend}}^{\mathrm{diff}}$ of Eq 13 (in au)[a]

C_i	n_1	n_2	n_3
−0.0040713277155	2	2	0
−0.023659941406	4/2	2/4	0/0
−0.00024053133094	5/1	1/5	1/1
0.0062884157463	2/0	0/2	1/1
−0.0040251186102	4/0	0/4	1/1
0.0075451909845	2	2	1
−0.00037799117499	6/0	0/6	1/1
−0.015284625238	1	1	0
0.055519559631	3	3	0

[a] When $n_4 \neq n_5$, the C_i coefficients are common to triplets in which n_4 and n_5 are inverted.

Table 5. Coefficients of the Analytic Representation of the Difference PES, $V_{\mathrm{sb}}^{\mathrm{diff}}$ of Eq 14 (in au)[a]

E_i	n_1	n_2	n_3	n_4	n_5	n_6
−0.0016889980478	1/0	0/1	0/0	2/0	0/2	1/1
−0.00085785320357	1/0	0/1	1/1	1/1	1/1	0/0
0.00020147090259	0/1	1/0	0/0	2/0	0/2	1/1
0.029003257627	0/0	0/0	1/1	2/0	0/2	1/1
−0.0058076857757	0	0	1	1	1	0
−0.0014955409812	2/0	0/2	0/0	2/0	0/2	1/1
0.00059036218845	0/2	2/0	0/0	2/0	0/2	1/1
−0.013421739988	0/0	0/0	2/2	2/0	0/2	1/1
−0.0015540891813	1/0	0/1	1/1	2/0	0/2	1/1

[a] When $(n_1, n_4) \neq (n_2, n_5)$, the E_i coefficients are common to sextuplets in which (n_1, n_4) and (n_2, n_5) are inverted.

Table 4 lists the coefficients for the pure bending degrees of freedom part of the difference surface C_{lmn}.

Table 5 lists the coefficients for the couplings between bending and stretching degrees of freedom of the difference surface E_{ijklmn}.

The polynomial expansion is used in the area of the configuration space in which ab initio points were computed, corresponding to the energy range of our present study. This range must at least correspond to the sum of the global zero-point energy (ZPE), the maximum of the excitation energy, and a margin due to a tunneling effect. However, the only way of verifying the energy range is to plot the final wave functions, which have to be well localized in the range of ab initio points. We chose the range [π:1.92], appearing (more than) sufficient regarding the shapes of stationary states up to ≃1800 cm^{-1} from the ZPE.

Figures 2 and 3 show that the average PES has a standard shape. The harmonic terms are predominant for θ_1 or θ_2 less than 60° from linearity. Non-negligible crossing terms are present between the bending and the central stretching modes in the trans conformation (see Figure 3).

The behavior of the difference PES is less intuitive and differs strongly from cis to trans conformations, as shown in Figure 4. In the trans conformation, both electronic components, V^X and V^Y, are well separated. The A' one (correlates to B in the C_2 symmetry point group) lies higher in energy than the A'' one (correlates to A in C_2). In the cis conformation, V^X and V^Y are almost degenerate, especially if $\theta_1 = \theta_2$, as already noted in our previous papers.[1,2]

The PES shapes are slightly different from Jacobi to valence coordinates, but the same conclusions remain. The definition of Renner−Teller parameters is not recommended since the perturbative approach, derived from the harmonic approximation, is unable to describe the near degeneracy of both electronic components in the cis conformation. As in previous works,[1,2] the conical intersection is reproduced by connecting parts of the PESs in the cis and trans conformations associated with the same irreducible representation A or B in the C_2 symmetry point group. Explicit ab initio computations of electronic energies in the C_1 symmetry point group would require a diabatization process, which is here achieved by smoothing the curves at the vicinity of conical intersections.

3.3. Basis Functions and Contraction Scheme. The primitive basis set is composed by products of one-dimensional functions, one by degree of freedom in the assumption that a factorizable and partially separable Hamiltonian is used:

$$\Phi_{\mathrm{prim}}^{\mathrm{rovib}} = \Phi_{v_1}(R_1)\Phi_{v_2}(R_2)\Phi_{v_3}(R_3)\cdot P_{l_1}^{m_1}(\theta_1)\cdot P_{l_2}^{m_2}(\theta_2)\cdot \mathrm{e}^{i\omega\phi}\cdot |J,P\rangle\cdot \Phi_{\mathrm{e}}^{\pm}\cdot |S,\Sigma\rangle \qquad (15)$$

The diagonalization of the complete molecular Hamiltonian (eq 6) directly from this primitive basis set is by far too expensive in terms of memory and CPU time. A successive step contraction scheme is then settled starting with the diagonalization of subspaces using parts of $\hat{H}$. In the present study of the acetylene cation and its isotopomers, the following procedure was achieved. (1) $\Psi_{v_i}^{\mathrm{str3D}}(R_1, R_2, R_3)$ contracted functions are first optimized in the 3-dimensional space associated with the stretches, $\{R_1, R_2, R_3\}$. (2) For *each* stretching contraction $\Psi_{v_i}^{\mathrm{str3D}}(R_1, R_2, R_3)$, vibronic origins of bands are obtained

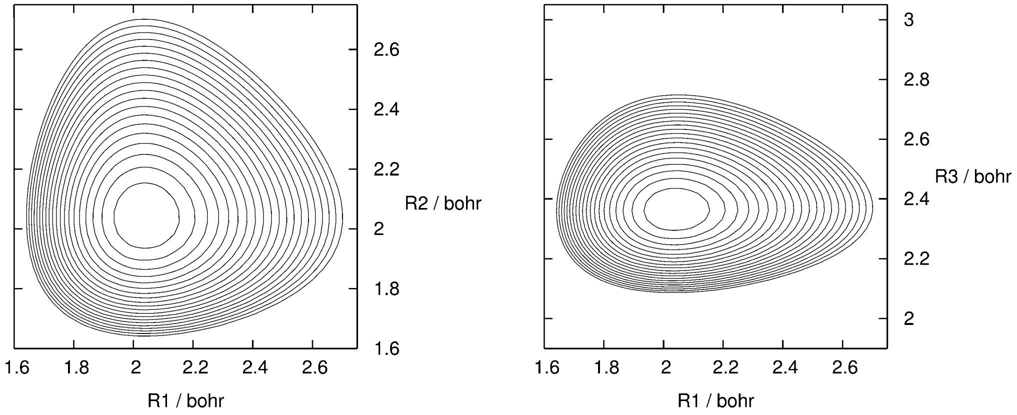


Figure 2. Two-dimensional contour plots of the average potential at linearity. The nonvarying bond length is fixed at its equilibrium geometry ($R_3 = 2.3627$ bohr and $R_2 = 2.0382$ bohr). The isolevel spacing is 500 cm^{-1}.

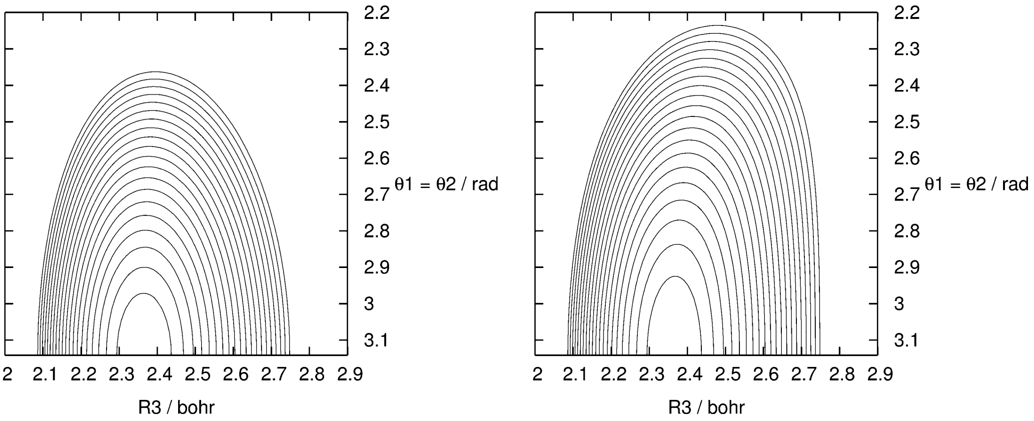


Figure 3. Two-dimensional contour plots of the average potential for R_1 and R_2 fixed at their equilibrium values. On the left-hand side, the torsion is fixed at 0 (cis conformation). On the right-hand side, the torsion is fixed at π (trans conformation). The isolevel spacing is 500 cm^{-1}.

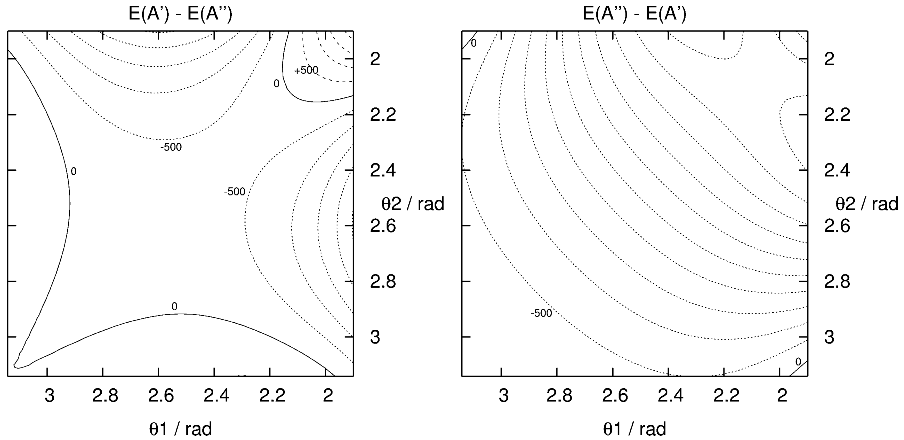


Figure 4. Two-dimensional contour plots of the difference potential for R_1, R_2, and R_3 fixed at their equilibrium values. On the left-hand side, the torsion is fixed at 0 (cis conformation). On the right-hand side, the torsion is fixed at π (trans conformation). Angles are in radians. The isolevel spacing is 500 cm^{-1}.

considering the four angles space, $\{\theta_1, \theta_2, \phi, \gamma\}$: $\Phi^{\rm vib}_{vi,bi,P,\Lambda,\Sigma} = \Psi^{\rm str3D}_{v_i}(R_1, R_2, R_3)\cdot\Psi^{\rm bend3D}_{b_i}(\theta_1, \theta_2, \phi)\cdot e^{i(P-\Lambda-\Sigma)\gamma}$. For these functions, the P and Σ quantum numbers are considered as good quantum numbers. Then each triplet (v_i, P, Σ) corresponds to independent calculations. (3) All previous contractions are collected together and coupled by the complete molecular Hamiltonian $\hat{H}$. The only good quantum number is J, associated with the total angular momentum.

While the only observable is the complete Hamiltonian $\hat{H}$, the physical meaning of subspaces as well as convergence criteria is not trivial. Different contraction schemes can be convenient depending on the stronger coupling terms in the PESs (see discussion in ref 4).

The following parts detail the different steps of this contraction scheme.

3.3.1. Stretching States. The first step of the contraction scheme is the diagonalization of the pure stretching Hamiltonian $\hat{H}_{\text{stretch}}$. The considered molecular geometries are linear ($\theta_1 = \theta_2 = \pi$ and ϕ is not defined). The reduced Hamiltonian is

$$\hat{H}_{\text{stretch}} = -\frac{1}{2}\left[\frac{1}{\mu_1}\frac{\partial^2}{\partial R_1^2} + \frac{1}{\mu_2}\frac{\partial^2}{\partial R_2^2} + \frac{1}{\mu_3}\frac{\partial^2}{\partial R_3^2}\right] \tag{16a}$$

$$+ \cos(\theta_1)\left[\frac{-1}{m_B R_1 R_3} - \frac{1}{m_B}\frac{\partial^2}{\partial R_1 \partial R_3}\right] \tag{16b}$$

$$+ \cos(\theta_2)\left[\frac{-1}{m_C R_2 R_3} - \frac{1}{m_C}\frac{\partial^2}{\partial R_2 \partial R_3}\right] \tag{16c}$$

$$+ V_{\text{stretch}} \tag{16d}$$

with

$$\begin{aligned}\mu_1 &= m_A m_B/(m_A + m_B)\\ \mu_2 &= m_C m_D/(m_C + m_D)\\ \mu_3 &= m_B m_C/(m_B + m_C)\end{aligned} \tag{17}$$

The relation (eq 16a) is very similar to the stretching part of the nuclear kinetic-energy operator in Jacobi coordinates[1] but with a different definition of R_3. In valence coordinates, R_3 is the B−C internuclear bond length while R_3 is the distance between both centers of mass of diatomic fragments A−B and C−D in Jacobi coordinates. Thus, the terms in eqs 16b and 16c are absent in $\hat{T}_N$ expressed in Jacobi coordinates. They are not simple couplings between the different valence stretches since they involve additionally the bending angles θ_1 and θ_2 by way of their cosines. For linear configurations, these cosines are fixed at −1, and as a consequence, the pure stretching vibrational energies become slightly overestimated by $\lesssim 2\%$. This issue will be raised in the following steps of the contraction scheme.

For the variational calculations, basis functions are products of three eigenfunctions of the harmonic oscillator

$$\Phi^{\text{str3D}}_{v_1,v_2,v_3}(R_1, R_2, R_3) = \Phi^{\alpha_1}_{v_1}(R_1 - R_{1,\text{eq}}) \cdot \Phi^{\alpha_2}_{v_2}(R_2 - R_{2,\text{eq}}) \cdot \Phi^{\alpha_3}_{v_3}(R_3 - R_{3,\text{eq}}) \tag{18}$$

where

$$\Phi^{\alpha}_{v}(X) = \frac{1}{\sqrt{2^v v!}}\left(\frac{\alpha}{\pi}\right)^{1/4} H_v(\sqrt{\alpha}X)\mathrm{e}^{-\alpha X^2/2} \tag{19}$$

H_v is the Hermite polynomial of order v, and α is linked to the force constant and to the reduced mass associated to a given one-dimensional binding potential

$$\alpha_i = \sqrt{k_i \mu_i} \text{ with } k_1 = k_2 = 3.775 \text{ au and } k_3 = 0.8897 \text{ au} \tag{20}$$

Only the values of the reduced masses are modified when different isotopomers are considered.

The diagonalization of $\hat{H}_{\text{stretch}}$ gives optimized three-dimensional stretching contractions

$$\Psi^{\text{str3D}}_{i}(R_1, R_2, R_3) = \sum_j C^i_j \Phi^{\text{str3D}}_{v_{1_j},v_{2_j},v_{3_j}}(R_1, R_2, R_3) \tag{21}$$

The convergence of the stretching vibrational energies is reached for HCCH$^+$, DCCH$^+$, and DCCD$^+$ with $v_{1_{\max}} = v_{2_{\max}} = 10$ and $v_{3_{\max}} = 15$ leading to 559 basis functions $\Phi^{\text{str3D}}_{v_1,v_2,v_3}$, the grid of ($v_1$, v_2, v_3) values being nonrectangular. The convergence of vibrational states associated with the central stretch R_3 requires more basis functions than the external stretches due to stronger anharmonicity.

3.3.2. Vibronic States. In the second step of the contraction scheme, all vibrational terms of the kinetic-energy operator $\hat{T}_N$ are used, including those depending on $\hat{J}_z$ after the following transformation

$$\hat{J}_z \rightarrow \hat{J}_z - \hat{L}_z - \hat{S}_z \tag{22}$$

in order to include the Renner−Teller and spin couplings. No coordinate is kept fixed in $\hat{T}_N$. Moreover, the spin−orbit coupling is introduced perturbatively at this step by the way of $\hat{H}_{\text{SO}}$. As mentioned in ref 1, the spin−orbit operator can be approximated as

$$\hat{H}_{\text{SO}} = A_{\text{SO}}\hat{L}_z \cdot \hat{S}_z \tag{23}$$

The spin−orbit constant A_{SO} is taken to be equal to −30.23 cm^{-1}.[1]

The variational basis functions are products of one-dimensional functions

$$\Phi^{\text{vib}}_{v_i,b_i,P,\Lambda,\Sigma} = \Phi^{\text{vib}}_{v_i,b_i} \cdot \mathrm{e}^{i(P-\Lambda-\Sigma)\gamma} \tag{24}$$

where

$$\Phi^{\text{vib}}_{v_i,b_i} = \Psi^{\text{str3D}}_{v_i}(R_1, R_2, R_3) \cdot \Psi^{\text{bend3D}}_{b_i}(\theta_1, \theta_2, \phi) \tag{25}$$

$\Psi^{\text{str3D}}_{v_i}(R_1, R_2, R_3)$ is optimized at the previous stretching contraction step and

$$\Psi^{\text{bend3D}}_{b_i}(\theta_1, \theta_2, \phi) = P^{m_1}_{l_1}(\theta_1) \cdot P^{m_2}_{l_2}(\theta_2) \cdot \mathrm{e}^{i\omega\phi} \tag{26}$$

Finally, vibronic energies are obtained independently for each triplet (v_i, P, Σ).

As already emphasized by Bramley et al.,[3] the quantum numbers m_1, m_2, and ω cannot be independent in order to avoid the singularities at linearity. Indeed, the kinetic-energy operator $\hat{T}_N$ contains several terms involving $1/\sin^2\theta_1$ and $1/\sin^2\theta_2$. Since the Jacobian is proportional to $\sin\theta_1 \cdot \sin\theta_2$, after multiplication with $\hat{T}_N$, $1/\sin\theta_1$ and $1/\sin\theta_2$ terms remain and diverge for $\theta_{1,2} = 0$ or π. On the other hand, the Legrendre polynomials $P^m_l(\theta)$ contain a factor $\sin^{|m|}\theta$, and singularities occur only if $m = 0$ for the bra *and* the ket. Bramley et al. canceled this problem for nondegenerate singlet electronic states by the introduction of dependencies between the values of m_1, m_2, and ω (eqs 34 and 35 in ref 3 or eq 5 in the present work). The same relationships are used in this work to remove singularities by replacing $\hat{J}_z$ by $\hat{J}_z - \hat{L}_z - \hat{S}_z$. Then the divergences due to the following terms

$$\frac{1}{2\sin^2\theta_1}\left[\frac{1}{\mu_1 r_1^2}+\frac{1}{\mu_3 r_3^2}-\frac{2\cos\theta_1}{M_2 r_1 r_3}\right]\cdot \left[\frac{\partial^2}{\partial\phi^2}-i(\hat{J}_z-\hat{L}_z-\hat{S}_z)\frac{\partial}{\partial\phi}-\frac{(\hat{J}_z-\hat{L}_z-\hat{S}_z)^2}{4}\right]$$
$$\frac{1}{2\sin^2\theta_2}\left[\frac{1}{\mu_2 r_2^2}+\frac{1}{\mu_3 r_3^2}-\frac{2\cos\theta_2}{M_3 r_2 r_3}\right]\cdot \left[\frac{\partial^2}{\partial\phi^2}+i(\hat{J}_z-\hat{L}_z-\hat{S}_z)\frac{\partial}{\partial\phi}-\frac{(\hat{J}_z-\hat{L}_z-\hat{S}_z)^2}{4}\right] \quad (27)$$

in $\hat{T}_N$ are canceled by introducing the constraints

$$m_1 = 0 \Rightarrow \omega = \frac{P-\Lambda-\Sigma}{2}$$
$$m_2 = 0 \Rightarrow \omega = -\frac{P-\Lambda-\Sigma}{2} \quad (28)$$

In the previous Jacobi coordinates development,[1,2] the singularities were removed using spherical harmonics $Y_{l_i}^{m_i}(\theta_i, \phi_i)$, $i = 1, 2$. ϕ_1 and ϕ_2 are independent azimuthal angles describing the rotation of each diatomic fragment along Z^{E2} with the corresponding dependence, $e^{im_1\phi_1}\cdot e^{im_2\phi_2}$ and the restriction

$$m_1 + m_2 = P - \Lambda - \Sigma \quad (29)$$

based on the additive property of the projections along Z^{E2} of the angular momenta.

The present choice of the reference plane for the definition of BF allows linking ϕ_1 and ϕ_2 with ϕ and γ

$$\phi = \phi_2 - \phi_1 \text{ and } \gamma = (\phi_1 + \phi_2)/2 \quad (30)$$

Then the collective variation of the bending basis function with rotation angles around Z^{BF} can be decomposed in terms of ϕ_1 and ϕ_2

$$e^{i\omega\phi}\cdot e^{i(P-\Lambda-\Sigma)\gamma} = e^{i\phi_1(P-\Lambda-\Sigma/2-\omega)}\cdot e^{i\phi_2(P-\Lambda-\Sigma/2+\omega)} \quad (31)$$

The use of spherical harmonics for both diatomic fragments such as $Y_{l_1}^{m_1}(\theta_1, \phi_1)\cdot Y_{l_2}^{m_2}(\theta_2, \phi_2)$ is then equivalent to the use of $P_{l_1}^{m_1}(\theta_1)\cdot P_{l_2}^{m_2}(\theta_2)\cdot e^{i\omega\phi}\cdot e^{i(P-\Lambda-\Sigma)\gamma}$ if

$$m_1 = \left|\frac{P-\Lambda-\Sigma}{2} - \omega\right|$$
$$m_2 = \left|\frac{P-\Lambda-\Sigma}{2} + \omega\right| \quad (32)$$

These equations are consistent with eq 27. To reduce the number of integral computations, we follow the same strategies as Bramley and Handy:[4] odd values of m_1 and m_2 ≥ 3 are replaced by 1 and even values of m_1 and $m_2 \geq 4$ are replaced by 2.

The simultaneous treatment of all bending degrees of freedom makes difficult the convergence of the energies,

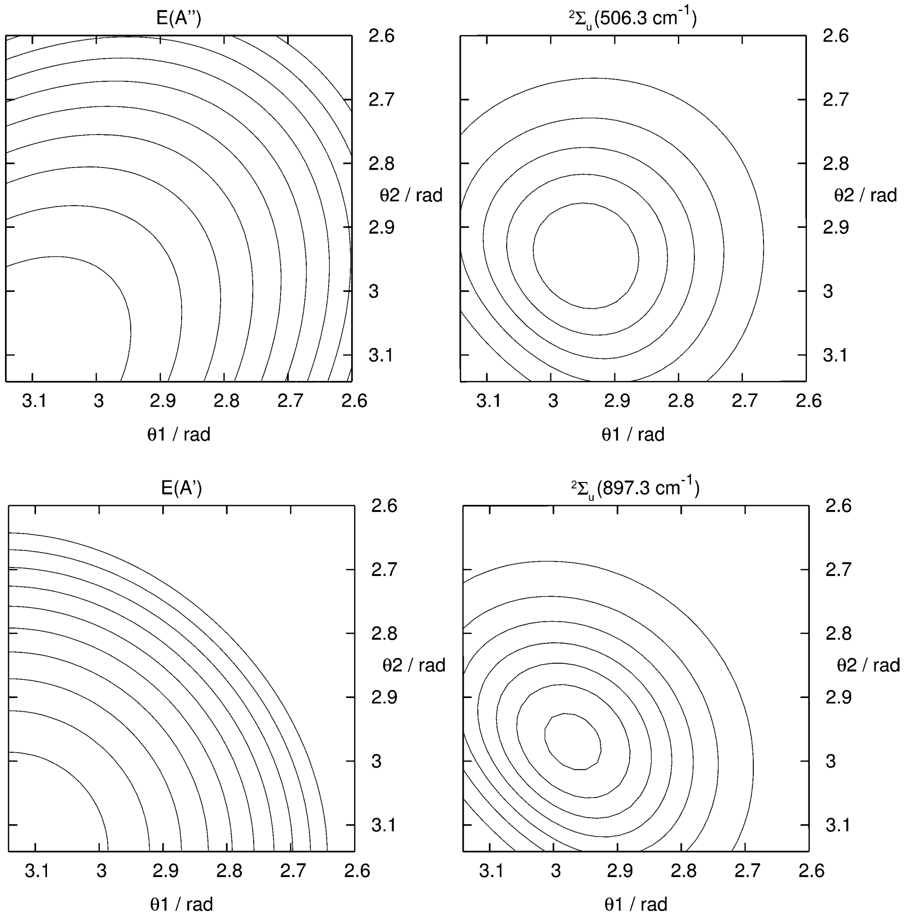


Figure 5. HCCH$^+$: Comparison between two-dimensional contour plots for the potential and both Σ states with one quantum in the trans bending mode. R_1, R_2, and R_3 are fixed at their equilibrium values, and $\phi = \pi$ (trans conformation) for all plots.

while it is not suitable to separate this ensemble of coordinates that are strongly coupled. In order to facilitate the convergence of the bending levels, factor $\exp[-10\cdot(\theta - \pi)^2]$ was introduced in the basis functions. Legende polynomials

$$P_l^m(\theta) = [\sum_{i=0}^{l-|m|} C_l^m \cos^i \theta] \cdot \sin^{|m|} \theta \tag{33}$$

are then replaced by

$$Q_l^m(\theta) = [e^{-10\cdot(\theta-\pi)^2}] \cdot [\sum_{i=0}^{l-|m|} {}^iD_l^m \cos^i \theta] \cdot \sin^{|m|} \theta \tag{34}$$

Coefficients ${}^iD_l^m$ have been determined by orthonormalization of basis functions for a given m, increasing l successively. As Legendre polynomials, these functions allow also solving the divergence problem at linearity but also are more concentrated around the linear geometries. Functions $Q_l^m(\theta)$ participate in a more general frame that is detailed in ref 24.

The basis set used in this work comprised 10 contractions for the stretches, $l_{1,\max} = l_{2,\max} = 15$ for the bending modes and $\omega_{\max} = 13/2$ for the torsion for all isotopomers. Cuts are introduced in energy at 10 000 cm^{-1} and in θ_1, θ_2 at 110°.

At this stage of our contraction scheme, the vibronic energies are overestimated for two reasons: (1) each vibronic contraction includes only one stretching function, $\Psi_{v_i}^{\mathrm{str3D}}(R_1, R_2, R_3)$ obtained from the first contraction step and (2) the Z^{BF} axis is defined by the direction of the central stretch. In this non-Eckart frame,[25,26] the vibration−rotation separation is not optimized. It is then important to allow full mixing of the different P sets of vibronic states, even to describe accurately ideal Hund's case a.

On the other hand, this contraction step appears as being most important for assignments. At least for the lower states, it is generally possible to assign most of quanta in each vibrational mode. Then, during the final contraction step, the wave functions will be (generally) quite simple combinations of these assigned levels. Resonances will appear with several significant weights on assigned vibronic contractions.

The assignment can also be facilitated by plotting some cuts of the rovibronic wave functions Ψ or of their squared norm $|\Psi|^2$. The present contraction scheme simplifies the sum of factorized amplitudes. For instance, cuts following (θ_1, θ_2) require the sum of amplitudes for which the stretch part is the same.

Two-dimensional cuts following (θ_1, θ_2) are especially useful for analysis of Hund's case b, which are localized on one of both electronic components.[2] As an example, the shapes of $|\Psi|^2$ for the first two Σ_u states of HCCH$^+$ shown in Figure 5 follow the shapes of the PES associated with the electronic component.

For HCCH$^+$, the difficulty of defining the Renner−Teller parameter ε_5 associated with the cis bending mode, since both electronic components are very close together, was

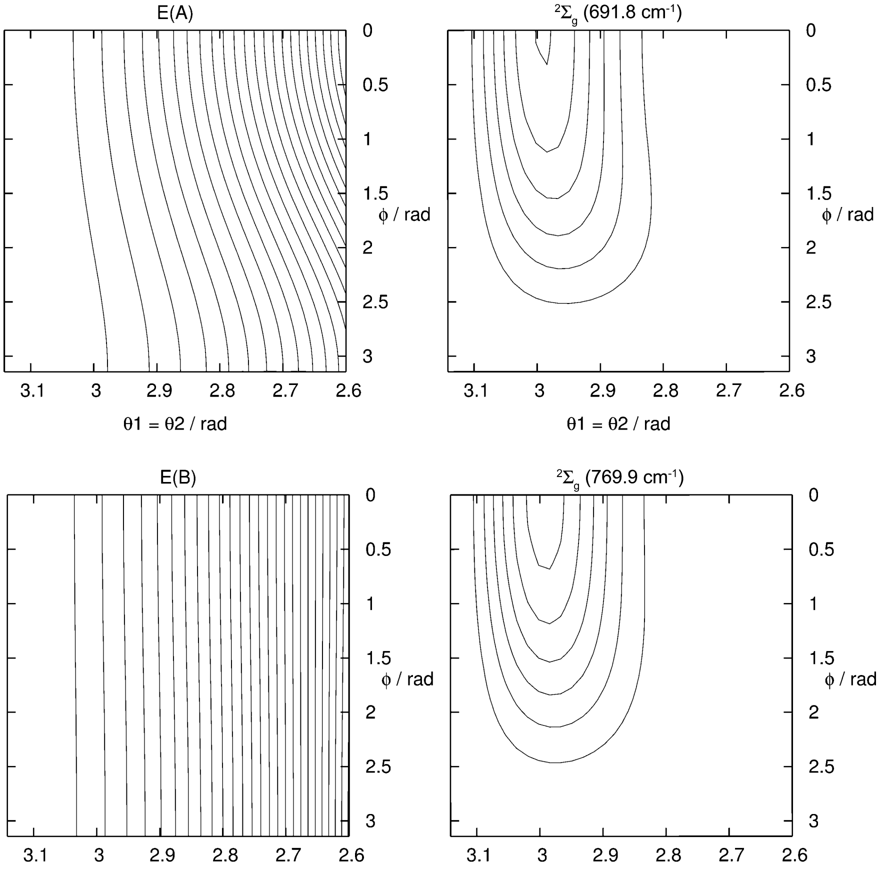


Figure 6. HCCH$^+$: Comparison between two-dimensional contour plots for the potential and both Σ states with one quantum in the cis bending mode. R_1, R_2, and R_3 are fixed at their equilibrium values.

mentioned in several experimental[7,8] and theoretical works.[2,10] Then the perturbative approach based on the harmonic approximation is not adapted to reproduce the cis bending energies. At the end, the assignment of the lower Σ state associated with one quantum in the cis bending mode can be helped by the shape of the electronic component associated with each Σ_g state. Whereas the cuts following (θ_1, θ_2) gave similar shapes for each of both $^2\Sigma_{g1/2}(\nu_5)$ states, the two-dimensional view following $[(\theta_1 = \theta_2), \phi]$ allows us to assign each Σ state with the corresponding electronic component, as shown in Figure 6.

3.3.3. Rovibronic States. In the final contraction step, all previous contractions are collected together and coupled by the complete molecular Hamiltonian $\hat{H}$. The only good quantum number is J, associated with the total angular momentum. The final basis set is now

$$\Phi^{\mathrm{rovib}}_{v_i,b_i,J} = \Phi^{\mathrm{vib}}_{v_i,b_i} \cdot |J, P\rangle \cdot \Phi_e^{\pm} \cdot |S, \Sigma\rangle \qquad (35)$$

where $\Phi^{vib}_{v_i,b_i}$ are the contracted vibronic functions of previous contraction steps. As the total kinetic-energy operator $\hat{T}_N$ adapted to the Renner−Teller and spin couplings is concerned, the full mixing between P states is allowed.

The calculations for each value of J are independent. The zero-order coefficient of the potential was removed, and all energy levels have their origin at the global minimum of the potential (at linearity in the case of $HCCH^+$). For each isotopomer, the zero-point energy is relative to the minimum of the PES and all other energies are relative to it.

The label of the final rovibronic states is based on the quantum numbers K_{space} and P, with $P = K_{\mathrm{space}} + \Sigma$. $K_{\mathrm{space}} = 0, \pm1, \pm2, ...$, corresponds to the Σ, Π, and Δ states, respectively. The value of P is given as the subscript, and the label u, g is the result of the combination the electronic (u) and vibrational (u, g for cis,trans bending mode) character with respect to the inversion of the rovibronic state for molecules associated with symmetry point group $D_{\infty h}$.

3.4. Preliminary Test. In order to check the largest contribution to the rovibronic energies, the $HCCH^{2+}$ dication in its fundamental state $X^3\Sigma_g^-$ is used as a testing system. The corresponding PES was obtained by Hochlaf et al.[27] at the RCCSD(T) level of theory and using the cc-pVQZ basis set. This electronic state is treated as a $^1\Sigma^+$ state in our new code, excluding the electronic angular momenta contributions, as well as in RVIB4 code based on Bramley et al.'s methodology.[28] We then neglect the spin−rotation and spin−bending couplings. Hochlaf et al. already used this PES for determination of the low-energy vibrational states at $J = 0$ and 1.[29]

The basis set used in our code was identical as the one used for $HCCH^+$, i.e., 10 contractions for the stretches, $l_{\max} = 15$ for the bending modes, and $\omega_{\max} = 13/2$ for the torsion. For RVIB4, we used 40, 70, and 70 integration points, 10, 31, and 12 initial basis functions, from which 10, 18, and 12 contracted functions are extracted for the stretches, bending modes, and torsion, respectively (see ref 4 for the details). The testing calculations are done for $J = 0$, 1, 2, and 3 and are compiled in Table 6 for rovibrational band origins up to ∼2600 cm^{-1} and for ν_1 and ν_3 fundamental states. The deviation between both sets of values is lower than 3 cm^{-1} for levels involving more than 3 quanta in the same bending mode and less than 1 cm^{-1} in most of cases. The rotational part of the methodology has been more accurately checked by comparing the energy difference, $\Delta E_{K,\,K+1}$, of the rovibrational states between $J = K_{\mathrm{space}}$ and $J = K_{\mathrm{space}} + 1$ calculations. As examples, $\Delta E_{K,\,K+1}$ values are given for some rovibronic levels in Table 7. In this way, the rovibrational part of our code has been validated.

Table 6. $HCCH^{2+}$ Rotational Band Origins (in cm^{-1}) from RVIB4 and the Present Code[a]

	$^3\Sigma_g$			$^3\Delta_g$	
$(v_4^{l_4}, v_5^{l_5})$	RVIB4	this work	$(v_4^{l_4}, v_5^{l_5})$	RVIB4	this work
$(0^0, 0^0)^-$	0.0[b]	0.0[c]			
$(0^0, 2^0)^-$	1280.2	1280.3	$(0^0, 2^2)$	1298.3	1298.4
$(2^0, 0^0)^-$	1348.2	1348.6	$(2^2, 0^0)$	1348.6	1349.6
ν_2	1517.6	1517.7			
$(0^0, 4^0)^-$	2536.7	2538.4	$(0^0, 4^2)$	2553.2	2555.3
ν_1	2737.3	2739.8			

	$^3\Sigma_u$			$^3\Delta_u$	
$(v_4^{l_4}, v_5^{l_5})$	RVIB4	this work	$(v_4^{l_4}, v_5^{l_5})$	RVIB4	this work
$(1^1, 1^1)^-$	1299.3	1299.4	$(1^1, 1^1)$	1318.4	1318.6
$(1^1, 1^1)^+$	1309.1	1309.3			
$(1^1, 3^1)^-$	2551.0	2555.2	$(1^1, 3^1)$[d]	2571.0	2570.7
$(1^1, 3^1)^+$	2566.7	2571.2	$(1^1, 3^3)$[d]	2586.4	2587.6
ν_3	2637.4	2637.7			

	$^3\Pi_g$			$^3\Gamma_g$	
$(v_4^{l_4}, v_5^{l_5})$	RVIB4	this work	$(v_4^{l_4}, v_5^{l_5})$	RVIB4	this work
$(1^1, 0^0)$	669.9	669.9			
$(1^1, 2^0)$[d]	1933.3	1933.7			
$(1^1, 2^2)$[d]	1946.3	1946.8	$(1^1, 2^2)$	1967.8	1968.5
$(3^1, 0^0)$	2035.4	2037.6	$(3^3, 0^0)$	2036.3	2037.8
$\nu_2(1^1, 0^0)$	2172.2	2172.8			

	$^3\Pi_u$			$^3\Gamma_u$	
$(v_4^{l_4}, v_5^{l_5})$	RVIB4	this work	$(v_4^{l_4}, v_5^{l_5})$	RVIB4	this work
$(0^0, 1^1)$	648.3	648.3			
$(0^0, 3^1)$	1916.3	1916.6	$(0^0, 3^3)$	1950.4	1949.9
$(2^0, 1^1)$[d]	1968.1	1968.8			
$(2^2, 1^1)$[d]	1989.8	1991.1	$(2^2, 1^1)$	1998.6	1998.7
$\nu_2(0^0, 1^1)$	2165.3	2165.3			

[a] ν_4, ν_5 correspond to the trans and cis bending modes, respectively. [b] ZPE: 4905.6 cm^{-1}. [c] ZPE: 4905.8 cm^{-1}. [d] Tentative assignment.

Table 7. $HCCH^{2+}$ $\Delta E_{K,\,K+1}$ (in cm^{-1}) from RVIB4 and the Present Code[a]

	$\Delta E_{K,K+1}$	
state $(v_4^{l_4}, v_5^{l_5})$	RVIB4	this work
$^3\Sigma_g^-(0^0, 0^0)^-$	1.9483	1.9486
$^3\Sigma_u^-(1^1, 1^1)^-$	1.9529	1.9534
$^3\Pi_g(1^1, 2^0)$	3.9571	3.9533
$^3\Pi_u(0^0, 1^1)$	3.9168	3.8992
$^3\Delta_g(2^2, 0^0)$	6.9288	5.8456

[a] ν_4, ν_5 correspond to the trans and cis bending modes, respectively.

Table 8. $HCCH^+$ Rotational Band origins (in cm^{-1})

state	assignment	previous work[2]	Tang[7]	Yang[8] [b]	this work
$^2\Pi_{u3/2}$	0	0.0	0.0	0.0	0.0[c]
$^2\Pi_{u1/2}$		28.5	30.8	29.8	28.5
$^2\Sigma_{u1/2}$	ν_4	496.2	502.7	499.5	506.3
$^2\Delta_{u5/2}$		658.6	666.4	672.9	665.0
$^2\Delta_{u3/2}$		685.8	695.8	701.4	692.3
$^2\Sigma_{u1/2}$		902.3	912.6	909.9	897.3
$^2\Sigma_{g1/2}$	ν_5	685.4	697.5	694.9	691.8
$^2\Delta_{g5/2}$		718.9	715.1	713.4	718.7
$^2\Delta_{g3/2}$		747.5	746.0	743.0	747.4
$^2\Sigma_{g1/2}$		776.2	746.6	738.2	769.9
$^2\Pi_u$	$2\nu_4$	1090.7	1109.4	1108.3	1105.1
$^2\Phi_{u7/2}$		1313.5	1327.0	1316.0	1323.2
$^2\Phi_{u5/2}$		1338.8	1354.3	1342.7	1348.8
$^2\Pi_u$		1685.5		1683.5	1682.3
$^2\Pi_g$	$\nu_4 + \nu_5$	1214.9	1210.8[a]	1210.2	1236.8
$^2\Pi_{g3/2}$		1365.6	1361.6	1373.1	1373.2
$^2\Pi_{g1/2}$		1392.9	1390.7	1401.6	1403.8
$^2\Phi_{g7/2}$		1384.5	1384.1	1370.4	1392.0
$^2\Phi_{g5/2}$		1411.9	1414.2	1398.9	1420.0
$^2\Pi_g$		1613.7	1616.8[a]	1620.6	1608.5
$^2\Pi_u$	$2\nu_5$	1392.5	1393.5	1404.8	1399.3
$^2\Phi_{u7/2}$		1423.8	1432.7	1410.7	1439.3
$^2\Phi_{u5/2}$		1452.6	1462.8	1440.5	1468.1
$^2\Pi_u$		1496.4	1459.0[a]	1451.2	1487.4
$^2\Pi_{u3/2}$	ν_2	1819.0		1817.5	1818.9
$^2\Pi_{u1/2}$		1847.5			1846.8
$^2\Pi_{g3/2}$	ν_3	3151.9[d]			3134.3
$^2\Pi_{u3/2}$	ν_1	3236.4[d]			3221.8

[a] Undetermined A_{SO}. [b] Data from Table 7 of ref 8. [c] ZPE: 5572.9 cm^{-1}. [d] Stretching contractions without coupling with the other degrees of freedom.

4. Results

We present here the rovibronic levels computed for $HCCH^+$, $DCCH^+$, and $DCCD^+$ up to two quanta in the bending modes.

4.1. Rotational Band Origins for the Symmetrical $HCCH^+$ and $DCCD^+$. The final rovibronic contracted state energy levels corresponding to rotational band origins for $HCCH^+$ and $DCCD^+$ are displayed in Tables 8 and 9. They are both associated with the $D_{\infty h}$ symmetry point group, and the same assignment is adopted.

In Table 8, the rovibronic energies of $HCCH^+$ coming from valence coordinates development roughly agree with photoelectron spectroscopy data[7,8] and the ones computed previously by Jacobi coordinates treatment. A value of P is given only when it corresponds almost to a good quantum number: (i) for Σ states $P = 1/2$, (ii) in degenerate Hund's case a, for which we obtain a separation between both components $P = K_{\text{space}} \pm 1/2$. On the other hand, for degenerate Hund's case b, spin−rotation couplings play a crucial role in the wave functions, so that both spin components are strongly mixed together except for the lowest value of J.

For both symmetrical isotopomers, the classification of Hund's cases a and b is very standard for systems having a weak spin−orbit constant: Hund's case b go by pairs, with one state on the lower electronic component and the other one on the upper component. Single states are Hund's case a, delocalized on both electronic components. It is worth noting that Σ states are generally Hund's case b, and they always are if the spin−orbit constant is negligible.

Table 9. $DCCD^{\ +}$ Rotational Band origins (in cm^{-1})

state	assignment	Perić[11] [a]	this work
$^2\Pi_{u3/2}$	0	0	0.0[b]
$^2\Pi_{u1/2}$		0 + 28.46	28.8
$^2\Sigma_{u1/2}$	ν_4	418 − 0.56	420.9
$^2\Delta_{u5/2}$		573 − 27.04/2	552.9
$^2\Delta_{u3/2}$		573 + 27.04/2	580.2
$^2\Sigma_{u1/2}$		782 + 0.54	751.7
$^2\Sigma_{g1/2}$	ν_5	539 − 8.97	511.7
$^2\Delta_{g5/2}$		546 − 28.46/2	530.0
$^2\Delta_{g3/2}$		546 + 28.46/2	558.9
$^2\Sigma_{g1/2}$		553 + 8.95	574.4
$^2\Pi_u$	$2\nu_4$	923	917.6
$^2\Phi_{u7/2}$		1077 − 28.45/2	1060.9
$^2\Phi_{u5/2}$		1077 + 28.45/2	1090.0
$^2\Pi_u$		1435	1397.1
$^2\Pi_g$	$\nu_4 + \nu_5$	950	959.2
$^2\Pi_{g3/2}$		1103 − 27.04/2	1078.0
$^2\Pi_{g1/2}$		1103 + 27.04/2	1105.8
$^2\Phi_{g7/2}$		1089 − 27.04/2	1087.1
$^2\Phi_{g5/2}$		1089 + 27.04/2	1114.8
$^2\Pi_g$		1314	1274.4
$^2\Pi_u$	$2\nu_5$	1068	1034.4
$^2\Phi_{u7/2}$		1115 − 25.15/2	1098.0
$^2\Phi_{u5/2}$		1115 + 25.15/2	1123.7
$^2\Pi_u$		1087	1104.7
$^2\Pi_{u3/2}$	ν_2		1637.6
$^2\Pi_{u1/2}$			1665.9
$^2\Pi_{g3/2}$	ν_3		2326.4 − 2327.3[c]
$^2\Pi_{u3/2}$	ν_1		2571.4

[a] $E_{\text{niv}} = E_{\text{rovib}} + E_{\text{SO}}$; the effect of the spin−orbit coupling is added to the rovibronic energies. Except for states $^2\Pi_{u3/2}$ and $^2\Pi_{u1/2}$, all values of E have been shifted by 14 cm^{-1}. [b] ZPE: 4405.5 cm^{-1}. [c] Resonance with a Hund's case b state. Both states energies are given.

For $DCCD^+$, the present results are in good agreement with the values determined by Perić et al.,[11] who separated the effects of the spin−orbit coupling as emphasized in Table 9. Except for the fundamental states, all rovibronic energies, without the effect of the spin−orbit coupling, are shifted by 14 cm^{-1} in order to facilitate the comparison. As in $HCCH^+$ (see Figure 8 in ref 2), the non-negligible spin−orbit coupling between the first two Σ states with one quantum in the cis bending mode (almost ±8.97 cm^{-1} for Perić et al., while we find ±5.6 cm^{-1}) is the signature of an intermediate nature between Hund's cases a and b due to the weak Renner−Teller splitting between both electronic components in the cis conformation (see section 3.2). Indeed, factorization of the total wave function by one of each electronic component involves a cancellation of the spin−orbit coupling, while factorization by one of each eigenfunction of $\hat{L}_z$, giving rise to a maximum contribution of the spin−orbit coupling, corresponds to a complete delocalization on both electronic components. Renner−Teller effect and spin−orbit couplings are then directly in competition for the Σ states. Moreover, we notice that the rotational excited states of Π states at 1105.8 and 1104.7 cm^{-1} are strongly mixed.

We also give the first excitation of the stretching modes. For $HCCH^+$, $\nu_3 = 3134.3$ cm^{-1} and $\nu_1 = 3221.8$ cm^{-1}. These values are in remarkable agreement with the best experimental results at 3135.9813 and 3226.6 cm^{-1}, respectively, from refs 30 and 31. From photoelectron spectroscopy, Reutt et al. deduced $\nu_2 = 1829.0(2.5)$, 1651(4) cm^{-1} agreeing within 10, 14 cm^{-1} with the present computed energies for

Table 10. HCCD$^+$ Rovibronic Levels (in cm^{-1})[a]

state	assignment	Perić[12] [b]	this work
${}^2\Pi_{3/2}$	0	14	0.0[c]
${}^2\Pi_{1/2}$		14	28.7
${}^2\Sigma_{1/2}$	ν_5	438	442.3
${}^2\Delta_{5/2}$		561	543.9
${}^2\Delta_{3/2}$		561	571.9
${}^2\Sigma_{1/2}$		621	582.3
${}^2\Sigma_{1/2}$	ν_4	656	692.6
${}^2\Delta_{5/2}$		711	690.5
${}^2\Delta_{3/2}$		711	718.7
${}^2\Sigma_{1/2}$		889	847.0
${}^2\Pi$	$2\nu_5$	946	953.4
${}^2\Phi_{7/2}$		1102	1088.3
${}^2\Phi_{5/2}$		1102	1115.5
${}^2\Pi$	resonance	1114	1099.9
${}^2\Pi_{3/2}$	$(\nu_4 + \nu_5)/(2\nu_5)$	1183	1140.1
${}^2\Pi_{1/2}$		1183	1160.5
${}^2\Phi_{7/2}$	$\nu_4 + \nu_5$	1250	1229.7
${}^2\Phi_{5/2}$		1250	1256.8
${}^2\Pi$		1264	1261.0
${}^2\Pi_{(1/2)}$		1466	1434.8
${}^2\Pi_{(3/2)}$		1466	1441.8
${}^2\Pi_{3/2}$	$2\nu_4$	1341	1347.3
${}^2\Pi_{1/2}$		1341	1368.0
${}^2\Phi_{7/2}$		1404	1385.8
${}^2\Phi_{5/2}$		1404	1413.7
${}^2\Pi$		1658	1596.1
${}^2\Pi_{3/2}$	ν_3		1717.2
${}^2\Pi_{1/2}$			1745.0
${}^2\Pi_{3/2}$	ν_2		2454.8
${}^2\Pi_{3/2}$	ν_1		3184.0

[a] ν_4, ν_5 correspond to the bending of the $C\hat{C}H$, $D\hat{C}C$ angles, respectively. [b] $E_{\mathrm{niv}} = E_{\mathrm{rovib}}$; the effect of the spin−orbit coupling has not been taken into account in ref 12. All values of E have been shifted by 14 cm^{-1}. [c] ZPE: 4990.6 cm^{-1}.

HCCH$^+$, DCCD$^+$, respectively.[9] For DCCD$^+$, they obtained $\nu_1 = 2572(14)$ cm^{-1}, which coincides with the present calculated value of 2571.4 cm^{-1}. They also extracted, from experimental data, a Renner−Teller multiplet origin for the trans bending mode ν_4 at 702(12) cm^{-1} for DCCD$^+$. From our experience, the ${}^2\Delta_{u5/2}$ is the closest rovibronic level to the perturbative Renner−Teller multiplet origin. We computed ${}^2\Delta_{u5/2}$ at 552.9 cm^{-1} quite far from this extracted experimental value. We cannot make a conclusion about the accuracy of either the experimental measurement or the theoretical calculations since, for Renner−Teller systems including spin−orbit coupling, the extraction of spectroscopic constants from observed spectra is sometimes quite problematic.[8,10] In particular, in more than three atom molecules, the number of interdependent perturbative parameters can make the accurate determination of each one difficult.

4.2. Rotational Band Origins for the Nonsymmetrical DCCH$^+$. For DCCH$^+$, the inversion symmetry disappears and the assignment of the bending levels raises further difficulties. Indeed, both bending modes belong to the same representation Π and new resonances are allowed. The ν_4 and ν_5 modes are mostly associated with the $\widehat{CCH}$, $\widehat{DCC}$ bending angles, respectively. As in Table 8, a value of P is given only when it corresponds almost to a good quantum number.

The final rovibronic state energy levels corresponding to rotational band origins for DCCH$^+$ are displayed in Table 10, in comparison with the results obtained by Perić et al. without spin−orbit coupling.[12] All their energies are shifted by 14 cm^{-1} in order to have a common reference. Most of these results are in good agreement. The main significant divergence is due to a resonance that we found between ν_4

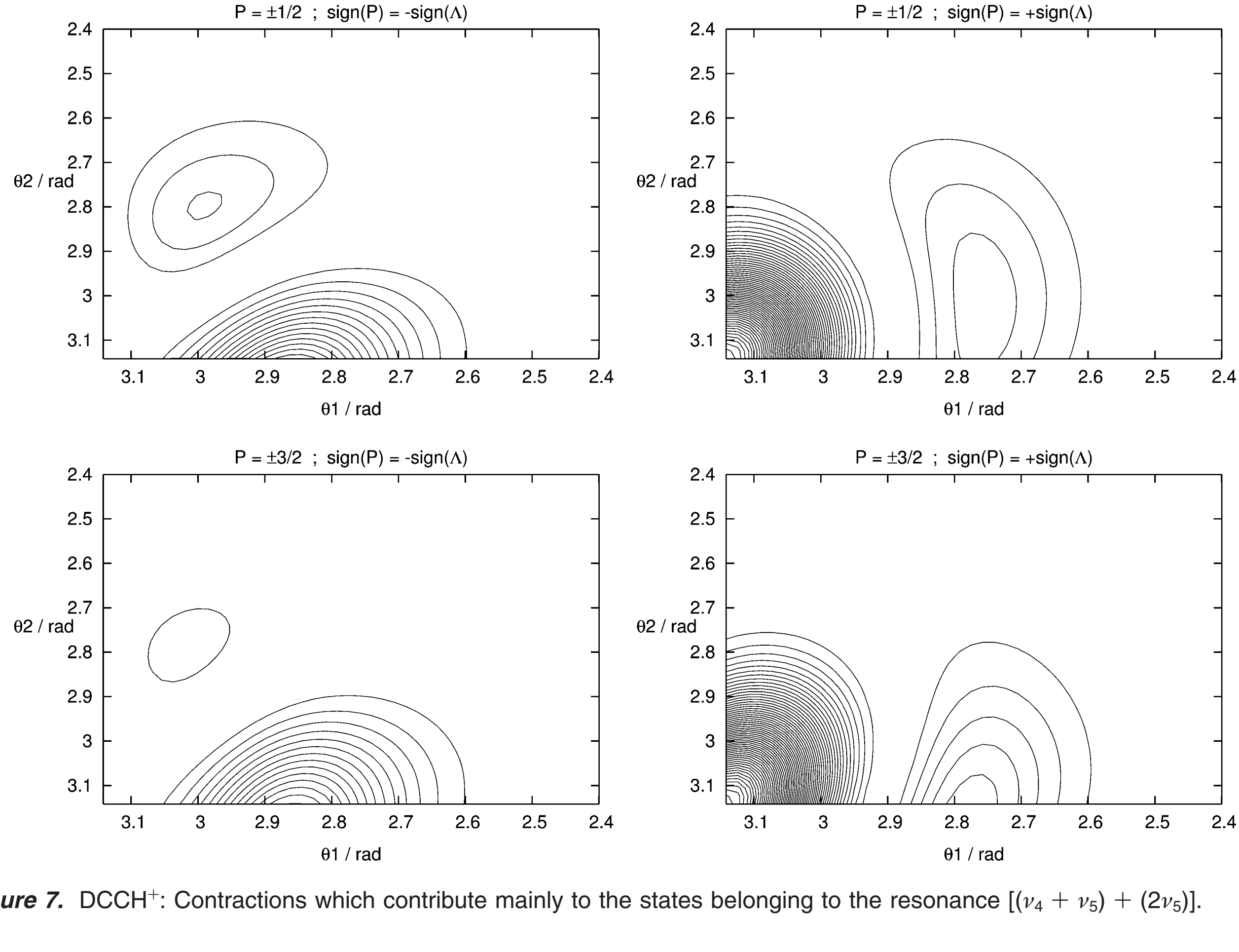


Figure 7. DCCH$^+$: Contractions which contribute mainly to the states belonging to the resonance $[(\nu_4 + \nu_5) + (2\nu_5)]$.

and ν_5. Indeed, the assignments of both Σ states at 582.3 and 692.6 cm^{-1} are approximative, while they are both coupled together. This leads to a minimization of the lower state energy and to an augmentation of the upper state energy. This fact explains the inversion between the energies of the Σ state at 692.6 cm^{-1} and of the Δ state at 690.5 cm^{-1}.

With two quanta in bending modes, the structure in Hund's cases a and b of Π states is not well conserved. Even an "anti-Hund's case a" contribution is observed in both states assigned Π at 1434.8 and 1441.8 cm^{-1}, for which component $|P| = 1/2$ possesses a lower energy than component $|P| = 3/2$. Moreover, the spin−orbit splitting for Hund's case a is reduced to around 20 cm^{-1} instead of 27 cm^{-1} in HCCH$^+$. On the other hand, a standard repartition such as the one obtained for both symmetrical isotopomers would lead to only one Hund's case a pair of states, i.e., ($^2\Pi_{g3/2}/^2\Pi_{g1/2}$), instead of two ($^2\Pi_{3/2}/^2\Pi_{1/2}$) pairs at energies (1140.1/1160.5) and (1347.3/1368.0) cm^{-1}. However, the averaged Hund's cases a and b are conserved.

The resonance between ($\nu_4 + \nu_5$) and $2\nu_5$ can be visualized by plotting each contribution $\pm\ \Lambda$ independently. As this resonance couples different values of Λ, the distorted wave function shapes usually obtained in the case of Fermi resonances as in ref 32 are avoided. In Figure 7, two-dimensional contours of the contractions which contribute mainly to the Hund's case b $^2\Pi$ state obtained at 1099.9 cm^{-1} are plotted. The rotational band origin, for which $|P| = 1/2$, is only concerned in the first two plots, while all states for which $J \geq 1/2$ are combinations of the four plots. Perić et al. also expect a significant contribution of $2\nu_4$ for this state. This last point could explain, for both left-hand side figures, the less important part of the wave function following the θ_2 axis, since both ($\nu_4 + \nu_5$) and $2\nu_4$ contractions contribute to the wave functions with $sign(P) = -sign(\Lambda)$, but they are in phase opposition. However the complete decomposition of the total spin-rovibronic functions is more complicated to analyze, since the contributions corresponding to $sign(P) = -sign(\Lambda)$ do not have a significant weight at linearity, in contrast with $2\nu_4$. For comparison, Figure 8 shows the main contributions of the more standard Hund's case b $^2\Pi$ state obtained at 1261.0 cm^{-1} assigned to ($\nu_4 + \nu_5$).

In conclusion, for nonsymmetrical systems, the diminution of the irreducible representations number entails much more resonance phenomena, even at low energy. In addition, because Renner−Teller systems involve two times more states for a given number of quanta in vibrational modes and because the hierarchy between the corresponding coupling and the spin−orbit one is not always sharply contrasted, it is often impossible to achieve a standard analysis in terms of assignments and Hund's cases as in perturbative approaches. Most of them are indicative.

4.3. Rotational Structures. As already discussed in previous articles,[1,2] the rotational structures attempted for Hund's cases a and b states are very different. For Hund's case a states, two independent rotational structures can be defined from both spin components $P = K_{\text{space}} \pm 1/2$. For Hund's case b states, P is not a good quantum number and the spin−rotation coupling has a crucial role as well as spin-bending couplings. However, spin−rotation couplings also have an effect on the rotational structure of Hund's case a. Indeed, small couplings between both spin components induce a mix between them. Even whether it concerns only few 10^{-2}%, it is sufficient to affect the effective rotational constants B_{eff} (defined as the energy difference of the $J = P$ and $J = P + 1$ states for a given rotational band origin divided by $(P + 1)(P + 2) - P(P + 1)$). For instance, in the case of the $^2\Pi$ fundamental state with no quantum in vibrational modes: $B_{\text{eff}(P=3/2)} = 1.064$ cm^{-1} and $B_{\text{eff}(P=1/2)} = 1.139$ cm^{-1} for HCCH$^+$, $B_{\text{eff}(P=3/2)} = 0.904$ cm^{-1} and $B_{\text{eff}(P=1/2)} = 0.958$ cm^{-1} for DCCH$^+$, and $B_{\text{eff}(P=3/2)} = 0.777$ cm^{-1} and $B_{\text{eff}(P=1/2)} = 0.818$ cm^{-1} for DCCD$^+$. For Hund's case b states, the rotational structures of the first three energy states are displayed in Table 11. In the case of HCCH$^+$, the present values are compared with the experimental work made by Yang et al.[8] and with our previous work. The global agreement makes us confident with our results for DCCH$^+$ and DCCD$^+$, which are predictive.

The most common rotational structures for Σ Hund's case b states is as follows: a band origin given by the ($J = 1/2$, $F1$) component; a succession of couple of states very close in energy, corresponding to (J, $F2$) and ($J + 1$, $F1$). Both states cannot be coupled without an external field, because J is still a good quantum number. Their energy order is often difficult to define and can vary under even very small couplings of any kind with other states.

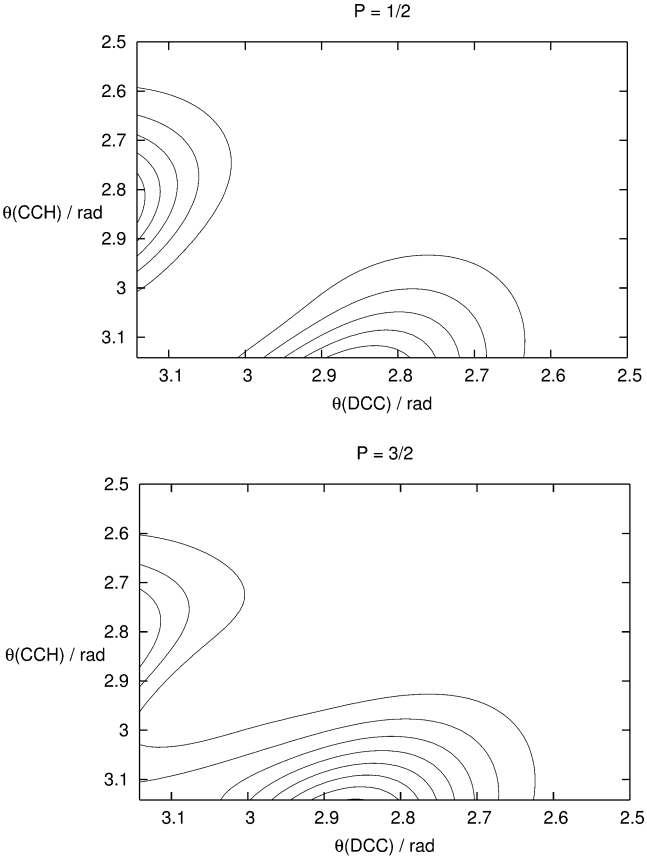


Figure 8. DCCH$^+$: Contractions which contribute mainly to the Hund's case b $^2\Pi$ state at 1261.0 cm^{-1} assigned to ($\nu_4 + \nu_5$).

Table 11. Rotational Structures (in cm^{-1}) from a Nondegenerate Σ and Two Degenerate Π Vibronic States (all Hund's case b states)[a]

lowest energy $^2\Sigma$ state						
HCCH$^+$					DCCH$^+$	DCCD$^+$
band origin-ZPE: Yang 499.5; previous work 496.2; this work 506.3					band origin-ZPE: this work 442.3	band origin-ZPE: this work 420.9
J	F	Yang	previous work	this work	this work	this work
1/2	F1	0.0	0.0	0.0	0.0	0.0
1/2	F2		2.2	2.2	1.8	1.6
3/2	F1	2.3	2.1	2.2	1.9	1.6
3/2	F2	6.7	6.7	6.6	5.6	4.8
5/2	F1		6.5	6.6	5.6	4.8
5/2	F2	13.4	13.3	13.2	11.2	9.6
7/2	F1		13.1	13.2	11.2	9.6
7/2	F2		22.1	22.0	18.6	15.9

lowest energy $^2\Pi$ state with two quanta in bending modes						
HCCH$^+$					DCCH$^+$	DCCD$^+$
band origin-ZPE: Yang 1108.3; previous work 1090.7; this work 1105.1					band origin-ZPE: this work 953.4	band origin-ZPE: this work 917.6
J	F	Yang	prev. work	this work	this work	this work
1/2	F1		1.9	1.8	0.9	1.8
3/2	F1	0.0	0.0	0.0	0.0	0.0
3/2	F2		5.9	5.9	4.4	4.7
5/2	F1	4.6	4.6	4.6	3.9	3.4
5/2	F2		12.3	12.3	9.9	9.3
7/2	F1	11.3	11.3	11.4	9.6	8.4
7/2	F2	20.9	21.0	21.0	17.4	15.7

second $^2\Pi$ state with two quanta in bending modes						
HCCH$^+$					DCCH$^+$	DCCD$^+$
(*J* = 1/2, F1)-ZPE: Yang 1210.2; previous work 1214.9; this work 1237.4					band origin-ZPE: this work 1099.9	band origin-ZPE: this work 959.2
J	F	Yang	previous work	this work	this work	this work
1/2	F1	0.0	0.0	0.0	0.0	0.4
3/2	F1		−0.8	−0.6	1.7	0.0
3/2	F2	4.2	4.2	4.2	5.2	3.5
5/2	F1	4.6	3.7	3.9	5.4	3.3
5/2	F2		10.8	10.6	10.9	8.2
7/2	F1		10.4	10.5	11.1	8.1
7/2	F2		19.6	19.7	18.4	14.6

[a] Yang and previous work are associated with refs 8 and 2, respectively.

The most common rotational structures for degenerate Π Hund's case b states are as follows: a band origin given by the (J = 3/2, $F1$) component, followed by (J = 1/2, $F1$). The energy difference between both is variable and depends on small spin−orbit couplings, spin−rotation, spin−bending, and a succession of couple of states very close in energy, corresponding to (J, $F2$) and (J + 1, $F1$), with the same remark as for Σ states.

The only case for which a deviation from this general scheme is observed concerns the rotational structure from the Π state of DCCH$^+$ obtained at 1099.9 cm^{-1}. We have already seen that this state belongs to a complicated resonance between several assignments with two quanta in the bending modes.

Actually, the rotational structures shown in Table 11 are simplified. Indeed, in the molecular symmetry groups $D_{\infty h}(MS)$ and $C_{\infty v}(MS)$, all irreducible representations are nondegenerate, in contrast with the corresponding symmetry point groups. The notations (J, F) should correspond to degenerate states, except for rotational structures from Σ states; however, the corresponding degeneracy is slightly raised. Then numbers in Table 11 are given with only one digit since inside each pair of states energies are very close together at low values of J but can be spaced by few times 0.1 cm^{-1} for $J \geq 7/2$. These effects should be pointed out for higher values of J and vary a lot particularly because the couplings with Σ states affect both components of a given (J, F) assignment independently.

5. Conclusions

In the present work, a new variational methodology for treatment of the Renner−Teller effect in tetra-atomic molecules is developed in valence coordinates. The kinetic-energy operator of Bramley et al.[3,4] for any sequentially bonded four-atom molecule, A−B−C−D, in a singlet nondegenerate electronic state has been adapted to the Renner−Teller and spin couplings by modifying the expression of the rotational angular momentum. The total Schrödinger equation is solved by diagonalizing the Hamiltonian matrix in a three-step contraction scheme. The present methodology has been checked by comparing rovibrational energies of the $X^3\Sigma_g^-$ electronic ground state of $HCCH^{2+}$ obtained from the present code, without taking into account the orbital and spin electronic angular momenta, and a variational code based on Bramley et al. works. Both sets of results are in remarkably good agreement for the rovibrational band origins as well as for rotational structures.

The main advantage of this new theoretical development is the possibility of studying different isotopomers using the same potential-energy surfaces. This procedure has been tested on $HCCH^+$ and its deuterated derivatives $DCCD^+$ and $DCCH^+$. The calculated rovibronic band origins have been compared with previous data deduced from Jacobi coordinates methodology,[1,2] dimensionality reduced variational treatment,[11,12] and photoelectron spectra[7−9] with an overall good agreement.

Finally, this new methodology will permit the variational treatment of systems for which Jacobi coordinates involve much too high crossing terms in the PESs such as HCCS or HCCO where at least one external atom possesses a comparable or higher weight than the central atoms directly linked to it.

Acknowledgment. This work was partially supported by the French "Centre National de la Recherche Scientifique" (CNRS). The authors thank Prof. M. Hochlaf for discussions on the $HCCH^{2+}$ PES.

References

(1) Jutier, L.; Léonard, C.; Gatti, F. *J. Chem. Phys.* **2009**, *130*, 134301.

(2) Jutier, L.; Léonard, C.; Gatti, F. *J. Chem. Phys.* **2009**, *130*, 134302.

(3) Bramley, M. J.; Green, W. H.; Handy, N. C. *Mol. Phys.* **1991**, *73*, 1183.

(4) Bramley, M. J.; Handy, N. C. *J. Chem. Phys.* **1993**, *98*, 1378.

(5) Carter, S.; Handy, N. C.; Rosmus, P.; Chambaud, G. *Mol. Phys.* **1990**, *71*, 605.

(6) Carter, S.; Handy, N. C.; Puzzarini, C.; Tarroni, R.; Palmieri, P. *Mol. Phys.* **2000**, *98*, 1697.

(7) Tang, S.-J.; Chou, Y.-C.; Lin, J. J.-M.; Hsu, Y.-C. *J. Chem. Phys.* **2006**, *125*, 133201.

(8) Yang, J.; Mo, X. *J. Phys. Chem. A* **2006**, *110*, 11001.

(9) Reutt, J. E.; Wang, L. S.; Pollard, J. E.; Trevor, D. J.; Lee, Y. T.; Shirley, D. *J. Chem. Phys.* **1985**, *84*, 3022.

(10) Perić, M.; Thümmel, H.; Marian, C. M.; Peyerimhoff, S. D. *J. Chem. Phys.* **1995**, *102*, 7142.

(11) Perić, M.; Radić-Perić, J. *J. Chem. Phys. Lett.* **1998**, *290*, 443.

(12) Perić, M.; Ostojić, B.; Radić-Perić, J. *J. Chem. Phys.* **1999**, *110*, 4783.

(13) Žabka, J.; Dolejšek, Z.; Hrušák, J.; Herman, Z. *Int. J. Mass Spectrom.* **1999**, *185/186/187*, 95.

(14) Bockelee-Morvan, D.; Gautier, D.; Lis, D.; Young, K.; Keene, J.; Phillips, T.; Owen, T.; Crovisier, J.; Goldsmith, P.; Bergin, E.; Despois, D.; Wootten, A. *Icarus* **1999**, *133*, 147.

(15) Zare, R. N. *Angular Momentum*; Wiley: New York, 1988.

(16) Gatti, F.; Nauts, A. *Chem. Phys.* **2003**, *295*, 167.

(17) Klein, O. Z. *Phys.* **1929**, *58*, 730.

(18) Brown, J. M.; Howard, B. J. *Mol. Phys.* **1976**, *31*, 1517.

(19) Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1993**, *99*, 5219.

(20) Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **2000**, *112*, 3106.

(21) Watts, J. D.; Gauss, J.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 8718.

(22) Dunning, J., T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(23) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M. MOLPRO: a package of ab initio programs, version 2008.1; http://www.molpro.net.

(24) Jutier, L. Manuscript in preparation.

(25) Eckart, C. *Phys. Rev.* **1935**, *47*, 552.

(26) Sayvetz, A. *J. Chem. Phys.* **1939**, *7*, 383.

(27) Hochlaf, M. Private communication.

(28) RVIB4 is a tetraatomic rovibrational variational code, see: Carter, S.; Handy, N. *J. Mol. Spectrosc.* **1998**, *192*, 263. *J. Phys. Chem. A*, **1998**, *102*, 6325. *Mol. Phys.*, **1997**, *90*, 729. *J. Mol. Spectrosc.*, **1993**, *157*, 301, and references therein.

(29) Hochlaf, M.; Palaudoux, J.; Ben Houria, A. *Recent Res. Dev. Chem. Phys.* **2004**, *5*, 403.

(30) Jagod, M.-F.; Rösslein, M.; Gabrys, C. M.; Rehfuss, B. D.; Scappini, F.; Crofton, M. W.; Oka, T. *J. Chem. Phys.* **1992**, *97*, 7111.

(31) Forney, D.; Jacox, M. E.; Thompson, W. E. *J. Mol. Spectrosc.* **1992**, *153*, 680.

(32) Stimson, S.; Evans, M.; Ng, C.; Hsu, C.-W.; Heimann, P.; Destandau, C.; Chambaud, G.; Rosmus, P. *J. Chem. Phys.* **1998**, *108*, 6205.

CT100071D

JCTC Journal of Chemical Theory and Computation

Convergence of Nuclear Magnetic Shieldings in the Kohn−Sham Limit for Several Small Molecules

Teobald Kupka,*,† Michał Stachów,† Marzena Nieradka,† Jakub Kaminsky,*,‡ and Tadeusz Pluta§

University of Opole, Faculty of Chemistry, Poland, Department of Molecular Spectroscopy, Institute of Organic Chemistry and Biochemistry, Prague, Czech Republic, and University of Silesia, Institute of Chemistry, Katowice, Poland

Received February 23, 2010

Abstract: Convergence patterns and limiting values of isotropic nuclear magnetic shieldings were studied for several small molecules (N_2, CO, CO_2, NH_3, CH_4, C_2H_2, C_2H_4, C_2H_6, and C_6H_6) in the Kohn−Sham limit. Individual results of calculations using dedicated families of Jensen's basis sets (pcS-n and pcJ-n) were fitted toward the complete basis set limit (CBS) using a simple two-parameter formula. Several density functionals were used; calculated vibrational corrections (ZPV) applied; and, for comparison purposes, similar calculations performed using RHF, MP2, SOPPA, SOPPA(CCSD), and CCSD(T) methods and additionally, the aug-cc-pVTZ-J basis set. Finally, the CBS estimated results were critically compared with earlier reported literature data and experimental results. Among 42 studied DFT methods, the KTn and "pure" functionals produced the most accurate heavy atom isotropic nuclear shieldings.

I. Introduction

Nuclear shieldings belong to the most important spectral features and are often nowadays predicted at several levels of theory (Hartree−Fock, HF; Density Functional Theory, DFT; Möller-Plesset second-order perturbation theory, MP2; or Coupled Cluster with singles, doubles, and perturbative treatment of triple excitations, CCSD(T)).[1–6] However, very sophisticated and expensive coupled-cluster methods and large basis sets appeared to be necessary to obtain quantitative ^{13}C, ^{19}F, or ^{17}O NMR parameters.[3,7,8]

Recently, the complete basis set limit (CBS) approach, typical for estimation of accurate energy,[9] has been adopted for NMR calculations, too.[10–13] Initially, Dunning's correlation-consistent basis sets (aug-cc-pVxZ, where x = D, T, Q, 5, 6, and sometimes 7) have been applied for accurate evaluations of energy and other molecular and spectroscopic properties.[14–16] Very recently, a detailed overview of estimating CCSD(T) structural parameters in the complete basis set limit was reported by Puzzarini.[17] Jensen proposed general purpose polarization-consistent basis sets,[18–24] pc-n (where n = 0, 1, 2, 3 and 4), capable of regular convergence. Later, he published modified versions of polarized-consistent basis sets, pcS-n,[24] designed for nuclear shieldings. The energy, and other parameters, including nuclear shieldings obtained with polarization-consistent basis sets, were estimated in the CBS limit with accuracy similar to those obtained with correlation-consistent basis sets.[11] Another family of regularly converging basis sets was proposed by Jorge et al.[25–27]

The electron correlation methods, MP2 and even more coupled-cluster methods, are computationally very expensive, and therefore density functional theory including some amount of electron correlation is very promising in studies of larger molecular systems.[28] For example, the BHandH hybrid density functional, capable of correctly reproducing π-stacking geometry and interactions,[29,30] was recently reported as superior to B3LYP and seemed to be the most accurate DFT functional among over 20 others for predicting water's CBS estimated isotropic shieldings.[31] On the other hand, theoretical methods are verified by comparison with reliable gas-phase experimental data.[32] However, it is not

* Corresponding author tel.: +48 665 921 475; fax: +48 77 452 7101; e-mail: teobaldk@yahoo.com (T.K.), kaminskj@gmail.com (J.K.).

† University of Opole.

‡ Institute of Organic Chemistry and Biochemistry.

§ University of Silesia.

10.1021/ct100109j © 2010 American Chemical Society
Published on Web 03/31/2010

easy to find accurate values of experimental nuclear shieldings in the gas phase.[32] For example, the experimental value of water oxygen shielding was recently significantly revised—from 344 ± 17.2[33] to 323.6[34] and 323.5 ± 6 ppm.[35] The latter value has been currently modified to 325.3 ± 3 ppm.[36]

Unfortunately, currently available density functionals are usually semiempirical, calibrated mainly on energy of selected molecular systems.[28] Thus, there is an open question in the literature of which density functional provides the most accurate nuclear shieldings.[12,31,37,38] And, is there a unique density functional, or a group of well performing ones, in predicting NMR properties in the Kohn−Sham limit? Therefore, a selection of high-quality DFT functional(s) for prediction of nuclear shieldings is of vital importance for a wide community, currently using predominantly B3LYP as the method of choice.

This study addresses the problem of standardization of DFT for predicting nuclear shieldings of several small and common inorganic and organic molecules (N_2, CO, CO_2, NH_3, CH_4, C_2H_2, C_2H_4, C_2H_6, and C_6H_6). Over 40 pure and hybrid density functionals were selected somewhat arbitrarily from the recent Gaussian 09 program edition.[39] The HF and MP2 as well as KT1, KT2, and KT3 density functionals, claiming to be the best in predicting oxygen nuclear shieldings, were studied (see also, refs 31, 37, 40–42). In addition, SOPPA and SOPPA(CCSD) methods, which produce accurate spin−spin coupling parameters, were tested. Additionally, for verification purposes, CCSD(T) calculations were used as a gold standard. In addition, for meaningful comparison of theory with experimental data in the gas phase, the vibrational corrections to the nuclear shieldings were applied. Among selected molecules, N_2 and CO are fairly challenging for computations, and they have been described previously using a multiconfigurational approach (MC-SCF[43]). However, to avoid extending the current work, this method was not considered in our studies.

Apart from the theoretical method, the selection of a regularly converging basis set family is important for the current studies. As a continuation of our recent works,[11,31] Jensen's basis set hierarchy pcS-*n*,[24] designed for accurate reproduction of nuclear shieldings, was chosen. Jensen's pcJ-*n*[44] series and aug-cc-pVTZ-J,[45,46] typically used in calculations of SSCC parameters, were additionally studied to check whether basis sets, solely designed for accurate calculations of spin−spin couplings, would correctly reproduce nuclear shieldings too. The latter basis set is relatively small, though efficient for calculating spin−spin couplings.[31,47]

II. Computational Details

Most DFT calculations (approximately 40 density functionals), as well as computations at the HF or MP2 levels, were performed using the Gaussian 09 program.[39] To be concrete, 7 "pure" and 31 "hybrid" density functionals were selected. In addition, three exchange-correlation density functionals KT*n* (where $n = 1$, 2, and 3),[37,40–42] recommended for isotropic nuclear shielding calculations, and the SOPPA and SOPPA(CCSD) methods were performed with the Dalton 2.0 code.[38] The CCSD(T) results were obtained with the Acess 2.0 program.[48] For some molecules, due to the demands of CCSD(T), SOPPA, and SOPPA(CCSD) calculations, the results obtained with the largest affordable basis set were used for comparison with CBS values obtained at DFT, RHF, and MP2 levels of theory. Both MP2 and CCSD(T) calculations were performed using the "Frozen-Core, FC" option (see reference 11 for a comparison of NMR accuracy for a water molecule calculated with "all-electrons" and "FC" schemes). In the subsequent parts of this paper, all 47 selected computational methods will appear in the following order: VXSC (1), HCTH (2), HCTH97 (3), HCTH147 (4), THCTH (5), M06L (6), B97D (7), B3LYP (8), B3P86 (9), B3PW91 (10), B1B95 (11), MPW1PW91 (12), MPW1LYP (13), MPW1PBE (14), MPW3PBE (15), B98 (16), B971 (17), B972 (18), PBE1PBE (19), B1LYP (20), O3LYP (21), BHandH (22), BHandHLYP (23), BMK (24), M06 (25), M06HF (26), M062X (27), tHCTHhyb (28), HSEh1PBE (29), HSE2PBE (30), PBEh1PBE (31), wB97XD (32), wB97 (33), wB97X (34), TPSSh (35), X3LYP (36), LC-wPBE (37), CAM-B3LYP (38), WP04 (39), RHF (40), MP2 (41), KT1 (42), KT2 (43), KT3 (44), SOPPA (45), SOPPA(CCSD) (46), and CCSD(T) (47).

For comparison with earlier studies, the experimental geometries from Bak and co-workers'[49] compilation were used in all NMR calculations. Nuclear shieldings were obtained using the Gauge Including Atomic Orbitals (GIAO) approach.[50–52] All NMR calculations were performed at the single level, and with no interacting (free) molecule, resembling the gas phase in the absence of intermolecular interactions (at zero gas density and without solvent present).

Two sets of Jensen's polarization-consistent basis set families, pcS-n[24] and pcJ-n,[44] were selected. These basis set hierarchies are dedicated to accurate calculations of nuclear shieldings and spin−spin coupling constants, respectively. The former are significantly smaller and somehow "pruned" from the latter ones, developed solely for accurate prediction of J-couplings. In case of the N_2 molecule, the calculations for $n = 0$, 1, 2, 3, and 4 were tested. Due to inaccuracies in smaller basis set designs, the initial results ($n = 0$ and sometimes 0 and 1) were considered meaningless, and the convergence of results obtained for $n = 2$, 3, and 4, and sometimes only for $n = 3$ and 4, were evaluated in the Kohn−Sham basis set limit using a simple two-parameter fit.[53] For comparison, additional single-point calculations employing the aug-cc-pVTZ-J basis set[45,46] were performed. All the nonstandard basis sets were downloaded from the EMSL basis set library.[54] The convention used in earlier works,[10,11,55] for graphical purposes, was applied also in the current study: pcS-n and pcJ-n, where n = 0, 1, 2, 3, and 4, were set equivalent to Dunning's $X = 2$, 3, 4, 5, and 6 and plotted at $X = 4$, 5, and 6. Individual plots of shielding convergence and fittings toward the complete basis set limit are similar to those observed in our earlier works[10,11,55] and therefore are not shown in this work.

Theoretical NMR values obtained at equilibrium or experimental geometry should be compared with experimental results[32,56,57] after inclusion of zero-point vibrational (ZPVC) and thermal corrections (TC). The latter term, being an order of magnitude smaller, has been neglected. Ruden et al.[58] subtracted the correction term from the total observed coupling, J_{tot}^{exp}, arriving at the so-called "empirical equilibrium" coupling constant, J_{eq}^{emp}: $J_{eq}^{emp} = J_{tot}^{exp} - J_{vib}^{B3LYP}$. In our study, the empirical "experimental" value of nuclear isotropic shielding σ was compared directly with theoretical equilibrium coupling, σ_{eq}^{theor}, obtained from our ab initio calculations using different methods. Thus, we want to underline that, in this work, the CBS predicted theoretical nuclear shieldings calculated at experimental equilibrium geometry are compared with empirical equilibrium values, which include vibrational correction terms, obtained from separate calculations.

Vibrational averaging of NMR chemical shieldings in semirigid molecules can be based on the expansion of the nuclear potential V and the chemical shielding δ in Taylor series of the coordinates. In this study, the potential was expanded up to fourth powers of the normal mode coordinates Q_i as[59,60]

$$V = \frac{1}{2}\sum_{i=1}\omega_i^2 Q_i^2 + \frac{1}{6}\sum_{i=1}\sum_{j=1}\sum_{k=1} c_{ijk}Q_iQ_jQ_k + \frac{1}{24}\sum_{i=1}\sum_{j=1}\sum_{k=1}\sum_{l=1} d_{ijkl}Q_iQ_jQ_kQ_l \quad (1)$$

where the summations run over all modes i with harmonic frequencies ω_i. All cubic (c_{ijk}) and the semidiagonal (d_{iijk}) quartic constants were considered.

Similarly, the shieldings were expanded as

$$\delta = \delta_0 + \sum_i \delta_{1,i}Q_i + \frac{1}{2}\sum_{i,j}\delta_{2,ij}Q_iQ_j \quad (2)$$

where δ_1 and δ_2 are the first and second normal mode shielding derivatives, respectively. The vibrationally averaged rotations were obtained from a vibrational function Ψ as

$$\delta_{ave} = \langle\Psi|\delta|\Psi\rangle \quad (3)$$

The function ψ was obtained using the second-order degeneracy-corrected perturbational formula[59,60] from harmonic-oscillator functions, or using limited vibrational configuration interaction (VCI). As observed before,[59] these two wave function approximations gave almost the same results for the NMR shielding corrections.

The cubic and quartic force constants we obtained numerically from Hessians calculated analytically by Gaussian program,[39] for geometries displaced in normal modes. Likewise, the first and diagonal ($\delta_{2,ii}$) second shielding derivatives were calculated numerically by Gaussian. Program S4[61] interfaced to Gaussian was used for the anharmonic vibrational averaging. The vibrational contributions were assessed at the BHandH/pcS-2, BHandH/pcS-3, and MP2/pcS-3 levels.

III. Results and Discussion

In the first step of our studies, Jensen's basis sets without and with additionally augmented polarization functions (significantly larger) were used: pcS-n and aug-pcS-n and pcJ-n and aug-pcJ-n. The basis sets with all possible values of n were tested (0, 1, 2, 3, and 4) for dinitrogen, carbon oxide, and carbon dioxide. Executing all calculations poses a considerable computational effort, and therefore, it was possible to decrease the number of calculations while saving the main information obtained from the study. The shieldings obtained with pcS-n and pcJ-n basis sets, and n = 2, 3, and 4 were fitted using a two-parameter formula.[49] In several cases, the shieldings for n = 2 were slightly off the trend of the two last points (n = 3 and 4), and therefore, for consistency, all the results were uniformly fitted with the two last points only. Obviously, the two largest basis sets should be the most complete and flexible ones, and the obtained nuclear shieldings, being the second derivatives of total energy of the atomic system, should be the least corrupted ones by the accidental error cancellation.

The nitrogen shieldings in the Kohn–Sham basis set limit for extended, larger basis set hierarchies (aug-pcS-n and aug-pcJ-n) were practically identical with those obtained with the corresponding parent basis sets (pcS-n and pcJ-n). Thus, this extensive study was limited to the calculations with pcS-n and pcJ-n basis sets only, and for $n = 2$, 3 and 4, and with the relatively small aug-cc-pVTZ-J basis set, resulting in considerable time savings.

III.1. Vibrational Corrections to Nuclear Isotropic Shielding. Molecular response to the electromagnetic field also includes the nuclear contribution. We estimated the vibrational parts of chemical shielding at the BHandH/pcS-2 and pcS-3, as well as at the MP2/pcS-3 levels of theory. For benzene the MP2/pcS-3 level was too computationally demanding, and the "cheaper" basis set pcS-2 was used. The calculated vibrational corrections to the nuclear magnetic shieldings of the title compounds are summarized in Table S1 in the Supporting Information.

In most cases, MP2 provided larger absolute corrections than those obtained at the DFT level. The difference is most significant in systems with multiple bonds. The corrections calculated at the BHandH level for N_2, CO, CO_2, and NH_3, respectively, seem to be well converged, while the difference in pcS-2 and pcS-3 values for hydrocarbons indicates an incomplete convergence. However, the usage of a higher (pcS-4) basis set is practically impossible because of the long CPU time needed for the calculation. Nevertheless, the ZPV results obtained with the pcS-3 basis set are assumed sufficiently accurate considering the error caused by different theoretical levels and are used in our study to obtain empirical shieldings. In general, the vibrational corrections in Table S1 are similar to earlier reported values.[3,8,67,73,77,78] In some cases there are some discrepancies due to different electronic and vibrational theoretical approaches.

Figure 1 summarizes the vibrational changes of NMR shielding tensors of all studied compounds caused by the first and second property derivatives, calculated at the BHandH/pcS-3 and MP2/pcS-3 levels. In all cases, the inclusion of the vibrational corrections leads to smaller shielding tensor values. The contribution of the first shielding derivatives is approximately half (~1.4 ppm for BHandH and 1.5 ppm for MP2) of that caused by the second derivatives (2.5 ppm for BHandH and 3.0 ppm for MP2). However, the vibrational corrections do not improve the overall agreement with experimental results: the mean average deviation changes

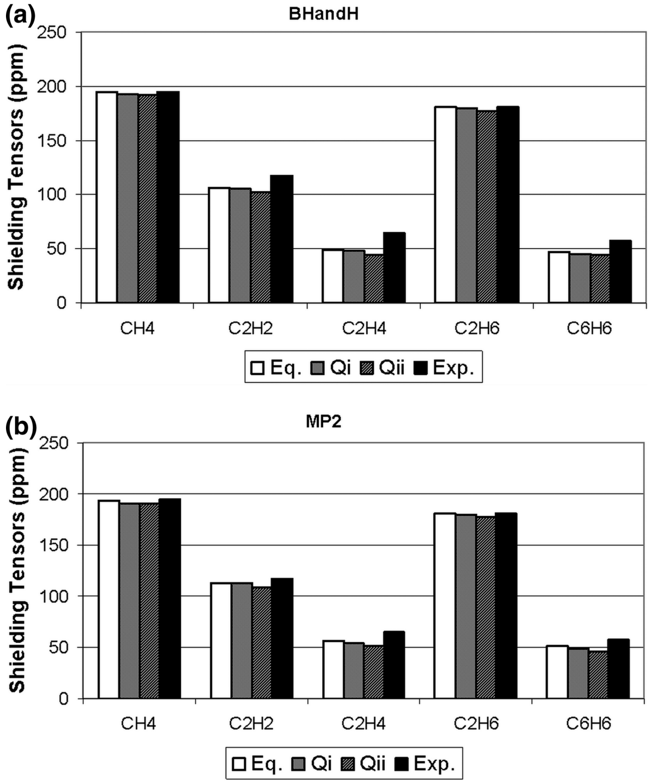


Figure 1. NMR shielding tensors (in ppm) calculated with the zero (Eq.), first (Qi), and second (Qii) shielding derivative corrections as compared to the experiment. NMR features were calculated at the (a) BHandH/pcS-3 and (b) MP2/pcS-3 (right) levels. Only in case of benzene was the MP2/pcS-3 level not available; thus the basis set used was pcS-2.

from 7.3 ppm (*Equilibrium*; BHandH) to 8.5 ppm (Q_i) or 11.0 ppm (Q_{ii}). For the MP2 method, the trend is the same, but the errors are lower than for DFT (4.1 ppm for *Equilibrium*, 5.4 ppm for Q_i, and 8.4 ppm for Q_{ii}). Similar observations have been already reported for NMR properties by Dračínský et al.,[59] and for the optical rotation, similar observations have been pointed out by Mort and Autschbach[62] and Kaminský et al.[60] The vibrational corrections thus seems important, but their contribution might be smaller than the error of the equilibrium values. Further improvement could be expected with higher electronic methods (such as coupled-cluster) and larger basis sets (e.g., pcS-4), which is, unfortunately, beyond our computational possibilities.

In order to better understand the role of individual vibrations in the averaging, in Figure 2, we plot the approximate contributions of individual modes defined as $\delta_{1,i}\langle Q_i\rangle + 1/2\delta_{2,ii}\langle Q_i^2\rangle$ (cf. eq 2) for methane. A similar analysis for acetylene, ethane, and benzene is shown in the Supporting Information (Figures S1A−S1D). As apparent from Figures 2 and S1A−S1D, most of the harmonic normal modes significantly contribute to the nuclear magnetic shielding.

The contributions of the lowest-energy mode in C_2H_6 cannot be considered reliable, as this mode (methyl rotation) exhibits a strongly anharmonic potential, for which the limited Taylor expansion (eq 1) is probably inappropriate.

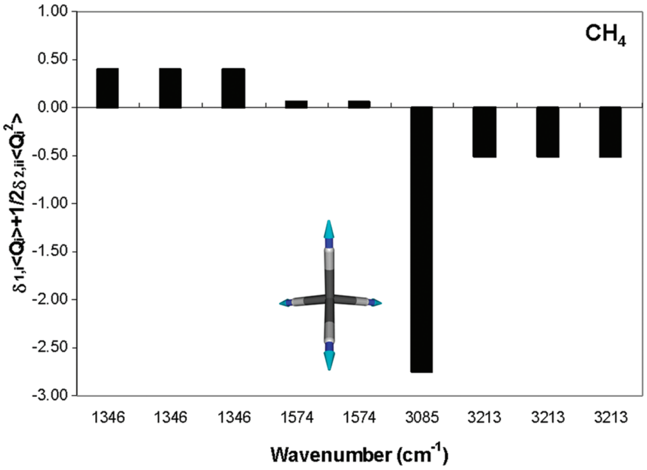


Figure 2. Contribution of individual normal modes to the magnetic shielding vibrational correction in methane. Picture represents the most contributive vibration.

The largest contributions come from the C−H stretch (for CH_4 and C_2H_6 at 3085 cm^{-1} and 3096 cm^{-1}, respectively), the CH bending (C_2H_2, 803 cm^{-1}), C=C stretching (C_2H_4, 1754 cm^{-1}), and benzene symmetric ring breathing (C−C stretch, 1063 cm^{-1}). Note that the four lowest modes in C_2H_2 are double-degenerated. The potential energy of such a linear symmetric molecule thus could be given, for example, as[63]

$$2V = k_1 Q_{23}^2 + k_2(Q_{12}^2 + Q_{34}^2) + k_\delta(\delta_{13}^2 + \delta_{24}^2) \qquad (4)$$

where δ_{13} is the deviation of the angle between atoms H1, C2, and C3 from 180° and δ_{13} is the corresponding deviation for C2−C3−H4. However, no simple judgment to predict the biggest contributions comes to our mind, and a complete estimation of the corrections for all the modes seems the only option.

III.2. Convergence of Nuclear Isotropic Shielding in N_2, CO, CO_2, and NH_3. N_2 nuclear isotropic shielding predicted in the CBS limit using pcS-n and pcJ-n basis set families and the single point aug-cc-pVTZ-J results were calculated at several theoretical levels and are gathered in Table S2 in the Supporting Information. The method numbers 1−7 refer to “pure” and 8−38 to “hybrid” density functionals and as such will be applied to all calculated results in the subsequent tables and figures. Analogous shielding data in the CBS limit for all studied compounds are contained in the Supporting Information (Tables S3−S8). The WP04 density functional (method No. 39 in Table S2) was recently designed[64,65] for better prediction of proton shieldings and executed in the Gaussian 09 program as a modification of the BLYP functional with IOp entries (see refs 64 and 65). Method numbers 40 and 41 (RHF and MP2), 42−44 (KT*n*), and 45−47 (SOPPA, SOPPA(CCSD), and CCSD(T)) close the list, being a kind of reference tool.

Obtained data are compared with experimental nitrogen shielding[2,66] and the estimated “empirical shielding”, containing the BHandH/pcS-3 calculated ZPV correction from Table S1. To distinguish the performance of the individual method, the deviations of calculated results from the empirical nitrogen shielding are plotted in Figure 3 (the methods

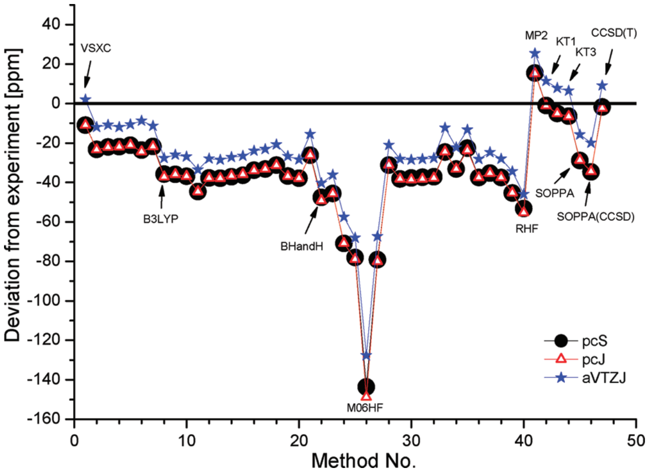


Figure 3. CBS estimated (pcS-n and pcJ-n) and SP aug-cc-pVTZ-J deviations of nitrogen isotropic shieldings from empirical values in N_2 (method numbers are listed in Table S2).

given on the horizontal axis are selected according to their order of appearance in Table S2).

Some methods, producing very poor (deviation of about −150 ppm for M06HF) or very good results (VSXC, KT1, KT2, and KT3 with deviations −11, −1, −5, and −6 ppm) as well as the popular B3LYP functional (deviation of about −37 ppm) are directly indicated in the plot. It is apparent from Figure 3 that both pcS-n and pcJ-n basis set hierarchies perform practically identically, and the corresponding results obtained with a significantly smaller basis set, aug-cc-pVTZ-J, are slightly closer (by about 10 ppm) to experimental results. Nitrogen shieldings predicted with RHF and BHandH methods (dev. −55 and −50 ppm) are worse than the B3LYP value. Moreover, the corresponding MP2 and CCSD(T) calculations produce smaller deviations from experimental results (+15 and −2 ppm). In general, the majority of density functionals underestimate N_2 isotropic nuclear shielding by −20 to −40 ppm. SOPPA and SOPPA(CCSD) results significantly deviate from experimental results (−30 to −35 ppm). The excellent predicting power of the CCSD(T) benchmark method is not surprising. In addition, the very good performance of KTn density functionals is encouraging, and fairly good results obtained with "pure" density functionals (method numbers 1−7) are remarkable.

In the Supporting Information (Figures S2A,B) are shown carbon and oxygen nuclear magnetic shielding deviations of CO from experimental results obtained with pcS-n, pcJ-n, and aug-cc-pVTZ-J basis sets. It is obvious that general trends, reflecting the performance of different density functionals and basis sets, are very similar to those observed for N_2 (Figure 3). Thus, the results produced with both of Jensen's basis sets are practically identical, and shieldings obtained with the compact aug-cc-pVTZ-J basis set are about 10 ppm closer to experimental values.[3,8,67,68] The majority of density functionals underestimate experimental carbon shieldings by 15 to 25 ppm (Figure S2A) and oxygen shieldings by 10 to 30 ppm (Figure S2B). The B3LYP and BHandH performance is similar to the majority of density functionals, and VSXC, KTn's, and MP2 reproduce experimental results significantly better than the remaining methods. The M06HF produces the worst result (carbon and oxygen deviations are about −80 and −185 ppm).

In the case of CO_2, very similar deviations to those of carbon monoxide patterns of shielding from the experiment[3,8,68] are observed. The corresponding graphs are placed in the Supporting Information (Figure S3A,B). The worst results are again observed for the M06HF density functional with carbon and oxygen shielding deviations of −23 and −44 ppm, respectively.

The general pattern of the studied method's performance is similar to the case of ammonia nitrogen shielding deviations from the experiment[2,69,70] (Figure S4A); the tHCTH, M06L, M06, and WP04 density functionals show the worst results (−15, −17, −27, −23 ppm), and the best performance is observed for VSXC, BHandH, BMK, and wB97 density functionals (−4, −4, −3, −5 ppm). As expected, proton shieldings of ammonia deviate from experimental results[70] less than the nitrogen ones (compare Figure S4A and B). Surprisingly, a very large deviation of the ammonia proton shielding from the empirical value is predicted with the BHandH and CCSD(T) methods (−0.6 and −0.4 ppm). The last result is difficult to explain taking into account the excellent performance of the CCSD(T) method in predicting nuclear shieldings of small molecules. On the other hand, we notice that accurate CCSD(T)/pz3d2f calculations reported by Gauss et al.[2] resulted in a similar value of this deviation (0.3 ppm) compared to the empirical value.

III.3. Convergence of Nuclear Isotropic Shielding in CH_4, C_2H_2, C_2H_4, C_2H_6, and C_6H_6. The carbon and proton nuclear isotropic shielding deviations from experimental values of methane[2,3,70] are shown in Figure 4.

Both the pcS-n and pcJ-n basis set hierarchies perform practically identically, and the corresponding results obtained using a significantly smaller aug-cc-pVTZ-J basis set give slightly better values.

In general, the majority of density functional methods underestimate the carbon shielding in CH_4 by −5 to −10 ppm. Some methods, producing very poor or good results in comparison to the popular B3LYP density functional, are again directly indicated in the plot. Carbon nuclear shieldings in methane predicted by the VSXC, BHandH, wB97, and MP2 methods are better than using the B3LYP. Corresponding deviations from the experiment are −4, −2.5, −2, 2.5, and −10 ppm, respectively. The worst methods (M06 and WP04) predict methane carbon shieldings significantly deviating from experimental results (−18 and −20 ppm).

In general, the pcS-n and pcJ-n families of basis sets overestimate the experimental proton nuclear shieldings of methane by 0.2−0.35 ppm. With the VSXC, BHandH, and MP2 methods, these basis sets provide the lowest deviations of about 0.15, −0.04, and 0.02. In addition, a compact aug-cc-pVTZ-J basis set predicts proton shieldings worse by about 0.10−0.15 ppm than those obtained in the complete basis set limit. Similarly to the previous compounds, the largest deviations (0.55 and 0.45 ppm) are observed for M06HF and WP04.

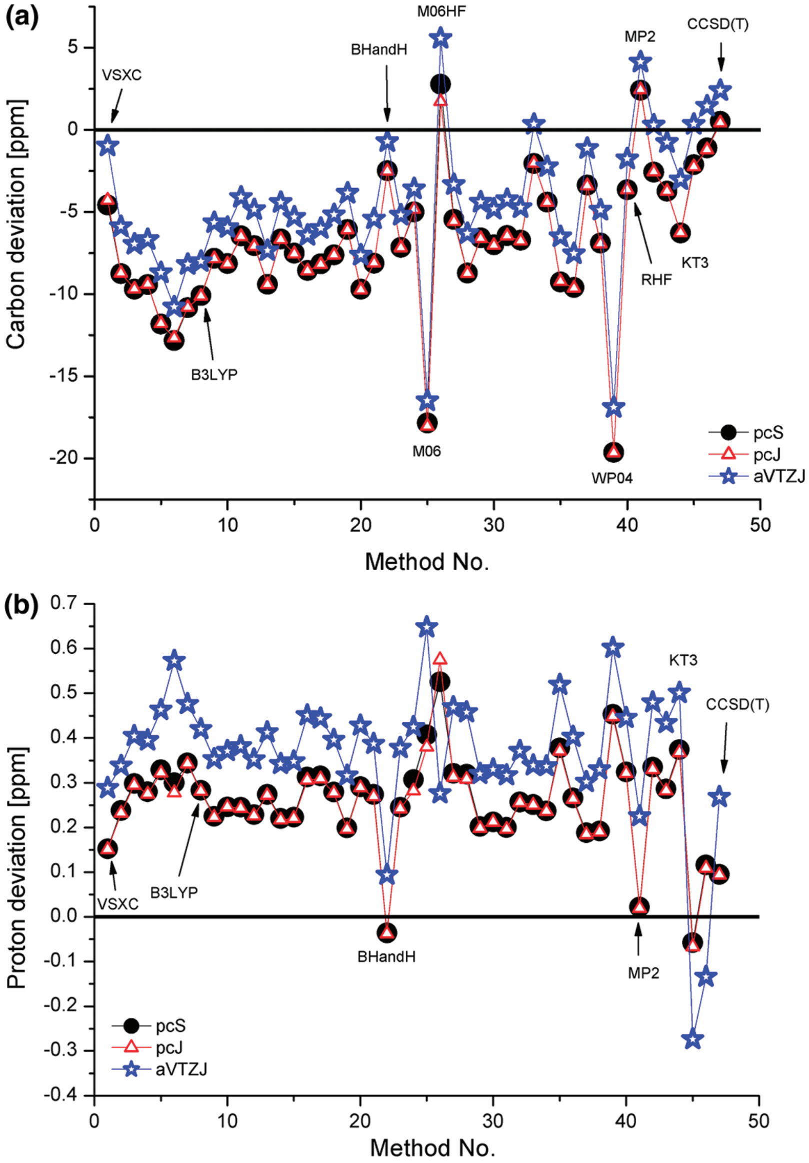


Figure 4. CBS estimated (pcS-n and pcJ-n) and SP aug-cc-pVTZ-J deviations of (a) carbon and (b) proton isotropic shieldings from empirical values in CH_4 (method numbers are listed in Table S2).

The majority of methods underestimate carbon shielding[71] in acetylene by −10 to −15 ppm (Figure S5A, Supporting Information), and the best performing methods are VSXC and MP2, deviating −5 and 2 ppm. On the other hand, the largest deviation is observed for M06HF (−33 ppm).

Acetylene proton shieldings[71,72] (Figure S5B) are overestimated by 0.1 to 0.2 ppm, and the best results are observed for VSXC, LC-wPBE, and RHF (dev. −0.09, 0.03, and 0.01 ppm). The worst performances are observed for BHandH and MP2 (dev. −0.27 and −0.26 ppm) and WP04 and M06 (dev. 0.48 and 0.57 ppm).

Carbon nuclear shieldings in ethylene (Figure S6A, Supporting Information) deviate from the experimental value (64.4 ppm reported by Auer and co-workers[3] and augmented with calculated ZPV correction) by about −15 to −25 ppm, and the best methods are VSXC, RHF, and MP2 (dev. −12, −11, and −2 ppm). The performance of B3LYP and BHandH (dev. −25 and −23 ppm) is similar to the majority of density functionals, and the worst case is observed for M06HF (dev. −58 ppm). The KTn methods underestimate the experiment by about 8 ppm.

Proton shieldings in ethylene (25.43 ppm,[72] see Figure S6B) are underestimated by −0.1 to −0.4 ppm, while the best method reproduces experimental results very well (−0.01 ppm deviation for TPSSh). M06HF, BHandH, VSXC, B3LYP, MP2, and RHF deviate from experimental results by −0.9, −0.7, −0.5, −0.2, −0.16, and 0.12 ppm.

Carbon nuclear shieldings in ethane (Figure S7A, Supporting Information) deviate from the experimental value (180.8 ppm by Auer and co-workers[3] and augmented with calculated ZPV correction) by about −5 to −15 ppm, and the best results are for MP2, RHF, wB97, M06HF, and BHandH (dev. 1, −3, −3, −3, and −5 ppm). B3LYP deviates by −14 ppm, and the M06, WP04, and BHandH density functionals are significantly worse (dev. −19, −22, and −25 ppm).

Experimental proton shieldings in ethane (experimental value (29.86 ppm) reported by Chesnut[73] and corrected with

Table 1. C_6H_6 Isotropic Shieldings Calculated Using a Few Selected Methods and Basis Sets Compared with Experimental Results before and after Inclusion of ZPV Correction

basis	method			
C shielding	RHF	B3LYP	BHandH	CCSD(T)
6-311G**	**59.930**	**51.330**	**52.604**	**71.559**
pcS-0	51.747	50.485	50.734	74.144
pcS-1	57.431	47.320	50.043	68.759
PcS-2	53.023	42.071	43.975	62.866
PcS-3	53.196	41.732	43.983	
PcS-4	53.222	41.673	44.002	
CBS	**53.258**	**41.591**	**44.028**	
pcJ-0	59.028	50.873	52.905	
pcJ-1	57.157	45.521	48.487	
PcJ-2	53.834	42.043	44.604	
PcJ-3	53.207	41.754	43.990	
PcJ-4	53.228	41.681	44.008	
CBS	**53.258**	**41.580**	**44.032**	
aVTZJ	57.929	47.020	48.800	
exp.		**57.105 ± 0.009**[a]		
emp.		**59.905**[b]		

H shielding	RHF	B3LYP	BHandH	CCSD(T)
6-311G**	24.675	24.568	23.994	24.765
pcS-0	25.502	25.487	24.899	26.002
pcS-1	24.364	24.244	23.707	24.481
PcS-2	24.222	24.020	23.468	24.105
PcS-3	24.196	23.984	23.434	
PcS-4	24.194	23.982	23.428	
CBS	**24.191**	**23.980**	**23.419**	
pcJ-0	25.672	25.703	25.080	
pcJ-1	24.379	24.271	23.706	
PcJ-2	24.236	24.046	23.494	
PcJ-3	24.198	23.989	23.435	
PcJ-4	24.193	23.983	23.427	
CBS	**24.186**	**23.975**	**23.415**	
aVTZJ	24.344	24.146	23.581	
exp.		**23.60**[c]		
emp.		**23.90**[b]		

[a] From ref 74. [b] Including ZPV corrections for C and H of −2.8 and −0.3 ppm (see Table S1). [c] From refs 75 and 76.

calculated ZPV contribution, see Figure S7B) are very well reproduced (dev. −0.1 to 0.1 ppm for most cases), and the least accurate results produce BHandH, MP2, RHF, WP04, and M06HF (dev. −0.25, −0.13, 0.33, 0.33, and 0.42 ppm).

Benzene is the largest molecule in the set of studied hydrocarbons, and DFT calculations of its shieldings using the pcS-4 basis set are very lengthy. Carbon isotropic shielding (57.105 ± 0.009 ppm) of the isolated benzene molecule in xenon gas was reported by Jackowski and coworkers.[74] Thus, it is a real challenge to get accurate shieldings of C_6H_6 at the level of CCSD(T) or to dream about estimating such results in the basis set limit. Therefore, we limited our CBS studies to a few methods only—RHF and BHandH with pcS-n and pcJ-n basis sets. In addition, we compared the CBS values obtained with pcS-n and pcJ-n basis sets using the popular B3LYP hybrid functional (Table 1). The corresponding results obtained with aug-pcS-n and aug-pcJ-n basis sets were almost identical (not shown in Table 1).

The RHF and DFT CBS fitted shieldings we compare directly with the CCSD(T) results with smaller basis sets pcS-n (n = 0, 1, and 2 only) and a relatively small Pople type basis set 6-311G**.

At first, we notice that RHF and CCSD(T) carbon and proton shieldings of benzene obtained with smaller basis sets (6-311**, pcS-0 and pcJ-0, pcS-1 and pcJ-1) deviate significantly from results obtained using larger basis sets (n = 2, and in some cases n = 3 and 4). Another important observation from Table 1 is that the CBS estimated B3LYP shieldings are practically identical for pcS-n and pcJ-n basis set families. It is also apparent from Table 1 that, by using a small basis set (6-311G**) and hoping for accidental error cancellation, we may obtain a perfect agreement of the theoretical result with experimental results (compare RHF and CCSD(T) carbon shielding of 59.930 and 71.559 ppm with an empirical value of 59.905 ppm). In such a drastic case, one could wrongly conclude that in practice it is enough to use a fast RHF calculation with a small basis set to confirm experimental data. On the other hand, this also shows that using a much elaborated method (for example, the coupled cluster wave function) with a deficient basis set may produce poor results.

The limited results from Table 1 confirm earlier reports that RHF is somehow more reliable in predicting carbon shieldings than DFT. The latter one tends to predict significantly lower heavy atom shieldings due to overestimation of their paramagnetic terms. The CBS predicted carbon shieldings of benzene calculated with RHF, BHandH, and B3LYP are 53.26, 44.03, and 41.6 ppm and are poor estimates of an empirical value (59.91 ppm). To the contrary, the corresponding CBS proton shieldings (24.2, 23.4, and 24.0 ppm for RHF, BHandH, and B3LYP) are significantly closer to experimental results (23.60 ppm[75,76] augmented with ZPV).

In spite of a limited number of studied systems, we would like to show some statistical data showing the general trends in performance of individual methods (due to very limited available calculation results, benzene is excluded from this analysis). Averaged nuclear shielding deviations from experimental results somehow mask the real performance of the methods (see Figures S8A and S8B). Hence, in Figure 5, the root-mean square (RMS) deviations of nuclear shieldings from experimental results are presented. In general, RMS deviations of 10 heavy atoms (Figure 5A) are between 10 and 20 ppm, and only a few advanced methods (MP2, KTn and CCSD(T)) produce better results. On the other hand, due to the large deviations of nuclear shieldings calculated using the Minnesota density functionals from experimental results, these methods are not recommended for NMR calculations in the studied molecules and in similar atomic systems. The performance of the pcS-*n* basis set is similar to pcJ-*n* hierarchy, and therefore, the first ones are computationally more accessible. Furthermore, the inexpensive aug-cc-pVTZ-J basis set generally produces slightly more accurate results at the DFT level of calculations (exceptions are observed for MP2, KTn, and CCSD(T)). RMS deviations of proton nuclear shieldings in the studied systems are about 0.2 ppm, and the aug-cc-pVTZ-J basis set produces slightly worse results (see Figure 5B).

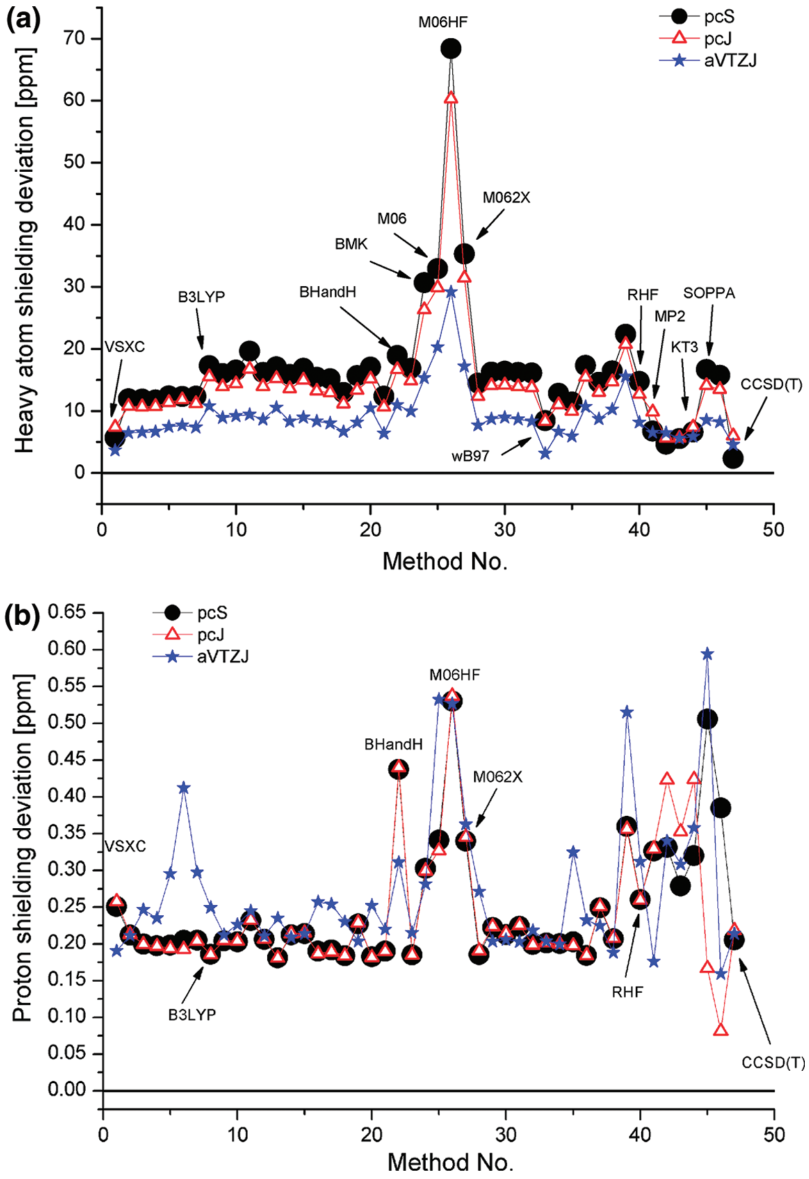


Figure 5. CBS estimated (pcS-n and pcJ-n) and SP aug-cc-pVTZ-J RMS deviations of (a) heavy atoms and (b) proton isotropic shieldings from empirical values in the studied systems. Results for benzene are excluded due to incompleteness of theoretical data, and method numbers are listed in Table S2.

IV. Conclusions

The performances of several density functionals were tested for predicting isotropic nuclear shieldings of nine small molecules (N_2, CO, CO_2, NH_3, CH_4, C_2H_2, C_2H_4, C_2H_6, and C_6H_6) using pcS-n and pcJ-n basis set hierarchies. The DFT nuclear shieldings estimated in the complete basis set limit were compared with empirical shieldings obtained from experimental values and calculated ZPV corrections and RHF, MP2, SOPPA, SOPPA(CCSD), and CCSD(T) results. RMS deviations of 10 heavy atoms reproduce experimental results by about ±10−20 ppm. Better results are obtained using MP2, KTn, and CCSD(T) methods. Jensen's pcS-n and pcJ-n basis sets work similarly; thus, the first ones are recommended. The inexpensive aug-cc-pVTZ-J basis set produces very accurate results. The studied DFT methods calculate proton isotropic shieldings with RMS deviations of about 0.2 ppm, and the aug-cc-pVTZ-J basis set gives slightly worse results. Surprisingly, the "pure" density functionals produce fairly accurate NMR shieldings, better than the popular B3LYP. The Minnesota density functionals are not suitable for shielding calculations of the selected molecules (and probably for similar molecules). However, it should be noted that our series of compounds is limited in structural diversity, and the general applicability of conclusions made in this work need to be verified in the future.

Acknowledgment. This work was supported by the Faculty of Chemistry, University of Opole (Grant 10/ICH/2009-S), and the calculation facilities and software in the Supercomputing and Networking Center ACK CYFRONET AGH in Krakow within a calculation grant MNiSW/SGI3700/UOpolski/061/2008 and calculation facilities and software at the Supercomputing and Networking Center in Wrocław. Michał Jaszuński, Piotr Lodowski, and Frank Jensen are thanked for helpful discussions.

Supporting Information Available: Detailed information about CBS estimated nuclear shieldings for all studied molecules using the pcS-n, pcJ-n, and aug-cc-pVTZ-J basis sets, as well as calculated vibrational corrections at BHandH and MP2 levels. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **1995**, *103*, 3561–3577.

(2) Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **1996**, *104*, 2574–2583.

(3) Auer, A.; Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **2003**, *118*, 10407–10417.

(4) Gauss, J. *Chem. Phys. Lett.* **1992**, *191*, 614–620.

(5) Gauss, J. *J. Chem. Phys.* **1993**, *99*, 3629–3643.

(6) Gauss, J. *J. Chem. Phys.* **2002**, *116*, 4773–4776.

(7) Harding, M. E.; Lenhart, M.; Auer, A. A.; Gauss, J. *J. Chem. Phys.* **2008**, *128*, 244111–10.

(8) Auer, A. *J. Chem. Phys.* **2009**, *131*, 024116–7.

(9) Feller, D. *J. Chem. Phys.* **1992**, *96*, 6104–6114.

(10) Kupka, T.; Ruscic, B.; Botto, R. E. *J. Phys. Chem. A.* **2002**, *106*, 10396–10407.

(11) Kupka, T.; Lim, C. *J. Phys. Chem. A* **2007**, *111*, 1927–1932.

(12) Kupka, T. *Magn. Reson. Chem.* **2009**, *47*, 959–970.

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

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

(15) Wilson, A.; van Mourik, T.; Dunning, T. H. *THEOCHEM* **1997**, *388*, 339–349.

(16) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

(17) Puzzarini, C. *J. Phys. Chem. B* **2009**, *113*, 14530–14535.

(18) Jensen, F. *J. Chem. Phys.* **1999**, *110*, 6601–6605.

(19) Jensen, F. *J. Chem. Phys.* **2001**, *115*, 9113–9125.

(20) Jensen, F. *J. Chem. Phys.* **2002**, *116*, 7372–7379.

(21) Jensen, F. *J. Chem. Phys.* **2003**, *118*, 2459–2463.

(22) Jensen, F.; Helgaker, T. *J. Chem. Phys.* **2004**, *121*, 3463–3470.

(23) Jensen, F. *Chem. Phys. Lett.* **2005**, *402*, 510–513.

(24) Jensen, F. *J. Chem. Theory Comput* **2008**, *4*, 719–727.

(25) Jorge, F. E.; Sagrillo, P. S.; de Oliveira, A. R. *Chem. Phys. Lett.* **2006**, *432*, 558–563.

(26) Canal Neto, A.; Muniz, E. P.; Centoducatte, R.; Jorge, F. E. *THEOCHEM* **2005**, *718*, 219–224.

(27) Barbieri, P. L.; Fantin, P. A.; Jorge, F. E. *Mol. Phys.* **2006**, *104*, 2945–2954.

(28) Foresman, J. B.; Frisch, A. *Exploring Chemistry with Electronic Structure Methods*, 2nd ed.; Gaussian Inc: Pittsburg, PA, 1996.

(29) Waller, M. P.; Robertazzi, A.; Platts, J. A.; Hibbs, D. E.; Williams, P. A. *J. Comput. Chem.* **2006**, *27*, 267–274.

(30) Dkhissi, A.; Ducéré, J. M.; Blossey, R.; Pouchan, C. *J. Comput. Chem.* **2008**, *30*, 1179–1184.

(31) Kupka, T. *Magn. Reson. Chem.* **2009**, *47*, 210–221.

(32) Jackowski, K. *J. Mol. Struct.* **2006**, *786*, 215–219.

(33) Raynes, W. T. *Mol. Phys.* **1983**, *49*, 443–447.

(34) Wasylishen, R. E.; Bryce, D. L. *J. Chem. Phys.* **2002**, *117*, 10061–10066.

(35) Wasylishen, R. E.; Mooibroek, S.; Macdonald, J. B. *J. Chem. Phys.* **1984**, *81*, 1057–1059.

(36) Puzzarini, C.; Cazzoli, G.; Harding, M. E.; Vázquez, J.; Gauss, J. *J. Chem. Phys.* **2009**, *131*, 234304–11.

(37) Keal, T. W.; Tozer, D. J.; Helgaker, T. *Chem. Phys. Lett.* **2004**, *391*, 374–379.

(38) Helgaker, T.; Jaszunski, M.; Ruud, K. *Chem. Rev.* **1999**, *99*, 293–352.

(39) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J. *Gaussian 09*, revision A.02; Gaussian, Inc.: Wallingford, CT, 2009.

(40) Teale, A. M.; Tozer, D. J. *Chem. Phys. Lett.* **2004**, *383*, 109–114.

(41) Kongstead, J.; Aidas, K.; Mikkelsen, K. V.; Sauer, S. P. A. *J. Chem. Theory Comput.* **2008**, *4*, 267–277.

(42) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2004**, *121*, 5654–5560.

(43) Olsen, J.; Jorgensen, P. *J. Chem. Phys.* **1985**, *82*, 3235–3264.

(44) Jensen, F. *J. Chem. Theory Comput.* **2006**, *2*, 1360–1369.

(45) Provasi, P. F.; Aucar, G. A.; Sauer, S. P. A. *J. Chem. Phys.* **2001**, *115*, 1324–1334.

(46) Peralta, J. E.; Scuseria, G. E.; Cheeseman, J. R.; Frisch, M. J. *Chem. Phys. Lett.* **2003**, *375*, 452–458.

(47) Maximoff, S. N.; Peralta, J. E.; Barone, V.; Scuseria, G. E. *J. Chem. Theory Comput.* **2005**, *1*, 541–545.

(48) Stanton, J. F.; Gauss, J.; Watts, J. D.; Lauderdale, W. J.; Bartlett, R. J. *Int. J. Quantum Chem.* **1992**, *44*, 879–894.

(49) Bak, K. L.; Gauss, J.; Jorgensen, P.; Olsen, J.; Helgaker, T.; Stanton, J. F. *J. Chem. Phys.* **2001**, *114*, 6548–6556.

(50) London, F. *J. Phys. Radium (Paris)* **1937**, *8*, 397–409.

(51) Wolinski, K.; Hinton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251–8260.

(52) Kutzelnigg, W.; Fleischer, U.; Schindler, M. *NMR Basic Principles and Progress*; Springer-Verlag: Berlin, 1990; Vol. 23.

(53) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.

(54) Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. *J. Chem. Inf. Model.* **2007**, *47*, 1045–1052.

(55) Kupka, T. *Chem. Phys. Lett.* **2008**, *461*, 33–37.

(56) Jackowski, K. *Int. J. Mol. Sci.* **2003**, *4*, 135–142.

(57) Ruud, K.; Astrand, P.-O.; Taylor, P. R. *J. Chem. Phys.* **2000**, *112*, 2668–2683.

(58) Ruden, T. A.; Lutnaes, O. B.; Helgaker, T.; Ruud, K. *J. Chem. Phys.* **2003**, *118*, 9572–9581.

(59) Dračínský, M.; Kaminský, J.; Bouř, P. *J. Chem. Phys.* **2009**, *130*, 094106–13.

(60) Kaminský, J.; Raich, I.; Tomčáková, K.; Bouř, P. *J. Comput. Chem.* **2010**, [Online] DOI: 10.1002/jcc.21511.

(61) Bouř, P. *Program S4*; Czech Academy of Sciences: Prague, 1994−2009.

(62) Mort, B. C.; Autschbach, J. *J. Phys. Chem. A* **2005**, *109*, 8617–8623.

(63) Herzberg, G. *Molecular Spectra and Molecular Structure. II. Infrared and Raman Spectra of Polyatomic Molecules*; Krieberg Publishing Company: Malabar, FL, 1945; p 181.

(64) Wiitala, K. W.; Hoye, T. R.; Cramer, C. J. *J. Chem. Theory Comput.* **2006**, *2*, 1085–1092.

(65) Jain, P.; Bally, T.; Rablen, P. R. *J Chem Theory Comput.* **2009**, *74*, 4017–4023.

(66) Jameson, C. J.; Jameson, A. K.; Oppusunggu, D.; Wille, S.; Burrel, P. M.; Mason, J. *J. Chem. Phys.* **1981**, *74*, 81–88.

(67) Sundholm, D.; Gauss, J.; Schafer, A. *J. Chem. Phys.* **1996**, *105*, 11051–11059.

(68) Makulski, W.; Jackowski, K. *J. Mol. Struct.* **2003**, *651−653*, 265–269.

(69) Kukolich, S. G. *J. Am. Chem. Soc.* **1975**, *97*, 5704–5707.

(70) Raynes, W. T. in Harris, R. K. *Nuclear Magnetic Resononance*; The Chemical Society: London, 1977; p 1.

(71) Jackowski, K.; Wilczek, M.; Pecul, M.; Sadlej, J. *J. Phys. Chem. A* **2000**, *104*, 5955–5958.

(72) Schneider, W. G.; Bernstein, H. J.; Pople, J. A. *J. Chem. Phys.* **1958**, *28*, 601–607.

(73) Chesnut, D. B. *Chem. Phys.* **1997**, *214*, 73–79.

(74) Jackowski, K.; Maciaga, E.; Wilczek, M. *J. Mol. Struct.* **2005**, *744−747*, 101–105.

(75) Katritzky, A. R. *Handbook of Heterocyclic Chemistry*; Pergamon Press: Oxford, 1985.

(76) Salsbury, F. R. J.; Harris, R. A. *Chem. Phys. Lett.* **1997**, *279*, 247–251.

(77) Dransfield, A. *Chem. Phys.* **2004**, *298*, 47–53.

(78) Ruud, K.; Astrand, P.-O.; Taylor, P. R. *J. Am. Chem. Soc.* **2001**, *123*, 4826–4833.

CT100109J

JCTC Journal of Chemical Theory and Computation

The Use of Anisotropic Potentials in Modeling Water and Free Energies of Hydration

Panagiotis G. Karamertzanis,*,[†] Paolo Raiteri,[‡] and Amparo Galindo[†]

Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, London, SW7 2AZ, United Kingdom, and Department of Chemistry and Nanochemistry Research Institute, GPO Box U1987, 6845 Perth, Western Australia

Received December 26, 2009

Abstract: We propose a novel, anisotropic rigid-body intermolecular potential model to predict the properties of water and the hydration free energies of neutral organic solutes. The electrostatic interactions of water and the solutes are modeled using atomic multipole moments up to hexadecapole; these are obtained from distributed multipole analysis of the quantum mechanically computed charge densities and include average polarization effects in solution. The repulsion−dispersion water−water interactions are modeled with a three-site, exp-6 model fitted to the experimental liquid water density and oxygen−oxygen radial distribution function at ambient conditions. The proposed water model reproduces well several water properties not used in its parametrization, including vapor−liquid coexistence densities, the maximum in liquid water density at atmospheric pressure, the structure of ordered ice polymorphs, and the liquid water heat capacity. The model is used to compute the hydration free energy of 10 neutral organic solutes using explicit-solvent free energy perturbation. The solute−solute repulsion−dispersion intermolecular potential is obtained from previous parametrizations on organic crystal structures. In order to calculate the free energies of hydration, water−solute repulsion−dispersion interactions are modeled using Lorenz−Berthelot combining rules. The root-mean-square error of the predicted hydration free energies is 1.5 kcal mol^{-1}, which is comparable to the error found using a continuum mean-field quantum mechanical approach parametrized using experimental free energy of hydration data. The results are also contrasted with explicit-solvent hydration free energies obtained with an atomic charge representation of the solute's charge density computed at the same level of theory used to compute the distributed multipoles. Replacing the multipole description of the solute's charge density with an atomic charge model changes the free energy of hydration by as much as 3 kcal mol^{-1} and provides an estimate for the effect of the modeling quality of the intermolecular electrostatic forces in free energy of solvation calculations.

Introduction

Computational chemistry techniques are often used to model protein folding, ligand recognition and binding, partition coefficients, solubility, reaction rates, pK_a, and tautomer ratios, all of which depend crucially on the accuracy with which solvation effects can be modeled. Water is the most important and widely studied solvent due to its ubiquity in biological and industrial processes.[1] Many water models, differing in the way they treat intramolecular distortions, electrostatic interactions,[2] polarization effects,[3−5] and their parametrization strategy using experimental[6,7] or *ab initio*[8] data, have been developed. Although these water models are

* Corresponding author e-mail: p.karamertzanis@imperial.ac.uk.

[†] Imperial College London.

[‡] Department of Chemistry and Nanochemistry Research Institute.

10.1021/ct900693q © 2010 American Chemical Society
Published on Web 04/02/2010

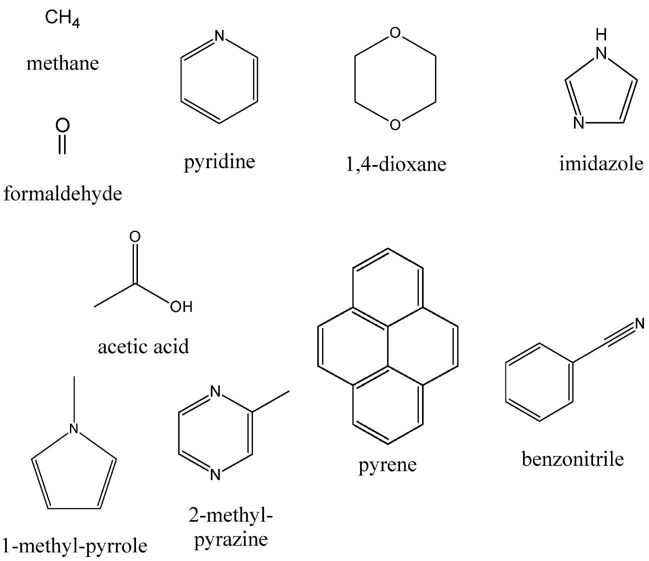
successful in predicting selected liquid water and ice polymorph properties, it is still debated what the indispensable aspects of an accurate water model are. This information is needed to construct a model of general applicability that is transferable and of tractable complexity to be of practical use in computationally intensive calculations, such as the hydration of biomolecules and modeling of nucleation.

The prediction of the free energy of hydration is a prime example of the use of water models and a stringent test of their accuracy. The free energy of hydration is associated with the tendency of the solute to leave the aqueous solution. Molecules with strong hydrogen bond donor and acceptor groups have a strongly negative ΔG_{hyd} (hydrophilic), while molecules that are poorly attracted to water (hydrophobic) have a positive ΔG_{hyd}, as the stabilization due to water−solute interactions is not sufficiently large to compensate the disruption of the energetically more favorable water−water interactions. The free energy of hydration can be experimentally determined by measuring the equilibrium constant for the solute transfer between vapor and aqueous solution, under experimental conditions and concentrations that eliminate solute self-association in both phases.[9] However, its computational prediction is of interest, as it can be used to compute partition coefficients[10] and solubility,[11,12] which are key quantities in pharmaceutical development. A wide range of methods to compute hydration free energies has been proposed.[13] The methods differ markedly in computational cost and extent of parametrization and include group additivity schemes,[14] continuum solvation models,[15] and explicit-solvent, free energy approaches[16] based on exhaustive sampling of all thermally accessible states using molecular dynamics or Monte Carlo simulations.

Explicit-solvent free energy methods have the potential to be systematically improved by minimizing the sources of statistical and systematic error until quantitative predictions of hydration free energies are obtained. Good progress has been achieved in designing better free energy methods and protocols[16−19] that minimize the error due to finite sampling. However, errors due to approximations in the intermolecular potential model have received less attention, although the accuracy of explicit-solvent hydration free energy calculations has been shown to be strongly dependent on the model for the water−water interactions[20,21] and the model for the solute's intermolecular electrostatic interactions.[22,23] Despite the known limitations in describing aqueous solutions, most water models used in hydration free energy calculations employ a point-charge approximation. Alternative representations include the use of smeared (Gaussian) charges[2,24,25] and multipoles,[3,26,27] which improve the representation of water's charge density, but their use in modeling hydration has thus far been limited.[28,29]

The modeling of the organic solid state has provided an impetus for developing accurate models for the intermolecular forces, with particular emphasis on the electrostatic[30] and induction contributions.[31] A key element of these models is the description of the intermolecular electrostatic forces with a distributed multipole model, which has been shown to improve accuracy when features such as lone-pair and π-electron densities[32,33] are present. These models have been used successfully in quantifying the small energy differences between polymorphs[33] and predicting bulk crystal properties[34,35] but have not been used to study the interaction of the crystal with the solvent in the context of nucleation and growth or solubility predictions, which would also require the development of an accurate model for the solvent.

Chart 1. Molecules Used in Hydration Free Energy Calculations



The objective of this work is to construct a high-rank multipole model for water and to evaluate its applicability in dynamic simulations that sample the whole range of water−water and solute−water molecular configurations. The model for the electrostatic forces for both water and the solute is derived from a distributed multipole analysis[36] of the quantum mechanically computed molecular charge densities and includes average (implicit) polarization effects in aqueous solutions. Hence, intermolecular electrostatic interactions are modeled accurately, avoiding the need for explicit polarization[37−39] that is computationally prohibitive in combination with a high rank, multipole model. Repulsion−dispersion parameters are fitted to liquid water experimental data in the case of water and taken from an earlier parametrization[30] for the solute. The proposed water model is found suitable for modeling the structure, and a range of properties of ordered ice polymorphs and liquid water not included in its parametrization.

We also use this anisotropic intermolecular potential model to compute the free energy of hydration of 10 rigid, uncharged organic molecules (see Chart 1) using an explicit-solvent free energy perturbation approach. The molecules are chosen so that their ΔG_{hyd} varies from −10 to +2 kcal mol^{-1}, practically covering the entire range of hydration free energies typically obtained for neutral organic solutes. We have deliberately not considered charged species because ionic hydration free energies are generally affected by strong electrostatic finite-size effects.[40] We contrast our hydration free energy calculations with predictions using a self-consistent reaction field quantum mechanical method and also with explicit-solvent free energy perturbation calculations using a point-charge model for the solute, computed at the same level of theory as used for the distributed multipole expansions. The purpose of the latter comparison is to examine if an atomic charge representation is sufficiently accurate to describe the hydration of molecules that are strongly hydrophilic due to their aromatic character and/or hydrogen bond acceptor and donor groups. Such tests are instrumental in establishing whether discrepancies from experiment[13,22,41] in the predicted ΔG_{hyd} are due to the limited accuracy of the atomic-charge representation of the molecular charge density or, alternatively, due to the lack of explicit polarization and errors in the repulsion−dispersion parametrization.

Methodology

The solutes were optimized in isolation at the MP2(fc)/6-31G(d,p) level of theory and were treated as rigid in all subsequent calculations. Water was also kept rigid in its TIP4P conformation (OH = 0.957 Å, HOH = 104.52°), as in the TIP4P[42] and TIP4P/2005[6] water models. The charge densities used in the distributed multipole analysis[36] (DMA) and to compute the molecular electrostatic potentials (ESP) to fit atomic-charge models were computed at the MP2(fc)/aug-cc-pVTZ level of theory for both water and the solutes, apart from pyrene, for which we used PBE0/aug-cc-pVTZ due to computational limitations. All molecular optimizations, charge density, and electrostatic potential calculations were carried out in Gaussian 03.[43]

Model for Intermolecular Forces. *Water−Water Interactions.* In the seminal work of Bukowski et al.,[8] the model for the water−water interactions was computed entirely from first principles using perturbation theory and dimer CCSD(T) calculations. Unfortunately, such models inevitably include nonadditive terms that are impractical for use in the long simulation runs required for accurate hydration calculations. On the other hand, several water pair potentials have been successfully developed by fitting both repulsion−dispersion and electrostatic components to experimental data,[6,7,42,44,45] despite the difficulty in extracting details of the potential energy function from bulk property measurements that represent only averages over the potential surface.[46] The predictive ability of these models often deteriorates for properties and thermodynamic conditions not included in the parametrization. To some extent, this is due to the large number of variables that can be altered and, perhaps, to the existence of multiple optimal solutions that differ in the quality of reproduction of different properties. In this work, we combine both approaches in developing a novel, anisotropic water model: the electrostatic component is computed quantum-mechanically and includes average polarization effects in liquid water at ambient conditions (see the section Electrostatic and Polarization Interactions), while the repulsion−dispersion interaction potential is fitted to experimental liquid water data (see the section Repulsion−Dispersion Interactions).

Electrostatic and Polarization Interactions. From early on in this work, it became apparent that the structure of ice polymorphs and hydrates, as well as the properties of liquid water, water vapor−liquid equilibria, and hydration free energies, could not all be modeled accurately using a distributed multipole model derived from the isolated-water charge density. Ideally, the response of the charge density to its surroundings in condensed phases should be treated using an accurate, distributed polarizability model.[47] However, the computational cost of including explicit polarization in conjunction with a multipole representation[48] in molecular dynamics is computationally expensive and impractical for explicit-solvent free energy perturbation calculations. Hence, we opted for an implicitly polarized model that would, on average, reproduce water's charge density in condensed phases.[27]

The wave function of a water molecule A in the vicinity of water molecule B can be approximated[49] by solving Schrödinger's equation:

$$(H_A + V_{AB})|\Psi_A\rangle = E_A \Psi_A\rangle \tag{1}$$

where H_A is the Hamiltonian of the isolated molecule A and V_{AB} represents the electrostatic interaction between molecules A and B. This interaction term requires computing the electrostatic potential field generated by molecule B, which in this work was approximated by an atomic charge representation

$$V_{AB} = \sum_{b \in B} Q_{00}^{b} \left\{ -\sum_{i \in A} \frac{1}{|\mathbf{r}_i - \mathbf{r}_b|} + \sum_{a \in A} \frac{Z_a}{|\mathbf{r}_a - \mathbf{r}_b|} \right\} \tag{2}$$

with the atomic charges Q_{00}^{b} of molecule B computed from its molecular electrostatic potential with the CHELPG scheme.[50] In eq 2, a and b refer to nuclei and i to electrons. Similarly, the charge density of molecule B will be perturbed due to molecule A, and hence eqs 1 and 2 need to be solved iteratively to self-consistency for both molecules.

We generated structurally uncorrelated, spherical water clusters by selecting 1996 configurations from a 298 K, 1 atm molecular dynamics run of an equilibrated 542-water molecule system with the TIP4P/2005[6] model. For each configuration, one water molecule (central) was arbitrarily chosen. All water molecules separated from the central molecule by an oxygen−oxygen distance of up to 12 Å were retained. These clusters of 225−250 water molecules were used to compute an average, distributed multipole moment model of water in the liquid state without structural relaxation of the cluster geometry.

For each cluster, eq 1 was solved for the central molecule. The central molecule was exposed to the field generated by the atomic charges of the surrounding molecules in the cluster. In the first iteration, the surrounding molecules were modeled with the CHELPG[50] ESP charges of isolated water. In the second iteration, the computed CHELPG ESP atomic charges of the central molecule were placed at the atomic position of the surrounding molecules, and eq 1 was solved repeatedly to self-convergence. Typically, this required fewer than six iterations for atomic charges to converge to 0.001e. The converged charge density of the central molecule was used in distributed multipole analysis[36] to compute water's atomic multipole moments up to hexadecapole. This method is limited in accuracy, because it does not include charge transfer effects, and additionally, the field that the central molecule is experiencing in the cluster is of limited accuracy due to its monopole representation. Nevertheless, it has been found to give comparable lattice energies to those obtained

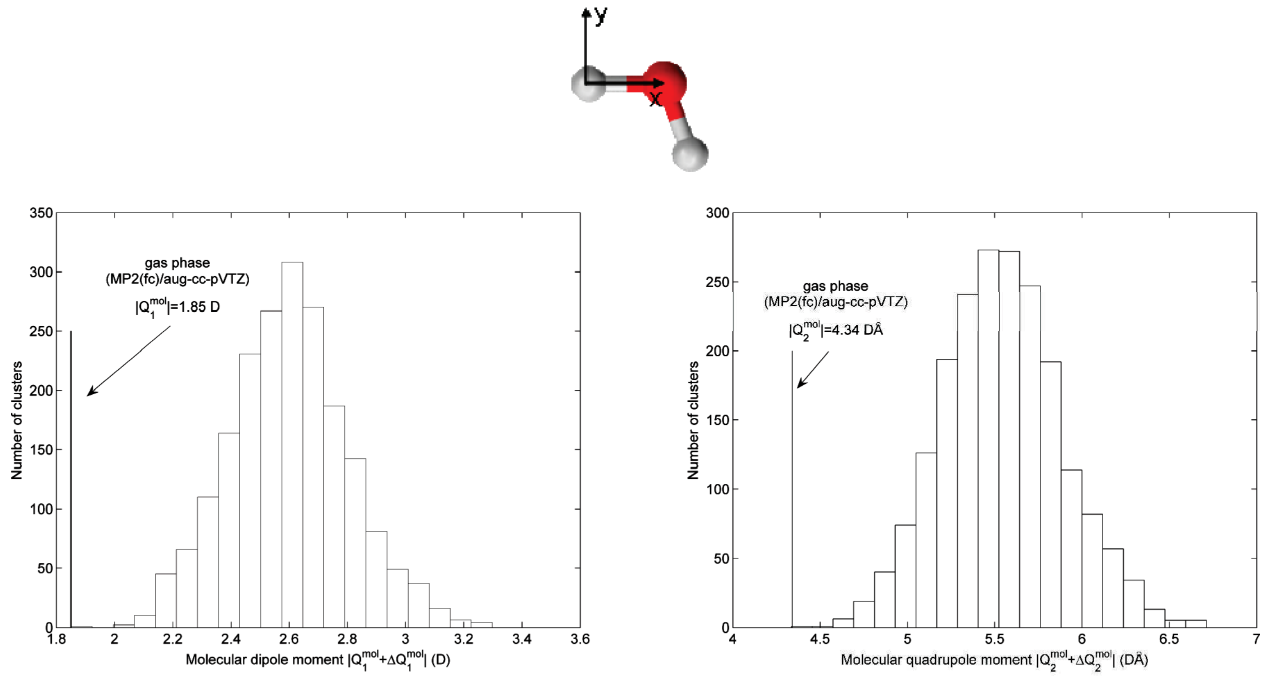


Figure 1. Distribution of magnitudes (defined as $|Q_l^{\rm mol}+\Delta Q_l^{\rm mol}| = \sqrt{[\sum_m(Q_{lm}^{\rm mol}+\Delta Q_l^{\rm mol})^2)]}$ of converged molecular dipole (in D) and quadrupole moments (in DÅ) of the central water molecule in water clusters taken from a TIP4P/2005[6] liquid water simulation at ambient conditions. Molecular multipole moments $Q_{lm}^{\rm mol}$ were computed[54] from the distributed multipoles and refer to the molecular axes system shown (see also Supporting Information).

with an elaborate distributed-polarizability model for organic crystal structures exhibiting hydrogen bonding.[37] Once all clusters had been processed, the atomic multipole moments of the central water molecule were averaged (Figure 1). These average multipole moments can be thought of as being the sum of the static (isolated water molecule) Q_{lm} and average induced multipole moments ΔQ_{lm}, where indices l and m refer to the component of the multipole moments 00, 10, 11s, ..., 44s. The average induced moments ΔQ_{lm} are readily obtained by subtracting the static multipole moments Q_{lm} computed from the distributed multipole analysis of the isolated-water charge density. The obtained average water dipole moment in the liquid is 2.59 D and is in reasonable agreement with the value of 2.7 D obtained for water clusters in the seminal work of Gregory et al.,[51] although the enhancement in the dipole moment of water in condensed phases is still a matter of contention.[52,53]

The induced multipole moments of a polarizable molecule A are determined by the competition between the lowering of the intermolecular energy due to the interaction of the induced moments with the field created by surrounding molecules and the energy cost (internal energy) to distort the molecule's charge density in zero field to the charge density in solution. If we assume a bilinear dependence of the internal energy on the induced moments, it can be shown[55] that the lowering of the intermolecular energy is twice the internal energy (in absolute terms). Hence, the overall lowering of the system's energy due to the polarization of molecule A (called induction energy), without damping, is

$$E_{\rm ind}(A) = (1/2)\sum_{\substack{B\\B\neq A}}\sum_{a\in A}\sum_{b\in B}\sum_{bm,l'm'}\Delta Q^a_{lm}T^{ab}_{lm,l'm'}Q^b_{l'm'}$$

where $T^{ab}_{lm,l'm'}$ is the interaction tensor[55] that depends on the relative position and orientation of sites a and b and the factor 1/2 accounts for the internal energy cost. All molecular dynamics runs were performed using DL_MULTI,[56] which allows the modeling of electrostatic interactions with standard Ewald summation for multipoles up to rank 4 (hexadecapole), although explicit induction is not included. However, the average effect of induction can be accounted for by performing simulations using fixed, effective moments $Q_{{\rm eff},lm} = Q_{lm} + (1/2)\Delta Q_{lm}$, which provides approximately the electrostatic and average induction energy of the system, so that

$$\frac{1}{2}\sum_{\substack{A,B\\B\neq A}}\sum_{a\in A}\sum_{b\in B}\sum_{lm,l'm'}\left(Q^a_{lm}+\frac{\Delta Q^a_{lm}}{2}\right)T^{ab}_{lm,l'm'}\left(Q^b_{l'm'}+\frac{\Delta Q^b_{l'm'}}{2}\right)$$
$$=\frac{1}{2}\sum_{\substack{A,B\\B\neq A}}\sum_{a\in A}\sum_{b\in B}\sum_{lm,l'm'}\left(Q^a_{lm}T^{ab}_{lm,l'm'}Q^b_{l'm'}+\frac{\Delta Q^a_{lm}}{2}T^{ab}_{lm,l'm'}Q^b_{l'm'}+\right.$$
$$Q^a_{lm}T^{ab}_{lm,l'm'}\frac{\Delta Q^b_{l'm'}}{2}+\frac{\Delta Q^a_{lm}}{2}T^{ab}_{lm,l'm'}\frac{\Delta Q^b_{l'm'}}{2}$$
$$=E_{\rm elec}+E_{\rm ind}+\Delta E_{\rm error} \qquad (3)$$

assuming that no damping is used. The last term in eq 3 can in principle be computed at each molecular dynamics time step using $\Delta Q_{lm}/2$ on each atom. However, this would practically double the computational cost and is not worthwhile given that the contribution of $\Delta E_{\rm error}$ to the lattice energy of organic hydrogen-bonded crystals was found to be less than a few percentage points of the lattice energy and around 8% of the induction energy. Hence, this term was omitted in this work, and its effects were absorbed in the water repulsion−dispersion parametrization (see the section "Repulsion−Dispersion Interactions") using the ef-

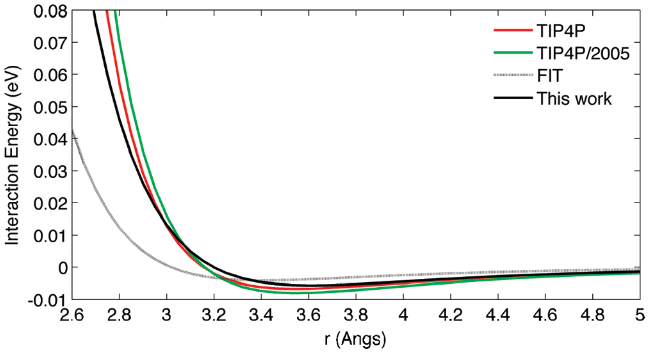


Figure 2. Fitted oxygen−oxygen exp-6 repulsion−dispersion interaction potential (black line) compared with TIP4P,[42] TIP4P/2005,[6] and FIT[30] parametrizations.

fective Q_{eff} multipole moments. In principle, the procedure outlined above can be repeated to compute new average induced multipole moments using the fitted repulsion−dispersion potential, and the whole scheme can be iterated to convergence. We did not, however, follow such an iterative scheme, since the TIP4P/2005 water model provides a sufficiently accurate reproduction of the liquid water structure.[6]

Repulsion−Dispersion Interactions. We considered the combination of the distributed-multipole description of water's electrostatic interactions obtained in this work with repulsion−dispersion parameters of well-established literature water models based on a monopole representation of the charge density,[6,42] but the resulting potential failed to reproduce liquid water properties in all cases. Hence, we carried out a parametrization of the repulsion−dispersion interactions of a three-site, exp-6 model for water using as a starting point the exp-6 repulsion−dispersion parameters of the oxygen and polar hydrogen atoms (hydrogen atoms connected to oxygen and nitrogen) of the FIT[30] empirical model, which was fitted to experimental structural data and sublimation energies of organic crystal structures in conjunction with distributed multipoles.

In order to keep the complexity of the parametrization manageable, the hydrogen parameters of the FIT potential were not altered, and the oxygen−hydrogen cross interaction parameters were computed using standard Lorentz−Berthelot combining rules. We fitted the oxygen parameters to reproduce the liquid water density and the oxygen−oxygen radial distribution function[57] at 298 K and 1 atm. Figure 2 shows a comparison of the obtained optimal oxygen−oxygen exp-6 potential with the starting FIT parametrization and the TIP4P[42] and TIP4P/2005[6] water models (see also the Supporting Information for a full set of repulsion−dispersion parameters and multipole moments for the proposed water model). We note that the water model obtained in this work appears less repulsive than the two potentials of the TIP4P family, but this is not the case, given that our potential also includes oxygen−hydrogen interactions that are strongly repulsive at typical hydrogen bonding distances.

In Figure 3, it can be seen that the proposed water model reproduces well the experimentally determined oxygen−oxygen radial distribution function at 298 K and 1 atm, given the experimental error[57] and the neglect of quantum effects[58,59] and molecular flexibility[60] in our classical calculations. The number of water molecules in the first hydration shell, obtained by integrating the oxygen−oxygen radial distribution function up to the distance of its first minimum, is 4.8 and also in good agreement with experimental results (Figure 3b). Moreover, the model reproduces well the second peak at 4.5 Å; this is indicative of the accuracy in modeling the hydrogen bond network in liquid water. It is encouraging that the water model also reproduces the oxygen−hydrogen and hydrogen−hydrogen radial distribution functions that were not included in the parametrization. The predicted liquid water density at 298 K and 1 atm is 0.994 g cm^{-3} compared to the experimental value[61] of 0.997 g cm^{-3}. The model was further tested by modeling several additional solid and liquid water properties as detailed in the section Testing of the Model for the Intermolecular Forces.

Solute−Water Interactions. For the solutes, we followed a simpler approach to derive the $Q_{lm} + \Delta Q_{lm}$ multipole moments, by computing the charge density for distributed multipole analysis using a polarizable continuum model[62] with the United Atom Topological Model (UA0) for the water cavity and the default parameters for water as a solvent in Gaussian 03.[43] As in the case of water, the solute's electrostatic intermolecular interactions were modeled using the effective multipole moments $Q_{\mathrm{eff},lm} = Q_{lm} + \Delta Q_{lm}/2$. Hydration free energy calculations were also performed using effective ESP atomic charges $Q_{\mathrm{eff},00} = Q_{00} + \Delta Q_{00}/2$ computed from the same dielectric continuum and gas-phase wave functions with the CHELPG scheme.[50]

The solute−water repulsion−dispersion interactions were obtained by applying standard Lorentz−Berthelot combining rules[63] to the original FIT parametrization[30] for the solute and the water model developed in this work (see the Supporting Information). In the case of the free energy perturbation calculations (see Repulsion−Dispersion Interactions), the exp-6 solute−water repulsion−dispersion potential was supplemented with a γ/r^{12} term to alleviate the divergence of the exp-6 potential to minus infinity at short distances. The coefficients γ were chosen for each interaction independently so as to make the added term negligible for all solute−water intermolecular distances sampled in a simulation with the unperturbed solute−water interaction potential.

Testing of the Model for the Intermolecular Forces. The developed water potential was first tested by modeling the five proton-ordered ice polymorphs XI,[64] II,[65] IX,[66] VIII,[67] and XIII.[68] Ice XIII is a hydrogen-ordered phase of disordered ice V prepared under high pressure but determined at ambient pressure by powder neutron diffraction.[68] In all ordered ice phases, the coordination of water molecules resembles the liquid phase with each water molecule tetrahedrally hydrogen bonded to four neighbors. The ice polymorphs were lattice energy minimized with respect to the relative position and orientation of the water molecules and the cell geometry using DMACRYS.[69] The minimizations were performed within the resulting space group symmetry constraints after lowering the symmetry so the asymmetric unit comprised complete molecules. Charge−charge, charge−dipole, and dipole−dipole interactions were calculated with Ewald summation, while repulsion−dispersion

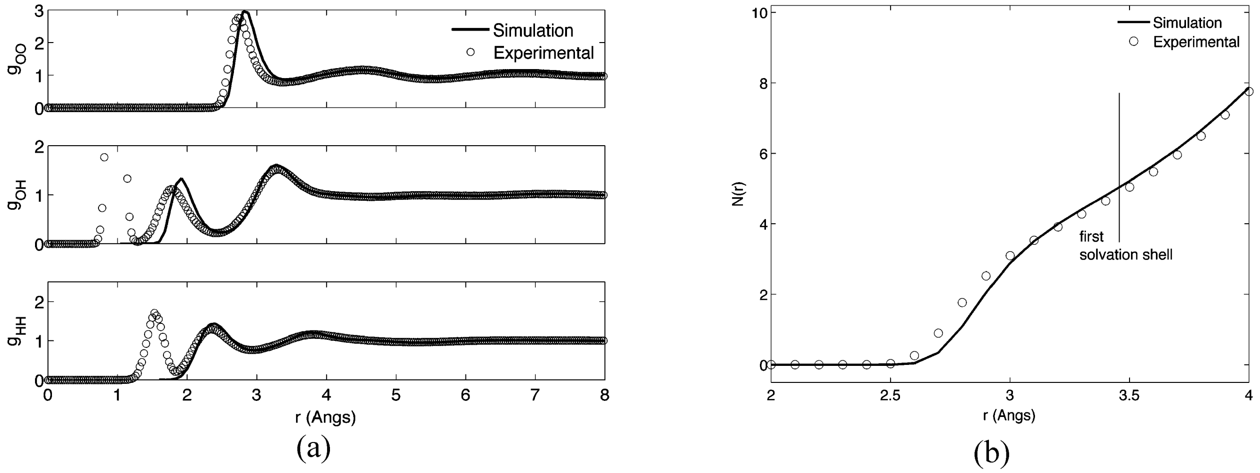


Figure 3. (a) Simulated site−site distribution functions for liquid water at 298 K and 1 atm compared with experimental data.[57] Simulation results for OH and HH do not include contributions from bonded atoms. (b) Average number of water oxygen atoms *N*(*r*) within distance *r* from any given water oxygen atom under the same conditions.

and higher multipole contributions were evaluated in direct space up to a 15 Å cutoff. All ice lattice energy minimizations were performed at 0 Pa pressure, apart from ice VIII, which was modeled at the determination pressure of 2.4 GPa. The neglected thermal expansion changes the volume of organic crystals by approximately one percentage point per 100 K, which is within the error margin of the proposed water model. The reproduction of the five ordered ice polymorphs was contrasted with three different intermolecular potential models: TIP4P,[42] TIP4P/2005,[6] and the original FIT parametrization[30] combined with gas-phase MP2(fc)/aug-cc-pVTZ water multipole moments. For the two water models of the TIP4P family, we included in the lattice energy calculations the oxygen partial charge at the non-nuclear position as described in the original force field specifications. The quality of the modeling of the ice structures was evaluated by computing[70] the root-mean-square discrepancy in overlaying the oxygen positions of a 20-water-molecule cluster ($\mathrm{RMS}_{\mathrm{cs}\text{-}20}$).

The Cambridge Structural Database[71] contains a trihydrate crystal structure of pyridine that was determined at 223 K, that is, 20 K below its decomposition temperature.[72,73] The structural reproduction of this crystal structure was used as an additional test of the quality of the potential model, and in particular of the solute−water cross interactions, by examining its thermal stability in an isothermal, isobaric ensemble with fully flexible cell (hereafter referred to as *NσT*) molecular dynamics simulations at its determination conditions. The solute−water interaction potential was further tested by assessing its ability to model hydration free energies as detailed in the section Free Energy of Hydration.

The liquid water density at 1 atm was computed in the temperature range 253−323 K in a series of isothermal, isobaric molecular dynamics simulations. The dependence of the simulated water density with temperature was fitted to a fourth-order polynomial that was subsequently used to compute the thermal expansion coefficient $\alpha_P = 1/V(\partial V/\partial T)_P = -1/\rho(\partial\rho/\partial T)_P$ at 298.15 K. The result was checked for consistency with the value obtained from the volume-enthalpy fluctuations[74] $\alpha_P = (\langle VH\rangle_{NPT} - \langle V\rangle_{NPT}\langle H\rangle_{NPT})/(kT^2\langle V\rangle_{NPT})$ in an isothermal, isobaric simulation at 298.15 K and 1 atm, where $\langle x\rangle$ denotes ensemble averages.

The experimental molar volume of water at 298.15 K depends linearly on pressure in the range 1−200 atm. Hence, the isothermal compressibility $k_T = -1/V(\partial V/\partial P)_T = 1/\rho(\partial\rho/\partial P)_T$ was obtained with a series of five isothermal isobaric simulations at 1, 10, 50, 100, and 200 atm at 298.15 K by computing the slope of a linear model fitted to the pressure dependence of the simulated molar volume. To check consistency, k_T was also computed from the volume fluctuations $K_T = (\langle V^2\rangle_{NPT} - \langle V\rangle_{NPT}{}^2)/(kT\langle V\rangle_{NPT})$ at 298.15 K and 1 atm.

The enthalpy of vaporization of liquid water was computed as the enthalpy difference $\Delta H_{\mathrm{vap}} = H_{\mathrm{vap}} - H_{\mathrm{liq}} = U_{\mathrm{vap}} - U_{\mathrm{liq}} + P(V_{\mathrm{vap}} - V_{\mathrm{liq}})$ where U is the configurational energy, ignoring the kinetic energy, which will be equal in the two phases at the coexistence (vapor) pressure and temperature. If we assume that the vapor behaves ideally, this equation simplifies to

$$\Delta H_{\mathrm{vap}} \approx -\langle U_{\mathrm{liq}}\rangle_{NPT} + RT \tag{4}$$

which was used to compute ΔH_{vap} at 298.15 K from an isothermal, isobaric simulation of liquid water at 1 atm. This pressure is higher than the vapor pressure at 298.15 K, but the effect on ΔH_{vap} can be ignored since the liquid phase enthalpy change over this pressure range can be considered negligible. It has been proposed that ΔH_{vap} calculated with effective pair potentials should also be corrected for self-polarization,[22,75] so that the energy cost to distort the molecular charge density to its polarized state is taken into account. This positive correction can be approximated[75] from the difference in the dipole moment of water in the liquid and gas phases and the isotropic scalar polarizability from the relation $E_{\mathrm{cor}} = 1/2(\mu_{\mathrm{liq}} - \mu_{\mathrm{gas}})/a$. However, this correction is not appropriate for our model, because the intermolecular electrostatic interactions are modeled using the effective multipole moments Q_{eff} that already account for the cost of distorting the charge density of water from its gas state to its average polarized state in liquid water.

The experimental liquid water heat capacity under constant pressure $C_P = (\partial H/\partial T)_p$ varies[61] in the temperature range 273−323 K by less than 1%; this is reflected in our model by an almost linear dependence of liquid water's enthalpy on temperature. Hence, C_P was computed from the slope of the liquid enthalpy with respect to temperature in six isothermal, isobaric simulations in the aforementioned temperate range. The liquid water enthalpy was computed as the sum of the average configurational energy, rotational and translational kinetic energy (difference from $3RT$ was within standard deviation), and the PV_{liq} term. To check for consistency, C_P was also computed from the enthalpy fluctuations $C_P = (\langle H^2\rangle_{NPT} - \langle H\rangle^2_{NPT})/(kT^2)$ in an isothermal, isobaric simulation at 298.15 K and 1 atm.

The heat capacity under constant volume $C_V = (\partial E/\partial T)_V$ at 298.15 K and 1 atm was computed from the almost perfectly linear dependence of the total energy E on temperature in a series of four constant volume, constant temperature simulations in the range 293.15−308.15 K in 5 K intervals. For these simulations, the density was constrained to the experimental density of water at 298.15 K. The heat capacity C_V was also computed from the liquid water's potential energy U fluctuations as

$$C_V = \frac{\langle E^2\rangle_{NVT} - \langle E\rangle^2_{NVT}}{kT^2} = \frac{\langle U^2\rangle_{NVT} - \langle U\rangle^2_{NVT}}{kT^2} + 3R \quad (5)$$

where the last term arises from the rotational and translational kinetic energy contribution of the rigid water molecules. We finally computed the self-diffusion coefficient using the Einstein relationship

$$D = \frac{1}{6(t - t_0)}\lim_{t\to\infty}\langle|\mathbf{r}(t) - \mathbf{r}(t_0)|^2\rangle$$

The water model was also tested by carrying out molecular simulations of direct liquid−vapor coexistence[76] to compute the saturated liquid and vapor densities as a function of temperature. A previously equilibrated cubic cell of 542 molecules was expanded 2.5 times in the z direction, and the additional volume was left empty. The system was allowed to evolve in the NVT ensemble for 300 ps following a 40 ps equilibration and for six temperatures spanning the range 300−550 K. The coexistence densities and interface thickness d were computed by fitting a hyperbolic tangent function[76] of the form

$$\rho(z) = \frac{1}{2}(\rho_{\text{liq}} + \rho_{\text{vap}}) - \frac{1}{2}(\rho_{\text{liq}} + \rho_{\text{vap}})\tanh[(z - z_0)/d] \quad (6)$$

where z_0 is Gibbs' dividing surface. For all temperatures studied, the two interfaces were symmetric and the liquid region was sufficiently wide to provide a reliable measure of the saturated liquid density. Longer simulation times and a larger number of water molecules gave statistically identical coexistence densities. The surface tension was estimated from the components of the pressure tensor.[76,77]

Free Energy of Hydration. Several explicit-solvent methods to compute the free energy of hydration have been proposed.[16,21,78] In the so-called free energy perturbation approach, the free energy difference[19] between two states A and B of a system defined with energy functions U_A and U_B can be computed as

$$\Delta G_{A,B} = G_B - G_A = -\frac{1}{\beta}\ln\langle\exp[-\beta\Delta U_{A,B}]\rangle_A = \frac{1}{\beta}\ln\langle\exp[\beta\Delta U_{A,B}]\rangle_B \quad (7)$$

i.e., by accumulating the energy differences $\Delta U_{\text{A,B}} = U_B - U_A$ (work distributions) in a simulation with either Hamiltonian (forward or reverse simulation). Exponential averaging in one direction does not necessarily give the minimum bias and variance of the free energy difference for a given set of $\Delta U_{A,B}$ measurements.[17] Bennet's acceptance ratio[79] is generally more efficient but requires sampling in both forward and reverse directions. As the purpose of our current work is not to identify the most efficient method to compute the free energy, but to test the effect of the accuracy of the intermolecular potential model in hydration calculations, all free energy estimates are based on forward exponential averaging (solute-creation only), apart from the free energy for the transfer of a single water molecule from the gas phase to the liquid state, which was independently computed with exponential averaging in both directions.

In hydration free energy calculations, state B in eq 7 corresponds to the solute interacting with the solvent, while in state A the solute−solvent interactions are fully annihilated. For practically all systems of interest, the overlap in phase space between the two states is sufficiently low that the free energy difference cannot be computed from data for only the end states A and B. Instead, the solute needs to be introduced into solution in stages by defining a series of intermediate states $i = 1, ..., N$ such that

$$\Delta G_{A,B} = \sum_{i=1}^{N-1}\Delta G_{i,i+1} = -\frac{1}{\beta}\sum_{i=1}^{N-1}\ln\langle\exp[-\beta\Delta U_{i,i+1}]\rangle_i \quad (8)$$

where states 1 and N are the end states A and B. In this work, the solute is introduced to a previously equilibrated water system, by first switching on the solute−solvent repulsion−dispersion interactions in a series of 29 simulations $U_{\text{solute−water}} = \lambda U^{\text{vdW}}_{\text{solute−water}}$, for $\lambda = $ (0.0, 1e−9, 1e−8, 1e−7, 1e−6, 1e−5, 1e−4, 3e−4, 7e−4, 1e−3, 2e−3, 4e−3, 7e−3, 0.01, 0.015, 0.025, 0.04, 0.05, 0.07, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9). The uneven spacing in λ reflects the much larger curvature $dG/d\lambda = \langle dU/d\lambda\rangle_\lambda$ for small λ. We did not observe numerical instabilities at low λ values, where the λ-scaled potential increases abruptly at short separations. However, designing a suitable[20,21] λ-dependent (soft-core) functional form for exp-6 potentials would be beneficial in alleviating this source of potential instability and also reducing the number of intermediate states required. Preliminary simulations using a soft-core solute−water repulsion−dispersion interaction[21] gave free energies of hydration that were within standard deviation from our predictions, as discussed in the Supporting Information.

Once the solute's van der Waals interactions were fully switched on, the solute's effective multipole moments Q_{eff} were gradually switched on in a series of 10 simulations with

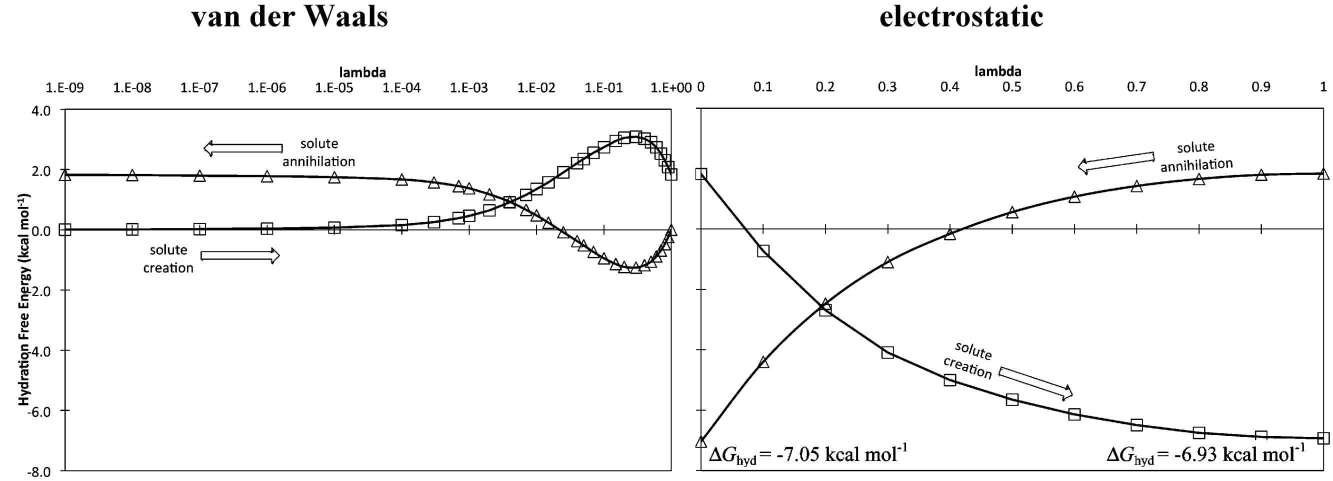


Figure 4. Repulsion−dispersion (left) and electrostatic (right) contributions to the free energy of hydration of water at ambient conditions as a function of the coupling parameter λ for solute-creation and solute-annihilation simulations.

the solute−water interaction potential being $U_{\text{solute-water}} = U^{\text{vdW}}_{\text{solute-water}} + \lambda U^{\text{elec+ind}}_{\text{solute-water}}$, for λ = (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9). By comparing with results using larger perturbation steps, we conclude that the hydration free energies obtained with this set of intermediate states are converged to accuracy that is within the statistical uncertainty due to finite sampling of the work distributions. The isothermal, isobaric simulations for all intermediate states were performed in parallel starting from the same equilibrated configuration with the unperturbed solute−water $U_{\text{solute-water}}$ interaction potential. The infinitesimal correction in the predicted hydration free energy due to the reduction in system volume when the solute is fully decoupled was ignored.[16]

The bias in exponential averaging has opposite signs for the forward and reverse simulations and, generally, depends on the size of the perturbation step. In solute-creation simulations, large perturbation steps lead to less negative ΔG_{hyd} caused by the overlap of the solute and water. In annihilation simulations, large perturbation steps lead to more negative ΔG_{hyd} due to the loss of van der Waals and electrostatic attraction between the solute and water. Figure 4 shows that, for the free energy of transfer of a water molecule from a gas to a liquid at ambient conditions, the result from forward and reverse exponential averaging differs by only 0.1 kcal mol^{-1}, which is less than the 0.3−0.4 kcal mol^{-1} statistical uncertainty in determining ΔG_{hyd} for our method (c.f. the section Modeling of Hydration) and confirms that the $\Delta\lambda$ increments are sufficiently small.

For comparison purposes, the hydration free energies were also computed using the polarizable continuum model[62] (PCM) at the B3LYP/aug-cc-pVTZ level of theory. The solvent cavity was built using the United Atom Topological Model with the recommended atomic radii for solvation free energy calculations (UAHF), and the default parameters for water solvent as in Gaussian 03.[43] The hydration free energies with the polarizable conductor calculation model[80] (CPCM) differed by less than 0.1 kcal mol^{-1} and are not reported. We note that the use of the UFF force field atomic radii (UA0) gave large errors, of up to 5 kcal mol^{-1}, in the calculated hydration free energies; this shows the pronounced sensitivity[81,82] to the method used to construct the cavity surface. The atomic radii had a much smaller effect in computing the solute's distributed multipole moments.

All reported experimental and computed free energies of hydration refer to ambient conditions and correspond to the use of molar concentration units for the solute both in the gas phase and in solution.[82,83]

Molecular Dynamics Simulations. Molecular dynamics simulations were performed with a 1.2 fs time step, the leapfrog integrator, and the Nose−Hoover thermostat and barostat[84] as implemented in DL_MULTI.[56] The Ewald precision was set to 5 × 10^{-7} for all multipole orders in reciprocal space and charges in direct space. The precision of all other multipole orders in direct space was 5 × 10^{-8}. With these settings, the integration of the equations of motion of liquid water at ambient conditions for 1 ns gives a drift in the total Hamiltonian per degree of freedom that is more than an order of magnitude smaller than kT.

All pure water simulations were performed with 542 water molecules. For hydration free energy calculations, a solute molecule was inserted in a previously equilibrated cell that contained 542 water molecules at ambient conditions, and 2−4 water molecules were removed, depending on the size of the solute, to alleviate overlaps and short contacts. Work distributions were accumulated every 10 time steps over a 250 ps isothermal, isobaric simulation following 50 ps of equilibration. This scheme is sufficient to obtain converged free energy differences to within a few tenths of a kilocalorie per mole; further sampling of the work distributions was limited by the availability of computing resources. The simulations carried out to obtain water properties varied in length from 1 ns to several nanoseconds, depending on the property under consideration; the longest runs performed were those required to compute the density of supercooled liquid water. A 542-water-molecule, 100 ps simulation required approximately 8 h on an eight-core Intel Xeon E5462 2.8 GHz node, resulting in the equivalent of one CPU year to compute the hydration free energy per solute. Roughly, the use of a distributed multipole model up to hexadecapole is an order of magnitude more expensive compared with simpler monopole charge models.

Table 1. Lattice Energy Minimization of Five Ordered Ice Polymorphs[a]

	% error lattice lengths						
model	a	b	c	density (g cm^{-3})	% error density	lattice energy (kcal mol^{-1})	RMS$_{cs\text{-}20}$[b] (Å)
ice XI,[64] *Cmc*2$_1$, Z' = 2, 5 K, ambient pressure, experimental density 0.930 g cm^{-3}							
TIP4P	−3.33	−0.35	−0.98	0.975	+4.84	−13.62	0.12
TIP4P/2005	−2.53	+0.33	−0.20	0.953	+2.46	−15.05	0.11
FIT+vacuo DMA[c]	+0.33	−11.15	−4.01	1.087	+16.87	−14.96	0.29
this work	**+2.74**	**−2.58**	**+0.95**	**0.921**	**−1.02**	**−15.79**	**0.12**
ice II,[65] $R\bar{3}$, Z' = 2, 110 K, ambient pressure, experimental density 1.180 g cm^{-3}							
TIP4P	−1.98	−1.98	−2.18	1.255	+6.40	−13.38	0.10
TIP4P/2005	−1.24	−1.24	−1.58	1.229	+4.18	−14.83	0.09
FIT+vacuo DMA[c]	−5.24	−5.24	−0.91	1.326	+12.38	−13.46	0.22
this work	**−0.14**	**−0.14**	**+2.16**	**1.158**	**−1.83**	**−14.91**	**0.10**
ice IX,[66] $P4_12_12$, Z' = 1.5, 110 K, ambient pressure, experimental density 1.160 g cm^{-3}							
TIP4P	−1.93	−1.93	−1.46	1.224	+5.52	−13.52	0.07
TIP4P/2005	−1.25	−1.25	−1.01	1.202	+3.60	−14.97	0.05
FIT+vacuo DMA[c]	−6.67	−6.67	+6.32	1.253	+7.97	−13.80	0.27
this work	**−0.30**	**−0.30**	**+3.82**	**1.125**	**−3.09**	**−14.99**	**0.10**
ice VIII,[67] $I4_1/amd$, Z' = 0.5, 10 K, 2.4 GPa, experimental density 1.629 g cm^{-3}							
TIP4P	−2.62	−2.62	+2.33	1.679	+3.03	−5.33	0.15
TIP4P/2005	−1.76	−1.76	+3.08	1.637	+0.50	−6.54	0.18
FIT+vacuo DMA[c]	−3.09	−3.09	−12.46	1.982	+21.62	−9.07	0.27
this work	**+0.04**	**+0.04**	**−3.30**	**1.684**	**+3.34**	**−8.99**	**0.10**
ice XIII,[68] $P2_1/a$, Z' = 7, 80 K, ambient pressure, experimental density 1.251 g cm^{-3}							
TIP4P	+0.80	−2.83	−2.22	1.309	+4.63	−13.20	0.16
TIP4P/2005	+1.60	−2.18	−1.52	1.282	+2.48	−14.64	0.15
FIT+vacuo DMA[c]	−3.85	−2.72	−3.65	1.441	+15.14	−13.59	0.26
this work	**+0.36**	**+0.94**	**+1.27**	**1.241**	**−0.80**	**−14.86**	**0.12**

[a] All minimizations were performed at 0 K and 0 Pa with water held rigid to the TIP4P conformation, apart from ice VIII, which was minimized at the experimental pressure of 2.4 GPa. [b] RMS overlay[70] of the oxygen positions of a 20-molecule water cluster. [c] Repulsion−dispersion interactions computed using FIT[30] repulsion−dispersion potential combined with multipole moments derived from the DMA of the isolated-water MP2(fc)/aug-cc-pVTZ charge density.

The hydration free energy depends on the exact protocol with which the solute creation is effected, which includes the treatment of long-range repulsion−dispersion interactions.[85] In this work, repulsion−dispersion interactions were computed up to a 10 Å cutoff, which corresponds to three times the oxygen−oxygen distance in liquid water's first hydration shell (Figure 3b). Long-range correction[84] to energy and pressure was only applied in pure water calculations. In the hydration free energy calculations, water−water and solute−water repulsion−dispersion interactions were smoothly switched off between 9 and 10 Å using a cubic spline. This alleviates the need to apply different long-range corrections to the two configurational energies when computing the work distributions, which we estimated to change the predicted free energies of hydration by less than the statistical uncertainty (Supporting Information) due to finite sampling of the work distributions.

Results

Accuracy of Water Model. *Modeling of Ordered Ice Polymorphs and Pyridine Hydrate.* Table 1 contrasts the lattice energies and lattice parameters for five proton-ordered ice polymorphs with the proposed water model and with other popular potentials for water. It can be seen that the proposed water model achieves the smallest overall RMS error in the reproduction of the 20-molecule coordination sphere of the five ordered ice polymorphs considered, which was in all cases smaller than 0.12 Å. The reproduction of hydrogen bonding geometries is also satisfactory. However, the differences in reproduction with the proposed water model compared to TIP4P[42] and TIP4P/2005[6] are small, and comparable to the effect of neglected thermal expansion and molecular distortions. Using the FIT[30] repulsion−dispersion potential with the isolated-water MP2(fc)/aug-cc-PVTZ multipole moments largely overestimates the ices' densities. This contrasts the successful modeling of four hydrogen-ordered ice polymorphs with FIT and multipole moments computed from the isolated-water MP2/6-31G(d,p) charge density by Hulme and Price.[86] The agreement in this case can be attributed to cancellation of errors due to the limited size of the 6-31G(d,p) basis set in computing the water's charge density. It is encouraging that the low-temperature ice XI corresponds to the most stable form with all the models in Table 1, and that the energy differences between all polymorphs modeled at 0 Pa lie in a narrow energy range of a couple of kilocalories per mole.

The isothermal bulk modulus of ice II at 0 K and 0.35 GPa with our water potential is approximately 20 GPa, compared with 23 GPa for TIP4P/2005 and 14 at 0.35 GPa and 225 K, experimentally.[87]

The pyridine trihydrate crystal structure is found to be stable in a $N\sigma T$ molecular dynamics simulation at the experimental determination conditions of 223 K and 1 atm. Pyridine molecules remain enclosed between water layers with limited π−π stacking despite the almost parallel arrangement of their molecular planes, in good agreement with the experimental crystal structure that is reflected in cell length errors of less than 2.5%. The first peak in the

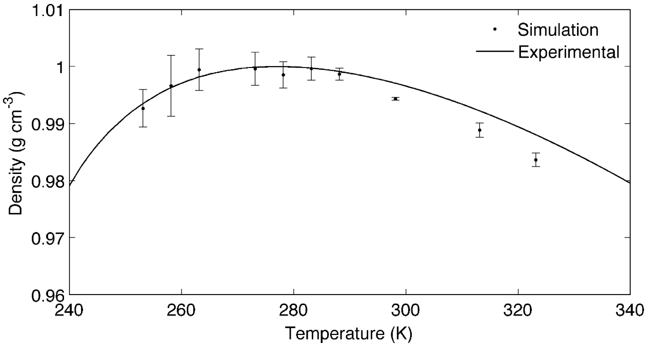


Figure 5. Simulated vs experimental[61,91] liquid water densities at 1 atm.

site−site correlation function between pyridine's nitrogen and water's oxygen atoms at these conditions is at 2.65 Å compared with the 2.79 Å N•••O hydrogen bond length in the experimentally determined crystal. This underestimation of hydrogen bond lengths is due to the use of implicit induction in conjunction with a repulsion−dispersion potential for pyridine that was parametrized using isolated-molecule multipole moments.

Modeling of Liquid Water and Water Vapor−Liquid Equilibria. In Figure 5, it can be seen that the proposed water model successfully predicts the density variation along the ambient pressure isobar including the temperature of maximum density. The thermal expansion coefficient computed from the volume-enthalpy fluctuations changes sign at approximately 279 K (6 °C), which is in good agreement with the 273 K temperature of maximum density obtained from a fourth-order polynomial fitted to density−temperature data and the experimental temperature of maximum density of 277 K. For comparison, from the TIP family of water models, TIP4P/2005,[6] TIP4P/Ew,[44] and TIP5P[7] (without Ewald summation) predict[88] this water density anomaly. It should be noted, however, that these models were fitted to do so, while the parametrization of our model only used the liquid water density at 298.15 K. From the slope of the fourth-order polynomial, we compute a thermal expansion coefficient α_P of 3.4×10^{-4} K^{-1} at 298.15 K compared with 3.1×10^{-4} K^{-1} calculated from the volume-enthalpy fluctuations and 2.6×10^{-4} K^{-1} obtained experimentally.[89] Despite predicting the temperature of maximum density, TIP5P overestimates[90] α_T by more than 100%, while TIP4P/2005[6] is only slightly better than our model with α_T equal to 2.8×10^{-4} K^{-1}. The average simulated isothermal compressibility in the pressure range 1−200 atm computed by fitting a linear model for the dependence of the simulated molar volume on pressure is 5.4×10^{-5} atm^{-1} compared with 5.6×10^{-5} atm^{-1} from volume fluctuations. The experimental isothermal compressibility in the same pressure range is 4.5×10^{-5} atm^{-1}, while the values for TIP5P[90] and TIP4P/2005[6] are 4.1×10^{-5} and 4.6×10^{-5} atm^{-1}, respectively.

The calculated enthalpy of vaporization at 298.15 K and 1 atm overestimates the experimental value by approximately 1.6 kcal mol^{-1}. For comparison, TIP4P/2005 and TIP4P/Ew overestimate[6] the experimental value by 1.5 and 1.2 kcal mol^{-1}, although this discrepancy is mainly because these models were parametrized by fitting to the experimental vaporization enthalpy after it was corrected for the polarization energy. In all these calculations, ΔH_{vap} is computed from eq 4 by treating the system classically. Strictly, the energy of a quantum-mechanical oscillator depends on the frequency, and hence there is a contribution to ΔH_{vap} due to the quantum-mechanical character of the intermolecular modes of liquid water and also due to the shifting of the intramolecular frequencies when a water molecule goes from the gas to the liquid phase. The vibrational corrections to ΔH_{vap} have been estimated[44] to be approximately −0.07 kcal mol^{-1} at 298 K. Moreover, the ΔH_{vap} increases by approximately 0.5 kcal mol^{-1} on going from H_2O to T_2O,[92] and the configurational energy of liquid water calculated with path-integral simulations is on the order of 1 kcal mol^{-1} less negative compared with classical water at ambient conditions.[59,93] All this suggests that inclusion of quantum effects would further reduce the deviation of the predicted ΔH_{vap} from experimental values.

The simulated enthalpy of liquid water in the temperature range 273.15−323.15 K is a linear function of temperature to a good approximation. From the slope of the fitted line, we compute that C_P is 101.4 J mol^{-1} K^{-1} compared to 75.3 J mol^{-1} K^{-1} experimentally and 100.8 J mol^{-1} K^{-1} obtained from the enthalpy fluctuations. For comparison, TIP4P/2005 and TIP4P/Ew predict C_P to be 88.3 and 89.5 J mol^{-1} K^{-1}, respectively, although once more these values do not include the polarization energy that, nevertheless, was included in their parametrization. Although the vibrational correction[44] to the enthalpy of vaporization is small, its temperature dependence is significant and results in a −9.4 J mol^{-1} K^{-1} correction to C_P at 298.15 K that significantly reduces the discrepancy of all three rigid-water models from experimental results. Quantization of classical models is also known to reduce the value of the heat capacity.[59] From the linear dependence of the simulated total energy with temperature in the range 293.15−308.15 K, we compute that the heat capacity under constant volume without any corrections is 101.2 J mol^{-1} K^{-1} compared to 74.5 J mol^{-1} K^{-1} obtained experimentally. The value obtained from the configurational energy fluctuations in a canonical ensemble simulation at 298.15 K and 1 atm is in excellent agreement and equal to 101.1 J mol^{-1} K^{-1}. The predicted self-diffusion coefficient is 1.4×10^{-9} m^2 s^{-1} and hence severely underestimates the experimental value (see Table 2) and the predictions of the two best performing TIP water models, TIP4P/2005 and TIP4P/Ew, that predict D to be 2.1 and 2.4×10^{-9} m^2 s^{-1}, respectively.[44,88] Nevertheless, the agreement of these models may be fortuitous given that quantum effects have been shown to increase[58] D by 50%.

A summary of the calculated properties of liquid water at 298.15 K and 1 atm with the proposed model is presented in Table 2. The computed liquid water properties are self-consistent, in that the difference in heat capacities obtained from the molar volume, thermal expansion coefficient, and isothermal compressibility satisfy $C_P - C_V = T\hat{V}\alpha_P^2/k_T$.

In Figure 6, the simulated water vapor−liquid coexistence envelope in the temperature range 298−550 K is presented compared to experimental data. The vapor density is

Table 2. Simulated and Experimental Properties of Liquid Water at 298.15 K and 1 atm[a]

property	simulated	experimental[b]
density (g cm^{-3})	0.994	0.997
$10^4\alpha_P$ (K^{-1})	3.4[c]	2.6
$10^5\kappa_T$ (atm^{-1})	5.4[d]	4.5
C_V (J mol^{-1} K^{-1})	101.2[e,f]	74.5
C_P (J mol^{-1} K^{-1})	101.4[e,f]	75.3
ΔH_{vap} (kcal mol^{-1})	12.1[f]	10.5
10^9D (m^2 s^{-1})	1.4	2.3

[a] α_P is the thermal expansion coefficient; κ_T, the isothermal compressibility; C_V and C_P, the heat capacities at constant volume and pressure, respectively; ΔH_{vap}, the enthalpy of vaporization; and D, the self-diffusion coefficient. [b] Experimental values from refs 61, 89, 91, 94, and 95. [c] Computed from the slope of a fourth-order polynomial fitted to simulated density vs temperature data. [d] Computed from the slope of a linear model fitted to simulated molar volume vs pressure data. [e] Computed assuming a linear dependence of enthalpy and total energy on the temperature in NPT and NVT simulations, respectively. [f] Computed without any dipole moment, vibrational, and quantum corrections.

reproduced excellently, while the saturated-liquid density is moderately underestimated at high temperatures. This is consistent with the slightly overestimated thermal expansion coefficient, given that the liquid density at coexistence is almost equal to the liquid density at ambient pressure at a given temperature. From the normal and tangential components of the pressure tensor, we found that the surface tension of our water model weakly underestimates the experimental values by 4−6 mJ m^{-2} in the temperature range 350−550 K and shows the experimental decreasing trend with increasing temperature. We note that the statistical uncertainty of the surface tension in our 300 ps simulations is on the same order of magnitude as this discrepancy (see the Supporting Information). The direct coexistence method is not sufficiently accurate at higher temperatures to determine the critical temperature T_c. However, from the temperature dependence[78] of the surface tension, we estimated the critical temperature to be 637 K, which compares favorably with the experimental value of 647 K. We note that our early attempts to model water using gas-phase multipole moments led to a severe underestimation of the critical temperature despite reproducing the liquid density and oxygen−oxygen radial distribution function at ambient conditions.

When the solute is the same as the solvent, the free energy of solvation (or free energy of vaporization) can be computed from[96] $\Delta G_{\mathrm{vap}} = \Delta G_{\mathrm{solv}} = -kT \ln (\hat{V}_{\mathrm{vap}}/\hat{V}_{\mathrm{liq}})$, where $\hat{V}$ is the molar volume of the two phases in equilibrium. Hence, the experimental free energy of hydration for water at 298.15 K and 1 atm is estimated to be −6.3 kcal mol^{-1}, in good agreement with the −6.9 kcal mol^{-1} value obtained from the free energy perturbation calculations (see Figure 4). The entropy of liquid water computed from $S_{\mathrm{liq}} = -(\Delta H_{\mathrm{vap}} + \Delta G_{\mathrm{vap}})/T + S_{\mathrm{vap}}$ is approximately 3 cal mol^{-1} K^{-1} lower than the experimental at these conditions. This is consistent with the simulated site−site distribution functions (Figure 3) being somewhat more structured compared with the neutron experimental results, although any such qualitative comparison is also subject to errors due to the neglect of molecular flexibility and quantum effects in our simulations.

Modeling of Hydration. In Figure 7, the predicted hydration free energies for the 10 organic solutes are compared with the experimental values.[97] The root-mean-square (RMS) error using the distributed multipole model is 1.50 kcal mol^{-1}. The maximum error of 2.49 kcal mol^{-1} is obtained for imidazole, which is the molecule with the most negative free energy of hydration in our set. For comparison, the polarizable continuum model, which has been parametrized using experimentally determined free energies of hydration, gives a root-mean-square error of 1.53 kcal mol^{-1} with a maximum error of 3.24 kcal mol^{-1} for pyrene.

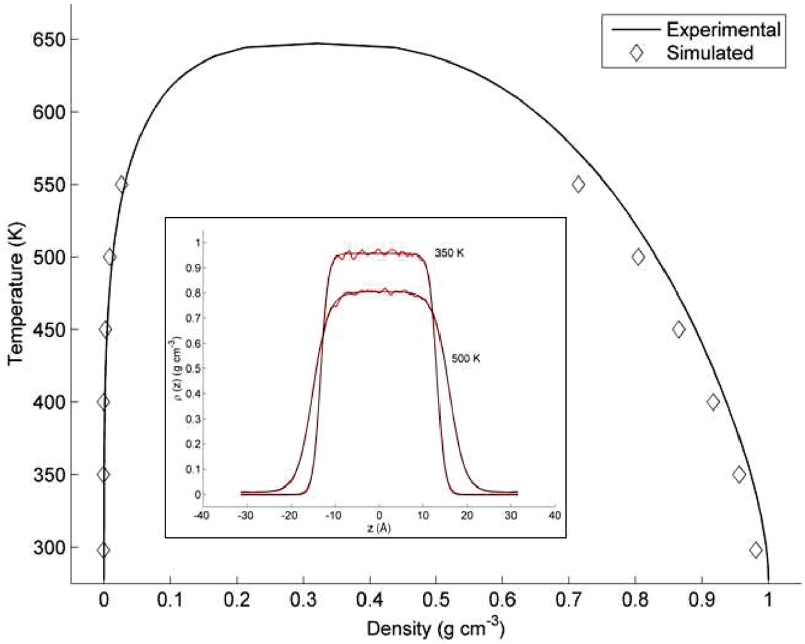


Figure 6. Water vapor−liquid equilibria. Experimental values from ref 61. The inset shows the *z*-density profile for 350 and 500 K; the molecular dynamics results are shown in red and the fitted tangent hyperbolic function in black (see also the Supporting Information).

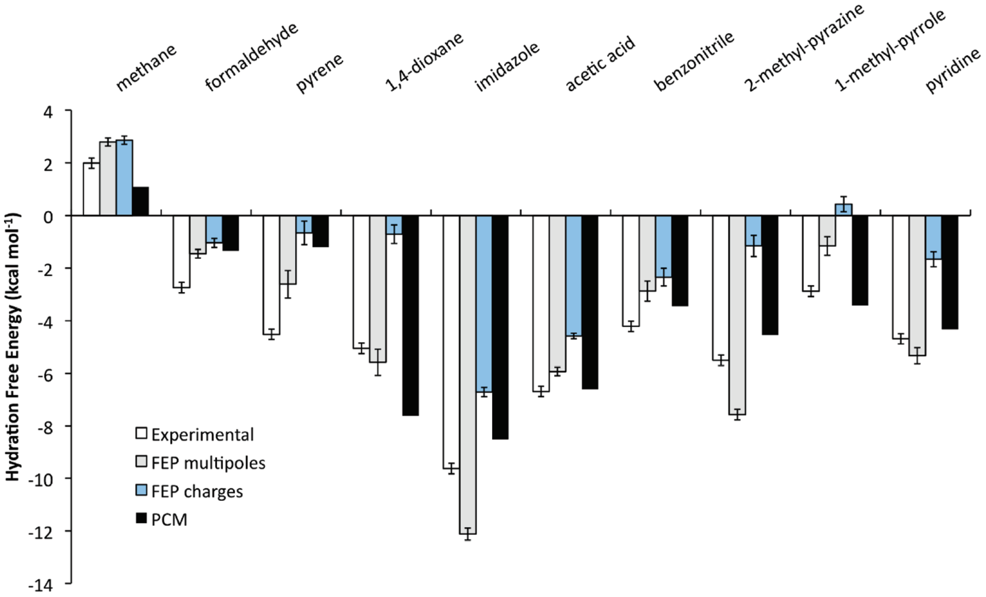


Figure 7. Predicted vs experimental[97] hydration free energies at ambient conditions using free energy perturbation (FEP) and a self-consistent reaction field method using the polarizable continuum model (PCM). Error bars in the FEP calculations correspond to one standard deviation computed by splitting the simulation into five equal parts; the error in experimental measurements was set to a nominal 0.2 kcal mol^{-1} value.[13]

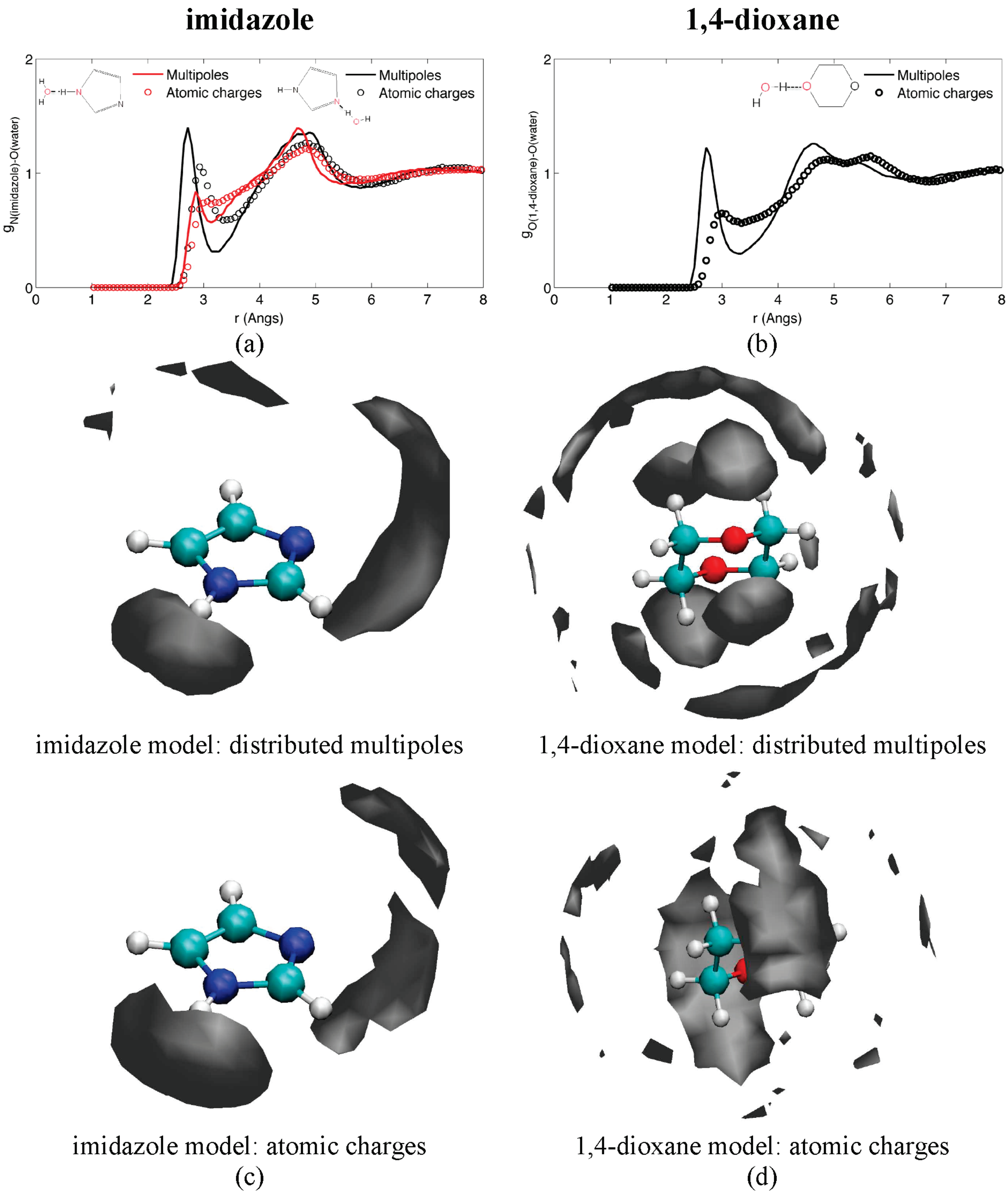
Changing the charge density representation of the solute from a distributed multipole to an atomic charge model and keeping all other potential and simulation parameters identical leads to more positive free energies of hydration for all molecules. However, the effect is very molecule-dependent and varies from +0.07 kcal mol^{-1} for methane to +6.42 kcal mol^{-1} for 2-methyl-pyrazine. With the atomic charge model, the RMS error of the hydration free energies increases to 3.05 kcal mol^{-1}, while the free energy of hydration of 1-methyl-pyrrole is predicted to have the wrong sign.

The quality of the solute's charge density representation has a profound effect on the calculated free energies of hydration that should be reflected in qualitative differences in the hydrogen bonding between the water and the solute's hydrogen bond donors and acceptors. Figure 8a,b shows the site−site radial distribution functions for water oxygen with imidazole's nitrogen and 1,4-dioxane's oxygen atoms as obtained in a 1 ns isothermal, isobaric molecular dynamics run at ambient conditions with the unperturbed water−solute interactions. The use of distributed multipoles results in more pronounced first peaks and hence more directional, spatially confined hydrogen bonding. These differences in solute−water hydrogen bonding propagate to the second hydration shell before leveling off. The gray regions in Figure 8c,d show the areas of the first hydration shell of 1,4-dioxane and imidazole in which the number density of water oxygen atoms is equal to 3.3 times the average number density in liquid water at ambient conditions. It is clear that, when 1,4-dioxane's electrostatic interactions are modeled with multipole moments, hydrogen bonding is found more localized at the oxygen lone pairs, compared with the more scattered and less directional hydrogen bonding contacts with atomic charges. On the other hand, the isodensity surface of imidazole's first hydration shell appears to depend only weakly on the electrostatic model. However, for both solutes, the use of a distributed multipole model results in significantly wider ranges of number densities for the water oxygen atoms in the first hydration shell compared to the overall more isotropic distribution of surrounding water with atomic charges.

We have finally computed the probability distribution of water-exchange times in the first hydration shell and in the vicinity of the solute's hydrogen bond donors and acceptors. By fitting the obtained function with an exponentially decaying function $p(t) = A \exp(-t/\tau)$, we estimate the residence time of water molecules in the first hydration shell. When multipole moments are used, τ is found to be 1.7 ± 0.2 ps for 1,4-dioxane's oxygen acceptor and 1.4 ± 0.2 ps for imidazole's nitrogen acceptor. The corresponding residence times when atomic charges are used are 3−4 times shorter, which suggests that the dynamic properties of the solute's hydration also depend strongly on the approach with which the electrostatic interactions are modeled.

Discussion

The first part of this study is concerned with the development of a rigid-body water model that comprises a quantum-mechanically derived distributed multipole representation of the charge density, which includes average polarization effects in liquid water at ambient conditions. Parameterization was limited to the repulsion−dispersion potential, which is fitted to the experimental liquid water density and oxygen−oxygen radial distribution function at ambient conditions. Despite restricting the fitting to a very narrow set of experimental data and to only one temperature, the model is found to be successful in modeling a wide range of liquid water and ordered ice properties, including the notably elusive temperature of maximum density at 1 atm and the vapor−liquid equilibrium densities. This contrasts simpler water models that employ a monopole representation of the charge density,[7,44,88] the computational efficiency of which



imidazole model: distributed multipoles

1,4-dioxane model: distributed multipoles

imidazole model: atomic charges

1,4-dioxane model: atomic charges

(c)

(d)

Figure 8. Site–site correlation functions demonstrating the differences in hydration environment of (a) imidazole's and (b) 1,4-dioxane's hydrogen bond donors and acceptors when the solute's charge density is modeled with distributed multipoles up to rank 4 (continuous lines) and atomic charges (open circles). (c and d) Isodensity surfaces corresponding to 3.3 times the average number density of water oxygen atoms in liquid water at ambient conditions for the first hydration shell of imidazole and 1,4-dioxane, respectively.

allows the parametrization of all potential parameters, including the charges and the position of off-nuclei interaction sites, using a much wider range of experimentally measured quantities. Despite its greater computational cost in molecular simulation, the derivation of a multipole model for the dominant electrostatic forces from first principles simplifies the paramerization procedure by reducing the number of independent variables that need to be optimized. Finally, the partial reliance on quantum mechanical calculations, instead of fitting all water potential terms to bulk water properties, is likely to increase the transferability[98] of the water model in hydration simulations.

The dipole moment of the proposed water model is 2.22 D compared with the 1.85 D dipole moment in the gas phase and our computed 2.59 D average dipole moment of liquid water at ambient conditions. By using the effective multipole moments $Q_{\text{eff}} = Q + \Delta Q/2$ we account for the cost of polarizing the water charge density from the gas to the liquid state and alleviate the need to include a self-polarization correction in the predicted vaporization enthalpy. The inclusion of a self-polarization correction in the vaporization enthalpy has been an important aspect in developing water models, such as in the derivation of SPC/E from SPC[76] and TIP4P/2005[6] from TIP4P.[42] TIP4P and TIP4P/2005 have a dipole moment[88] of 2.18 and 2.31 D, respectively, that is significantly smaller than the liquid water dipole moment in our calculations and other studies,[51,52] and despite the uncertainty in the average induced dipole moment in liquid water $\Delta\mu$,[53] they most likely approximate the $\mu + \Delta\mu/2$ value. This suggests that the self-polarization correction to the predicted vaporization enthalpy with these models may also be not applicable. This reflects the difficulty in developing a nonexplicitly polarizable, classical water model to predict accurately water properties without overestimating the vaporization enthalpy. Similarly, it has been shown that the melting temperature of ice Ih and the vaporization enthalpy and temperature of maximum density cannot be simultaneously predicted,[88] although, once more, no physical explanation is obvious apart from the limitations of the underlying water model. It has been suggested[21] that hydration free energies predicted with fixed (prepolarized) electrostatic models should also be corrected to account for the energy cost associated with the polarization of the solute by the electric field of the solvent. In this work, we do not, however, correct the calculated free energies of hydration, because the solute was also modeled using the effective $Q_{\mathrm{eff}} = Q + \Delta Q/2$ multipole moments.

Using a high-rank, multipole description of the electrostatic forces, the hydration energies of 10 solutes with diverse chemistries are predicted with an RMS error of 1.50 kcal mol^{-1}, which is similar to the error with the self-consistent reaction field method that has been parametrized for this task. A comparative test[13] of implicit- and explicit-solvent free energy approaches for 17 small solutes gave an RMS error in predicted ΔG_{hyd} that ranged between 1.3 and 2.6 kcal mol^{-1}. These results are consistent with a recent blind test,[99] which showed that the solvation free energy of complex, drug-like molecules can at present be predicted with an RMS error of 2.5−3.5 kcal mol^{-1}. In light of these comparative tests, and despite the limited size of the molecules, our predictions are encouraging given that the solute's repulsion −dispersion interactions have not been parametrized for hydration calculations,[98,100] nor in conjunction with explicit or implicit modeling of induction, which leaves a lot of scope for improvement. Indeed, using the effective multipole moments to account for polarization, we predict that the free energies of hydration of imidazole and 1,4-dioxane are 2.5 and 0.6 kcal mol^{-1} more negative compared to the experimental values, respectively. On the other hand, when the solutes are modeled with their gas phase multipole moments, the predicted hydration free energies are 1.8 and 1.7 kcal mol^{-1} more positive compared to experimental results. Hence, the errors of two predictions have opposite sign and show the sensitivity of the hydration free energies to the model for the intermolecular forces. Despite the challenge in predicting ΔG_{hyd} to a target accuracy of 1 $kT = 0.5$ kcal mol^{-1} or better, explicit-solvent free energy methods have the advantage that they can be systematically improved by employing theoretically justified models for the intermolecular forces derived from first principles.[101] In contrast, dielectric continuum methods depend strongly on parameters that have little physical meaning, such as the model for the solute's cavity, the permittivity at the cavity boundary, and the modeling of the nonelectrostatic contributions to ΔG_{hyd},[15] that limit their predictive power for molecules dissimilar from the training set. Unfortunately, the scope for more extensive parametrizations of such methods is limited because experimental data for hydration free energies are sparse,[82] especially for complex, polyfunctional molecules for which predictions are mostly needed.

A comparative, explicit-solvent hydration free energy study with a diverse set of charge models showed that agreement with experimental results does not improve with an increasing level of quantum theory to compute the atomic charges.[22] Our study demonstrates that representing the solute's charge density with an atomic charge model changes the hydration free energy by as much as 6 kcal mol^{-1} compared with a distributed multipole expansion, despite the two models being derived from the same wave function calculation. The discrepancy is greater for very polar molecules that have the most negative free energy of hydration. Hence, the inherent limitations of the monopole model are sufficiently large to suggest that the accurate modeling of electrostatic interactions in dynamic simulations should be a higher priority research goal and perhaps precede the development of isotropic polarizable models. We accept the view that the effect of modeling the electrostatic interactions with an atomic charge model can be partially absorbed in the repulsion−dispersion parametrization. However, the success of this strategy is likely to be limited for a diverse set of chemical functionalities, because the mathematical form of the models for the two intermolecular energy contributions is very different: electrostatic interactions are long-ranged and highly dependent on the relative molecular orientation. We note that the effect of including higher multipole moments may not always be evident in structural reproduction: the isotropic models TIP4P and TIP4P/2005 achieve comparable accuracy with our model in reproducing the structure of ice polymorphs and liquid water. The difference between a multipole and monopole representation of the charge density is more likely to be manifested in evaluating the relative stability of different molecular arrangements, including the relative energy and dynamics of a solute in a solvent and in a vacuum, as shown by the sensitivity of the predicted hydration free energy to the electrostatic model. Similar strong dependence on the electrostatic model has been well established in predicting the relative stability of different packing arrangements of organic molecules.[102,103]

Our ultimate goal is the prediction of the aqueous solubility of crystalline materials, which is defined as the solution concentration for which the chemical potential of the solute in solution and in the solid state are equal. The solubility depends on the relative strength of the solute's intermolecular interactions in the solid state and in solution.[104] Hence, a computationally viable route for computing solubility is through the thermodynamic cycle *crystal structure* → *gas* *phase* → *solution* that involves the calculation of the free energy of solvation and sublimation free energy of the crystal.[11] The latter can be computed using anisotropic model potentials within the harmonic approximation[35] and using the Einstein crystal methodology[84] at elevated temperatures compared with the melting point. In this first publication, we investigated the possibility of computing the hydration free energy using explicit-solvent free energy perturbation and the same anisotropic potentials we have been developing for modeling the crystal structure of organic molecules.[33] It is expected that computing the hydration free energy and sublimation free energy of the crystal using the same accurate model for the intermolecular forces would be advantageous compared with approaches that mix classical lattice energy with self-consistent reaction field quantum mechanical calculations.[11] Our results show that the prediction of solubility using such thermodynamic cycles would require further improvements in the models for the intermolecular forces, given that an error of 1.5 kcal mol^{-1} in hydration free energy alone would cause a discrepancy in the predicted solubility that is comparable with the accuracy of statistical QSPR methods[105,106] to predict solubility, which have negligible computational cost.

Conclusions

We present a rigid-body, implicitly polarized water model based on a high-rank, distributed multipole representation of the quantum-mechanically computed water charge density, which was computed to include average polarization effects in liquid water. The repulsion−dispersion water−water interactions are modeled with an exp-6 potential fitted only to the experimental density and oxygen−oxygen site correlation function of liquid water at ambient conditions. The model performs well in modeling a wide range of water properties not used in its parametrization, including the heat capacity, diffusion coefficient, density maximum of liquid water and vapor−liquid phase equilibria data. This water model was used in explicit-solvent free energy perturbation calculations to compute the hydration free energy of 10 organic solutes. The solute−water interactions were also modeled with an implicitly polarized, distributed multipole model and an empirical exp-6 repulsion−dispersion potential parametrized for organic crystal structures in conjunction with distributed multipoles. The root-mean-square error of the predicted hydration free energies is 1.50 kcal mol^{-1}, which is comparable with the accuracy of a self-consistent reaction field model that had been parametrized explicitly for this task. The free energy of hydration was found to be particularly sensitive to the accuracy of the model for the intermolecular electrostatic forces. Representing the solute's charge density using an atomic charge model changes the predicted hydration free energy by up to 6 kcal mol^{-1} compared with a distributed multipole model computed at the same level of theory. The discrepancy between the two models is very molecule-dependent and provides an estimate for the effect of the modeling quality of the intermolecular electrostatic forces in hydration free energy calculations.

Acknowledgment. P.G.K. thanks Dr. Carlos Avendaño (Imperial College) for useful discussions and Mr. Simon Burbidge for assistance with high-performance computing. Funding to the Molecular Systems Engineering group from the Engineering and Physical Sciences Research Council (EPSRC) of the U.K. [EP/E016340] is gratefully acknowledged. The majority of calculations were performed on the High Performance Computing Cluster of Imperial College. P.R. thanks the Australian Research Council for financial support through a Discovery grant (DP0986999) and NCI and iVEC for the provision of computing resources.

Supporting Information Available: The proposed model for the water−water intermolecular electrostatic and repulsion−dispersion interactions, detailed water vapor−liquid equilibria results, the hydration free energy of 1,4-dioxane and pyrene using a soft-core interaction potential for the solute−water repulsion−dispersion interactions, and the effect of long-range corrections in modeling the hydration free energy of 1,4-dioxane. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Guillot, B. A Reappraisal of What We Have Learnt During Three Decades of Computer Simulations on Water. *J. Mol. Liq.* **2002**, *101*, 219–260.

(2) Paricaud, P.; Predota, M.; Chialvo, A.; Cummings, P. From Dimer to Condensed Phases at Extreme Conditions: Accurate Predictions of the Properties of Water by a Gaussian Charge Polarizable Model. *J. Chem. Phys.* **2005**, *122*, art-244511.

(3) Ren, P.; Ponder, J. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.

(4) Lamoureux, G.; MacKerell, A.; Roux, B. A Simple Polarizable Model of Water Based on Classical Drude Oscillators. *J. Chem. Phys.* **2003**, *119*, 5185–5197.

(5) Yu, H.; van Gunsteren, W. Charge-on-Spring Polarizable Water Models Revisited: From Water Clusters to Liquid Water to Ice. *J. Chem. Phys.* **2004**, *121*, 9549–9564.

(6) Abascal, J.; Vega, C. A General Purpose Model for the Condensed Phases of Water: Tip4p/2005. *J. Chem. Phys.* **2005**, *123*, art-234505.

(7) Mahoney, M.; Jorgensen, W. A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential Functions. *J. Chem. Phys.* **2000**, *112*, 8910–8922.

(8) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; van der Avoird, A. Predictions of the Properties of Water from First Principles. *Science* **2007**, *315*, 1249–1252.

(9) Wolfenden, R. Waterlogged Molecules. *Science* **1983**, *222*, 1087–1093.

(10) Garrido, N. M.; Queimada, A. J.; Jorge, M.; Macedo, E. A.; Economou, I. G. 1-Octanol/Water Partition Coefficients of N-Alkanes from Molecular Simulations of Absolute Solvation Free Energies. *J. Chem. Theory Comput.* **2009**, *5*, 2436–2446.

(11) Palmer, D. S.; Llinas, A.; Morao, I.; Day, G. M.; Goodman, J. M.; Glen, R. C.; Mitchell, J. B. O. Predicting Intrinsic Aqueous Solubility by a Thermodynamic Cycle. *Mol. Pharm.* **2008**, *5*, 266–279.

(12) Westergren, J.; Lindfors, L.; Hoglund, T.; Luder, K.; Nordholm, S.; Kjellander, R. In Silico Prediction of Drug

Solubility: 1. Free Energy of Hydration. *J. Phys. Chem. B* **2007**, *111*, 1872–1882.

(13) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test for Computational Chemistry. *J. Med. Chem.* **2008**, *51*, 769–779.

(14) Hine, J.; Mookerjee, P. Intrinsic Hydrophilic Character of Organic Compounds - Correlations in Terms of Structural Contributions. *J. Org. Chem.* **1975**, *40*, 292–298.

(15) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Perspective on Foundations of Solvation Modeling: The Electrostatic Contribution to the Free Energy of Solvation. *J. Chem. Theory Comput.* **2008**, *4*, 877–887.

(16) 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.

(17) Shirts, M.; Pande, V. Comparison of Efficiency and Bias of Free Energies Computed by Exponential Averaging, the Bennett Acceptance Ratio, and Thermodynamic Integration. *J. Chem. Phys.* **2005**, *122*, art-144107.

(18) Jorgensen, W. L.; Thomas, L. L. Perspective on Free-Energy Perturbation Calculations for Chemical Equilibria. *J. Chem. Theory Comput.* **2008**, *4*, 869–876.

(19) Rodinger, T.; Pomes, R. Enhancing the Accuracy, the Efficiency and the Scope of Free Energy Simulations. *Curr. Opin. Struct. Biol.* **2005**, *15*, 164–170.

(20) Shirts, M.; Pande, V. Solvation Free Energies of Amino Acid Side Chain Analogs for Common Molecular Mechanics Water Models. *J. Chem. Phys.* **2005**, *122*, art-134508.

(21) Hess, B.; van der Vegt, N. F. A. Hydration Thermodynamic Properties of Amino Acid Analogues: A Systematic Comparison of Biomolecular Force Fields and Water Models. *J. Phys. Chem. B* **2006**, *110*, 17616–17626.

(22) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. Comparison of Charge Models for Fixed-Charge Force Fields: Small-Molecule Hydration Free Energies in Explicit Solvent. *J. Phys. Chem. B* **2007**, *111*, 2242–2254.

(23) Shivakumar, D.; Deng, Y.; Roux, B. Computations of Absolute Solvation Free Energies of Small Molecules Using Explicit and Implicit Solvent Model. *J. Chem. Theory Comput.* **2009**, *5*, 919–930.

(24) Chialvo, A.; Cummings, P. Simple Transferable Intermolecular Potential for the Molecular Simulation of Water over Wide Ranges of State Conditions. *Fluid Phase Equilib.* **1998**, *150*, 73–81.

(25) Burnham, C.; Xantheas, S. Development of Transferable Interaction Models for Water. III. Reparametrization of an All-Atom Polarizable Rigid Model (Ttm2-R) from First Principles. *J. Chem. Phys.* **2002**, *116*, 1500–1510.

(26) Liem, S. Y.; Popelier, P. L. A. Properties and 3d Structure of Liquid Water: A Perspective from a High-Rank Multipolar Electrostatic Potential. *J. Chem. Theory Comput.* **2008**, *4*, 353–365.

(27) Walsh, T. R.; Liang, T. A Multipole-Based Water Potential with Implicit Polarization for Biomolecular Simulations. *J. Comput. Chem.* **2009**, *30*, 893–899.

(28) Jiang, H.; Jordan, K. D.; Taylor, C. E. Molecular Dynamics Simulations of Methane Hydrate Using Polarizable Force Fields. *J. Phys. Chem. B* **2007**, *111*, 6486–6492.

(29) Liang, T.; Walsh, T. R. Simulation of the Hydration Structure of Glycyl-Alanine. *Mol. Simulat.* **2007**, *33*, 337–342.

(30) Coombes, D.; Price, S.; Willock, D.; Leslie, M. Role of Electrostatic Interactions in Determining the Crystal Structures of Polar Organic Molecules. A Distributed Multipole Study. *J. Phys. Chem.* **1996**, *100*, 7352–7360.

(31) Price, S.; Leslie, M.; Welch, G.; Habgood, M.; Ls, P.; Karamertzanis, P.; Day, G. Modelling Organic Crystal Structures Using Distributed Multipole and Polarizability-Based Model Intermolecular Potentials. Submitted, 2010.

(32) Price, S.; Andrews, J.; Murray, C.; Amos, R. The Effect of Basis Set and Electron Correlation on the Predicted Electrostatic Interactions of Peptides. *J. Am. Chem. Soc.* **1992**, *114*, 8268–8276.

(33) Price, S. L. Computational Prediction of Organic Crystal Structures and Polymorphism. *Int. Rev. Phys. Chem.* **2008**, *27*, 541–568.

(34) Day, G.; Price, S.; Leslie, M. Elastic Constant Calculations for Molecular Organic Crystals. *Cryst. Growth Des.* **2001**, *1*, 13–26.

(35) Day, G.; Price, S.; Leslie, M. Atomistic Calculations of Phonon Frequencies and Thermodynamic Quantities for Crystals of Rigid Organic Molecules. *J. Phys. Chem. B* **2003**, *107*, 10919–10933.

(36) Stone, A. Distributed Multipole Analysis: Stability for Large Basis Sets. *J. Chem. Theory Comput.* **2005**, *1*, 1128–1132.

(37) Welch, G. W. A.; Karamertzanis, P. G.; Misquitta, A. J.; Stone, A. J.; Price, S. L. Is the Induction Energy Important for Modeling Organic Crystals. *J. Chem. Theory Comput.* **2008**, *4*, 522–532.

(38) Misquitta, A. J.; Stone, A. J.; Price, S. L. Accurate Induction Energies for Small Organic Molecules. 2. Development and Testing of Distributed Polarizability Models against Sapt(Dft) Energies. *J. Chem. Theory Comput.* **2008**, *4*, 19–32.

(39) Handley, C. M.; Hawe, G. I.; Kell, D. B.; Popelier, P. L. A. Optimal Construction of a Fast and Accurate Polarisable Water Potential Based on Multipole Moments Trained by Machine Learning. *Phys. Chem. Chem. Phys.* **2009**, *11*, 6365–6376.

(40) Kastenholz, M. A.; Huenenberger, P. H. Computation of Methodology-Independent Ionic Solvation Free Energies from Molecular Simulations. Ii. The Hydration Free Energy of the Sodium Cation. *J. Chem. Phys.* **2006**, *124*, art-224501.

(41) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Dill, K. A. Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations. *J Phys Chem B.* **2009**, *113*, 4533–4537.

(42) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(43) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Montgomery, J.; Vreven, T.; Kudin, K.; Burant, J.; Millam, J.; Iyengar, S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G.; 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.; Hratchian, H.; Cross, J.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R.; Yazyev, O.; Austin, A.; Cammi, R.; Pomelli, C.; Ochterski, J.; Ayala, P.; Morokuma, K.; Voth,

G.; Salvador, P.; Dannenberg, J.; Zakrzewski, V.; Dapprich, S.; Daniels, A.; Strain, M.; Farkas, O.; Malick, D.; Rabuck, A.; Raghavachari, K.; Foresman, J.; Ortiz, J.; Cui, Q.; Baboul, A.; Clifford, S.; Cioslowski, J.; Stefanov, B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R.; Fox, D.; Keith, T.; Al Laham, M.; Peng, C.; Nanayakkara, A.; Challacombe, M.; Gill, P.; Johnson, B.; Chen, W.; Wong, M.; Gonzalez, C.; Pople, J. *Gaussian 03*; Gaussian Inc.: Wallingford, CT, 2003.

(44) Horn, H.; Swope, W.; Pitera, J.; Madura, J.; Dick, T.; Hura, G.; Head-Gordon, T. Development of an Improved Four-Site Water Model for Biomolecular Simulations: Tip4p-Ew. *J. Chem. Phys.* **2004**, *120*, 9665–9678.

(45) Liem, S.; Popelier, P.; Leslie, M. Simulation of Liquid Water Using a High-Rank Quantum Topological Electrostatic Potential. *Int. J. Quantum Chem.* **2004**, *99*, 685–694.

(46) Stone, A. J. Water from First Principles. *Science* **2007**, *315*, 1228–1229.

(47) Misquitta, A.; Stone, A. Distributed Polarizabilities Obtained Using a Constrained Density-Fitting Algorithm. *J. Chem. Phys.* **2006**, *124*, art-024111.

(48) Millot, C.; Soetens, J.; Costa, M.; Hodges, M.; Stone, A. Revised Anisotropic Site Potentials for the Water Dimer and Calculated Properties. *J. Phys. Chem. A* **1998**, *102*, 754–770.

(49) Tu, Y.; Laaksonen, A. The Electronic Properties of Water Molecules in Water Clusters and Liquid Water. *Chem. Phys. Lett.* **2000**, *329*, 283–288.

(50) Breneman, C.; Wiberg, K. Determining Atom-Centered Monopoles from Molecular Electrostatic Potentials - the Need for High Sampling Density in Formamide Conformational-Analysis. *J. Comput. Chem.* **1990**, *11*, 361–373.

(51) Gregory, J.; Clary, D.; Liu, K.; Brown, M.; Saykally, R. The Water Dipole Moment in Water Clusters. *Science* **1997**, *275*, 814–817.

(52) Silvestrelli, P.; Parrinello, M. Water Molecule Dipole in the Gas and in the Liquid Phase. *Phys. Rev. Lett.* **1999**, *82*, 3308–3311.

(53) Handley, C. M.; Popelier, P. L. A. The Asymptotic Behavior of the Dipole and Quadrupole Moment of a Single Water Molecule from Gas Phase to Large Clusters: A Qct Analysis. *Synth. React. Inorg. Met.* **2008**, *38*, 91–99.

(54) Stone, A. J. *Orient*, 4.6; University of Cambridge: Cambridge, 2009. http://www-stone.ch.cam.ac.uk/programs.html (accessed Mar 2010).

(55) Stone, A. *The Theory of Intermolecular Forces*; Clarendon Press: Oxford, 1996.

(56) Leslie, M. Dl_Multi - a Molecular Dynamics Program to Use Distributed Multipole Electrostatic Models to Simulate the Dynamics of Organic Crystals. *Mol. Phys.* **2008**, *106*, 1567–1578.

(57) Soper, A. The Radial Distribution Functions of Water and Ice from 220 to 673 K and at Pressures up to 400 MPa. *Chem. Phys.* **2000**, *258*, 121–137.

(58) de la Pena, L.; Kusalik, P. Quantum Effects in Light and Heavy Liquid Water: A Rigid-Body Centroid Molecular Dynamics Study. *J. Chem. Phys.* **2004**, *121*, 5992–6002.

(59) Mahoney, M.; Jorgensen, W. Quantum, Intramolecular Flexibility, and Polarizability Effects on the Reproduction of the Density Anomaly of Liquid Water by Simple Potential Functions. *J. Chem. Phys.* **2001**, *115*, 10758–10768.

(60) Allesch, M.; Schwegler, E.; Gygi, F.; Galli, G. A First Principles Simulation of Rigid Water. *J. Chem. Phys.* **2004**, *120*, 5192–5198.

(61) Lemmon, E.; McLinden, M.; Friend, D. Thermophysical Properties of Fluid Systems. In *NIST Chemistry Webbook, NIST Standard Reference Database Number 69*; Linstrom, P., Mallard, W., Eds.; National Institute of Standars and Technology: Gaithersburg, MD. http://webbook.nist.gov (retrieved September 20, 2009).

(62) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999–3093.

(63) Haslam, A. J.; Galindo, A.; Jackson, G. Prediction of Binary Intermolecular Potential Parameters for Use in Modelling Fluid Mixtures. *Fluid Phase Equilib.* **2008**, *266*, 105–128.

(64) Leadbetter, A.; Ward, R.; Clark, J.; Tucker, P.; Matsuo, T.; Suga, H. The Equilibrium Low-Temperature Structure of Ice. *J. Chem. Phys.* **1985**, *82*, 424–428.

(65) Kamb, B.; Hamilton, W.; Laplaca, S.; Prakash, A. Ordered Proton Configuration in Ice-Ii, from Single-Crystal Neutron Diffraction. *J. Chem. Phys.* **1971**, *55*, 1934–1945.

(66) Laplaca, S.; Hamilton, W.; Kamb, B.; Prakash, A. Nearly Proton-Ordered Structure for Ice Ix. *J. Chem. Phys.* **1973**, *58*, 567–580.

(67) Kuhs, W.; Finney, J.; Vettier, C.; Bliss, D. Structure and Hydrogen Ordering in Ice-Vi, Ice-Vii, and Ice-Viii by Neutron Powder Diffraction. *J. Chem. Phys.* **1984**, *81*, 3612–3623.

(68) Salzmann, C.; Radaelli, P.; Hallbrucker, A.; Mayer, E.; FINNEY, J. The Preparation and Structures of Hydrogen Ordered Phases of Ice. *Science* **2006**, *311*, 1758–1761.

(69) Willock, D.; Price, S.; Leslie, M.; Catlow, C. The Relaxation of Molecular-Crystal Structures Using a Distributed Multipole Electrostatic Model. *J. Comput. Chem.* **1995**, *16*, 628–647.

(70) Chisholm, J.; Motherwell, S. Compack: A Program for Identifying Crystal Structure Similarity Using Distances. *J. Appl. Crystallogr.* **2005**, *38*, 228–231.

(71) Allen, F.; Taylor, R. Research Applications of the Cambridge Structural Database (Csd). *Chem. Soc. Rev.* **2004**, *33*, 463–475.

(72) Mootz, D.; Wussow, H. A Novel Proton-Ordered, Two-Dimensional Cross-Linking of Water-Molecules in Pyridine Trihydrate. *Angew. Chem., Int. Ed.* **1980**, *19*, 552–553.

(73) Mootz, D.; Wussow, H. Crystal-Structures of Pyridine and Pyridine Trihydrate. *J. Chem. Phys.* **1981**, *75*, 1517–1522.

(74) Allen, M.; Tildesley, D. *Computer Simulation of Liquids*; Oxford University Press: New York, 1992.

(75) Berendsen, H.; Grigera, J.; Straatsma, T. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(76) Alejandre, J.; Tildesley, D.; Chapela, G. Molecular-Dynamics Simulation of the Orthobaric Densities and Surface-Tension of Water. *J. Chem. Phys.* **1995**, *102*, 4574–4583.

(77) Vega, C.; de Miguel, E. Surface Tension of the Most Popular Models of Water by Using the Test-Area Simulation Method. *J. Chem. Phys.* **2007**, *126*, art-154707.

(78) Deng, Y.; Roux, B. Hydration of Amino Acid Side Chains: Nonpolar and Electrostatic Contributions Calculated from Staged Molecular Dynamics Free Energy Simulations with

Explicit Water Molecules. *J. Phys. Chem. B* **2004**, *108*, 16567–16576.

(79) Bennett, C. Efficient Estimation of Free-Energy Differences from Monte-Carlo Data. *J. Comput. Phys.* **1976**, *22*, 245–268.

(80) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. Energies, Structures, and Electronic Properties of Molecules in Solution with the C-Pcm Solvation Model. *J. Comput. Chem.* **2003**, *24*, 669–681.

(81) Cramer, C. J.; Truhlar, D. G. A Universal Approach to Solvation Modeling. *Acc. Chem. Res.* **2008**, *41*, 760–768.

(82) Guthrie, J.; Povar, I. A Test of Various Computational Solvation Models on a Set Of "Difficult" Organic Compounds. *Can. J. Chem.* **2009**, *87*, 1154–1162.

(83) Jorgensen, W.; Blake, J.; Buckner, J. Free-Energy of Tip4p Water and the Free-Energies of Hydration of Ch4 and Cl- from Statistical Perturbation-Theory. *Chem. Phys.* **1989**, *129*, 193–200.

(84) Frenkel, D.; B., S. *Understanding Molecular Simulation*; Academic Press: San Diego, 2002.

(85) Wescott, J.; Fisher, L.; Hanna, S. Use of Thermodynamic Integration to Calculate the Hydration Free Energies of N-Alkanes. *J. Chem. Phys.* **2002**, *116*, 2361–2369.

(86) Hulme, A. T.; Price, S. L. Toward the Prediction of Organic Hydrate Crystal Structures. *J. Chem. Theory Comput.* **2007**, *3*, 1597–1608.

(87) Fortes, A.; Wood, I.; Alfredsson, M.; Vocadlo, L.; Knight, K. The Incompressibility and Thermal Expansivity of D2o Ice Ii Determined by Powder Neutron Diffraction. *J. Appl. Crystallogr.* **2005**, *38*, 612–618.

(88) Vega, C.; Abascal, J. L. F.; Conde, M. M.; Aragones, J. L. What Ice Can Teach Us About Water Interactions: A Critical Comparison of the Performance of Different Water Models. *Faraday Discuss.* **2009**, *141*, 251–276.

(89) Kell, G. Precise Representation of Volume Properties of Water at One Atmosphere. *J. Chem. Eng. Data* **1967**, *12*, 66–69.

(90) Jorgensen, W.; Tirado-Rives, J. Potential Energy Functions for Atomic-Level Simulations of Water and Organic and Biomolecular Systems. *Proc. Natl. Acad. Sci U. S. A.* **2005**, *102*, 6665–6670.

(91) Hare, D.; Sorensen, C. The Density of Supercooled Water 0.2. Bulk Samples Cooled to the Homogeneous Nucleation Limit. *J. Chem. Phys.* **1987**, *87*, 4840–4845.

(92) Jancso, G.; Vanhook, W. Condensed Phase Isotope-Effects (Especially Vapor-Pressure Isotope-Effects). *Chem. Rev.* **1974**, *74*, 689–750.

(93) Billeter, S.; King, P.; Vangunsteren, W. Can the Density Maximum of Water Be Found by Computer-Simulation. *J. Chem. Phys.* **1994**, *100*, 6692–6699.

(94) Krynicki, K.; Green, C.; Sawyer, D. Pressure and Temperature-Dependence of Self-Diffusion in Water. *Faraday Discuss.* **1978**, *66*, 199–208.

(95) Mills, R. Self-Diffusion in Normal and Heavy-Water in Range 1−45 Degrees. *J. Phys. Chem.* **1973**, *77*, 685–688.

(96) Hermans, J.; Pathiaseril, A.; Anderson, A. Excess Free-Energy of Liquids from Molecular-Dynamics Simulations - Application to Water Models. *J. Am. Chem. Soc.* **1988**, *110*, 5982–5986.

(97) Rizzo, R.; Aynechi, T.; Case, D.; Kuntz, I. Estimation of Absolute Free Energies of Hydration Using Continuum Methods: Accuracy of Partial, Charge Models and Optimization of Nonpolar Contributions. *J. Chem. Theory Comput.* **2006**, *2*, 128–139.

(98) Docherty, H.; Galindo, A.; Vega, C.; Sanz, E. A Potential Model for Methane in Water Describing Correctly the Solubility of the Gas and the Properties of the Methane Hydrate. *J. Chem. Phys.* **2006**, *125*, art-074510.

(99) Guthrie, J. P. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. *J. Phys. Chem. B* **2009**, *113*, 4501–4507.

(100) Oostenbrink, C.; Villa, A.; Mark, A.; van Gunsteren, W. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The Gromos Force-Field Parameter Sets 53a5 and 53a6. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

(101) Misquitta, A. J.; Welch, G. W. A.; Stone, A. J.; Price, S. L. A First Principles Prediction of the Crystal Structure of C6br2clfh2. *Chem. Phys. Lett.* **2008**, *456*, 105–109.

(102) Karamertzanis, P. G.; Kazantsev, A. V.; Issa, N.; Welch, G. W. A.; Adjiman, C. S.; Pantelides, C. C.; Price, S. L. Can the Formation of Pharmaceutical Cocrystals Be Computationally Predicted? 2. Crystal Structure Prediction. *J. Chem. Theory Comput.* **2009**, *5*, 1432–1448.

(103) Day, G.; Motherwell, W.; Jones, W. Beyond the Isotropic Atom Model in Crystal Structure Prediction of Rigid Molecules: Atomic Multipoles Versus Point Charges. *Cryst. Growth Des.* **2005**, *5*, 1023–1033.

(104) Huang, L.; Tong, W. Impact of Solid State Properties on Developability Assessment of Drug Candidates. *Adv. Drug Delivery Rev.* **2004**, *56*, 321–334.

(105) Johnson, S. R.; Chen, X.-Q.; Murphy, D.; Gudmundsson, O. A Computational Model for the Prediction of Aqueous Solubility That Includes Crystal Packing, Intrinsic Solubility, and Ionization Effects. *Mol. Pharm.* **2007**, *4*, 513–523.

(106) Delaney, J. Predicting Aqueous Solubility from Structure. *Drug Discovery Today.* **2005**, *10*, 289–295.

CT900693Q

JCTC Journal of Chemical Theory and Computation

Rapid Prediction of Solvation Free Energy. 1. An Extensive Test of Linear Interaction Energy (LIE)

Traian Sulea, Christopher R. Corbeil, and Enrico O. Purisima*

Biotechnology Research Institute, National Research Council Canada,
6100 Royalmount Avenue, Montreal, Quebec H4P 2R2, Canada

Received November 17, 2009

Abstract: The present study provides a comprehensive systematic analysis on the applicability of the linear interaction energy (LIE) approximation to the prediction of gas-to-water transfer (hydration) free energy. The study is based on molecular dynamics simulations in explicit solvent for an extensive and diverse hydration data set comprising 564 neutral compounds with measured hydration free energies, including a "traditional" data set and the more challenging drug-like SAMPL1 data set. A highly correlative LIE model was achieved without empirical scaling of the solute−solvent interaction energy terms along with a cavity term calibrated to the experiment. This model was particularly accurate for the "traditional" data set and of acceptable accuracy for the SAMPL1 data set, with mean-unsigned-errors below 1 kcal/mol and slightly above 2 kcal/mol, respectively. We have analyzed the sensitivity of the LIE model to several parameters such as continuum correction terms applied outside the explicit water shell, the impact of various charging methods, the applicability of single-conformer representation of the solute, and the inclusion of internal energy terms. The parameters with the greatest sensitivity are the charging methods used, with AM1BCC-SP (without AM1 geometry optimization) charges favored over AM1BCC-OPT and RESP charges. The inclusion of the change in intramolecular van der Waals and electrostatic energies between the solution and gas phases can also lead to improved prediction accuracies. Functional group based error analysis identified several chemical classes as minor outliers with systematic errors. A direct comparison of the LIE and free energy perturbation (FEP) approaches using the same force field and charging method shows that the LIE approximation is at least as accurate as the FEP approach with a reduction of computing time by at least 1 order of magnitude.

Introduction

Hydration (aqueous solvation) of molecules plays an important role in biological, chemical, and industrial processes, as exemplified by the change in hydration upon complex formation. This change in hydration is a critical component of binding affinities in aqueous solution;[1−3] hence, simulation methods that predict absolute binding free energies require accurate hydration models. Implementation of accurate hydration models in scoring functions would benefit promising applications such as virtual screening in drug discovery. Over the years, much effort has been expended in developing and parametrizing solvation models at various levels of theory.[4−10] Accurate prediction of hydration free energies still remains a challenge for computational methods, as underscored by scientific community efforts like the SAMPL1 blind prediction challenge organized recently by OpenEye, Inc.[11] In general, simulation methods fall into two groups depending on whether they treat the water implicitly or explicitly.

Implicit hydration models (also known as continuum models) have the benefit of speed, but they break down on describing effects that relate to the discrete nature of water

* To whom correspondence should be addressed. Phone: 514-496-6343. Fax: 514-496-5143. E-mail: enrico.purisima@cnrc-nrc.gc.ca.

10.1021/ct9006025 Published 2010 by the American Chemical Society
Published on Web 04/02/2010

due to ordering in the first solvation shell around solutes, e.g., charge asymmetry or dependence of ion pairing on molecular shape.[12,13] Such effects are captured by explicit solvation models, which are more transferable than the continuum models but are generally prohibitively slow for many routine applications. The current challenge for the computational chemistry community is to develop hydration models that are as physics-based as the explicit models but have the speed of continuum models, as both accuracy and speed are required for practical applications in drug discovery. One such example is restoring charge asymmetry observed with explicit models within continuum models, which has been recently reported.[14] Even with these advances, implicit models still require further calibration on either experimental hydration free energies or computed energies from explicit-solvent simulations.

The free energy of solvation can be calculated in explicit solvent from molecular dynamics (MD) simulations using path methods like free energy perturbation (FEP) and thermodynamic integration (TI),[15] which require slow transformations between the end points of the hydration process (the gas phase and water-solvated state). Typically, such methods require long simulation times to reach convergence. Applications of the FEP approach to the hydration of small molecules have been recently reported with good success.[16−18] The linear interaction energy (LIE) is an attractive approximation of the rigorous full FEP methodology, where only the average interaction energies are calculated at the end-points of the process. This approximation has a significant impact on shortening the required MD simulation time. The LIE approximation has been applied to the problem of protein−ligand binding in solution with adjustable parameters for the interaction terms.[19−25] Linear response theory, which is deeply rooted in fundamental Gibbs inequalities, can provide remarkably accurate descriptions of the process of filling aqueous cavities with nonpolar, polar, or charged molecules.[26] Hydration of a small set of organic molecules has been studied previously with LIE in explicit solvent using Monte Carlo simulations and a calibrated cost of cavity formation[27] and was later shown not to require empirical scaling of the solute−solvent interaction terms.[24] More recently, linear response theory has been also applied to a hydration data set consisting of 194 simple neutral and ionic compounds. These semiempirical LIE models of hydration consisted of functional class-dependent empirical scaling of solute−solvent electrostatic interaction energies from fitting to simulated FEP data, and empirical scaling of the nonpolar term from fitting directly to the experimental data for the entire hydration data set.[28]

While the application of the LIE approximation to the calculation of hydration free energy is not a new idea, the present study is based on molecular dynamics simulations in explicit solvent for a significantly larger and more diverse hydration data set than previously analyzed. The set is comprised of 564 neutral compounds with measured hydration free energies, including 501 "traditional" simpler compounds used by Mobley et al.,[16,29] and the drug-like SAMPL1 data set.[11] The inclusion of the 63 highly diverse, densely polyfunctional, neutral polar compounds, which encompass larger magnitudes of hydration free energies and molecular weights, make the newer SAMPL1 testing data set more challenging than previous testing data sets.[11] The SAMPL1 blind challenge operated by first releasing the molecules to the public followed by the release of experimental transfer free energies after a few months. A number of continuum and explicit-solvent methods were tested, most of them in the prospective mode,[17,30−32] while results with two other methods were included retrospectively.[33,34]

We have analyzed the sensitivity of the LIE model to several parameters that we believe are important for the accuracy of the model and for calibration of a continuum model based on these results. These include continuum corrections to infinity, charging method, solute flexibility, and internal energy terms. We have opted not to fit the scaling parameters for the solute−solvent interaction energy terms to FEP data as in a previous study[28] but rather maintain the idealized theoretical values for these scaling parameters (1 for dispersive and 0.5 for electrostatic solute−solvent interactions) derived from the linear response theory.[24,26] By assuming an idealized linear response, we simplify the model and avoid the danger of model overfitting, while maintaining the theoretical rigor of a physically sound if albeit idealized model. It is a surprisingly good approximation. It is also important to compare the results from the LIE approach directly with those from the more rigorous FEP calculations. That is, the speedup advantage of LIE over FEP is attractive only if the accuracy of the prediction is maintained. This is possible since both data sets have been very recently studied with the FEP method with the same force field and charges,[16,17] thus allowing a direct comparison between LIE and FEP data.

A major motivation for this work is to generate the individual components of hydration free energy, electrostatics and van der Waals energies, both of which are experimentally inaccessible. This data will then be used to train the respective components of purely continuum solvation models with the aim of having this model mimic an explicit solvent simulation (see accompanying paper[35]).

Materials and Methods

Hydration Data Sets. A data set consisting of experimental hydration free energies for 501 neutral organic small molecules were taken from the Supporting Information provided by Mobley et al.[16] We followed the same approach of Mobley et al.,[16] not to include molecular ions in the study data set due to uncertainties in experimental data. The data set was split into a training set of 200 compounds used only for calibrating the cost of cavity formation in water and a testing data set of 301 compounds. In the training set, we included mostly rigid representatives of the various chemical classes, with the majority of compounds being monofunctional, and only a few polyfunctional compounds were included to increase coverage of some functional groups. The testing data set mirrors the training data set in terms of chemical class representation for monofunctional compounds but differs from the training analogs by having increased flexibility and containing a larger collection of polyfunctional compounds. The SAMPL1 data set[11] consists of 63 drug-like, diverse, polyfunctional, neutral polar compounds and spans a wider range of transfer free energies and molecular weights in comparison to the training and testing data sets. Details on the composition of the training and testing and SAMPL1 data sets are provided in the Supporting Information (Table S1).

Hydration Models. The following implementation of the LIE approach[19,20,24,26,27] was used to describe electrostatic, van der Waals, and cavity contributions to solvation:

$$\Delta G_{\mathrm{hyd}}^{\mathrm{LIE}} = \underbrace{\alpha(\langle E_{\mathrm{S-W}}^{\mathrm{Coul}}\rangle_{\leq 12\,\mathrm{Å}} + \langle G_{\mathrm{S}}^{\mathrm{RF}}\rangle_{12\,\mathrm{Å}\text{-}\infty})}_{\text{electrostatic}} + \underbrace{\beta(\langle E_{\mathrm{S-W}}^{\mathrm{vdW}}\rangle_{\leq 12\,\mathrm{Å}} + \langle E_{\mathrm{S}}^{\mathrm{cvdW}}\rangle_{12\,\mathrm{Å}\text{-}\infty})}_{\text{van der Waals}} + \underbrace{\gamma_{\mathrm{cav}}\langle MSA\rangle + C}_{\text{cavity}} \qquad (1)$$

where all terms represent averages (⟨...⟩) over snapshots from MD simulations. Here, $\langle E_{\mathrm{S-W}}^{\mathrm{Coul}}\rangle_{\leq 12\,\mathrm{Å}}$ and $\langle E_{\mathrm{S-W}}^{\mathrm{vdW1}}\rangle_{\leq 12\,\mathrm{Å}}$ are the average Coulomb and van der Waals interaction energies of the solute with explicit water molecules within 12 Å, respectively. We applied the ideal theoretical values of 0.5 and 1 for the α and β scaling factors of electrostatic and van der Waals average interaction energy terms,[26] respectively, in eq 1, shown to yield good predictions of experimental hydration free energies on a limited data set.[24,27] Following previous work,[27] the cavity contribution is expressed as a linear dependence on the average molecular surface area of the solute, ⟨MSA⟩, with the cavity surface coefficient, γ_{cav}, and a constant, C. These parameters were determined from a linear fit using Microsoft Excel to pseudoexperimental cavity energies obtained by subtracting the electrostatic and van der Waals contributions from experimental hydration free energies for the training data set.

Continuum-model correction terms were added outside the 12 Å shell of explicit water in order to capture the bulk solvent contribution to infinity. The average implicit electrostatic solvation outside the explicit water shell, $\langle G_{\mathrm{S}}^{\mathrm{RF}}\rangle_{12\,\mathrm{Å}-\infty}$, includes reaction field energy contributions from the interactions between (a) the solute charges and their induced surface-charge density, $G_{\mathrm{S}-(\sigma_{\mathrm{S}})_{12\,\mathrm{Å}}}^{\mathrm{RF}}$, (b) the solute charges and the induced surface charge density due to the explicit water-shell charges, $G_{\mathrm{S}-(\sigma_{\mathrm{W}})_{12\,\mathrm{Å}}}^{\mathrm{RF}}$, and (c) the explicit water-shell charges and the induced surface-charge density due to solute charges, $G_{\mathrm{W}-(\sigma_{\mathrm{S}})_{12\,\mathrm{Å}}}^{\mathrm{RF}}$ (eq 2).

$$\begin{aligned}\langle G_{\mathrm{S}}^{\mathrm{RF}}\rangle_{12\mathring{A}-\infty} &= \langle (G_{\mathrm{S}-(\sigma_{\mathrm{S}})_{12\mathring{A}}}^{\mathrm{RF}} + G_{\mathrm{S}-(\sigma_{\mathrm{W}})_{12\mathring{A}}}^{\mathrm{RF}} + G_{\mathrm{W}-(\sigma_{\mathrm{S}})_{12\mathring{A}}}^{\mathrm{RF}})\rangle \\ &= \langle (G_{\mathrm{SW}-(\sigma_{\mathrm{SW}})_{12\mathring{A}}}^{\mathrm{RF}} - G_{\mathrm{W}-(\sigma_{\mathrm{W}})_{12\mathring{A}}}^{\mathrm{RF}})\rangle \qquad (2)\end{aligned}$$

Operationally, this external electrostatic solvation term was calculated by subtracting the reaction field energy between the explicit water charges and their induced surface-charge density, $G_{\mathrm{W}-(\sigma_{\mathrm{W}})_{12\,\mathrm{Å}}}^{\mathrm{RF}}$ (calculated by setting solute charges to zero) from the entire reaction field energy of the solute plus the explicit water shell, $G_{\mathrm{SW}-(\sigma_{\mathrm{SW}})_{12\,\mathrm{Å}}}^{\mathrm{RF}}$. All induced surface-charge densities were calculated at the molecular surface of the 12 Å shell of explicit water solvating the solute at each particular MD snapshot, using the boundary element method (BEM) implemented in the BRI-BEM program.[36,37] All molecular surface calculations were carried out with a variable-radius probe.[38]

The continuum solute−solvent van der Waals interaction outside the explicit water shell, averaged over the MD trajectory, $\langle E_{\mathrm{S}}^{\mathrm{cvdW}}\rangle_{12\,\mathrm{Å}-\infty}$, was calculated as described by Floris et al.[39,40] Briefly, the discrete surrounding water molecules are replaced by a continuum of uniform density distribution, and the solute−solvent van der Waals interaction is taken to be proportional to the integral of the solute-continuum interaction over all of space. For ease of computation, the volume integral is transformed into a surface integral typically at the solute−solvent boundary defined by the solvent-accessible surface. Here, the surface integral was evaluated at the solvent-accessible surface around the 12 Å shell of explicit water. For the dispersion (attractive) component of the 6−12 Lennard-Jones potential, that leads to

$$\langle E_{\mathrm{S}}^{\mathrm{cvdW}}\rangle = -\rho_{\mathrm{N}} \sum_{i}^{\substack{\text{solute}\\\text{atoms}}} \sum_{j}^{\text{patches}} \frac{1}{3}\frac{B_{iw}}{d_{ij}^{6}}\mathbf{r}_{ij}\cdot\mathbf{n}_j\mathrm{SA}_j \qquad (3)$$

where $\mathbf{r}_{ij}$ is the vector from solute atom i to patch j of the solvent-accessible surface around the 12 Å shell of explicit water, $\mathbf{n}_j$ is the surface normal at j, SA_j is the area of j, and ρ_{N} is the solvent number density. The atomic dispersion parameters B_{iw} were taken from the TIP3P and GAFF force field[41] without scaling.

We also extended the LIE approach to include the difference in the average intramolecular energy of the solute in the aqueous phase, $E_{\mathrm{aq}}^{\mathrm{intra}}$, and in the gas phase, $E_{\mathrm{gas}}^{\mathrm{intra}}$:

$$\Delta G_{\mathrm{hyd}}^{\mathrm{LIE}} = \underbrace{\alpha(0.5\langle E_{\mathrm{S-W}}^{\mathrm{Coul}}\rangle_{\leq 12\,\mathrm{Å}} + \langle G_{\mathrm{S}}^{\mathrm{RF}}\rangle_{12\,\mathrm{Å}\text{-}\infty})}_{\text{electrostatic}} + \underbrace{\beta(\langle E_{\mathrm{S-W}}^{\mathrm{vdW}}\rangle_{\leq 12\,\mathrm{Å}} + \langle E_{\mathrm{S}}^{\mathrm{cvdW}}\rangle_{12\,\mathrm{Å}\text{-}\infty})}_{\text{van der Waals}} + \underbrace{\gamma_{\mathrm{cav}}\langle MSA\rangle + C}_{\text{cavity}} + \underbrace{\langle E_{\mathrm{aq}}^{\mathrm{intra}}\rangle - \langle E_{\mathrm{gas}}^{\mathrm{intra}}\rangle}_{\text{intramolecular}} \qquad (4)$$

where the intramolecular energies are comprised of molecular mechanics force field bond stretching, angle bending, torsional (including improper corrections), 1−4 and 1−5 van der Waals, and 1−4 and 1−5 Coulombic electrostatic energies. We also considered excluding the covalent terms (bond stretching, angle bending, and dihedral and improper torsions) from the intramolecular energy.

Explicit-Solvent Simulations and Solute Parameters. Single-conformation molecular geometries for the training, testing, and SAMPL1 data sets were downloaded from the appropriate sources[11,16] and refined by energy minimization with the Merck Molecular Force Field (MMFF94)[42] and a 4R distance-dependent dielectric constant, up to a gradient of 0.01 kcal mol^{-1} Å^{-1}, in SYBYL 8 (Tripos, Inc. St. Louis, MO). These geometries were then used to calculate partial atomic charges with different methods. Solute atomic partial charges were calculated with the AM1BCC method[43,44] implemented within QUACPAC (OpenEye, Inc., Santa Fe, NM), with AM1 geometry optimization (AM1BCC-OPT) or without AM1 geometry optimization (single-point calculation, AM1BCC-SP), as well as with the two-stage RESP method.[45,46] RESP atomic partial charges were fitted to the electrostatic potential from *in vacuo* single-point calculations at the 6-31G* level (3-21G level for iodine-containing compounds) using GAMESS.[47]

Molecular dynamics simulations were carried out using AMBER 9 software[48] with the general AMBER force field (GAFF) parameters[41] assigned to the solutes with PARMCHK and ANTECHAMBER,[49] and using the TIP3P water parameters[50,51] for the solvent. Solute molecules were solvated in a truncated octahedron of water extending 12 Å away from the solute. Applying harmonic restraints with force constants of 10 kcal mol^{-1} Å^{-2} to all solute atoms, the system was energy-minimized first, followed by heating from 100 K to 300 K over 25 ps in the canonical ensemble (NVT), and by equilibrating to adjust the solvent density under 1 atm of pressure over 25 ps in the isothermal−isobaric ensemble (NPT) simulation. The harmonic restraints were then reduced to zero, and a 2 ns production NPT run was obtained with snapshots collected every 10 ps, using a 2 fs time-step and 9 Å nonbonded cutoff. The Particle Mesh Ewald (PME) method[52] was used to treat long-range electrostatic interactions, and bond lengths involving bonds to hydrogen atoms were constrained by SHAKE.[53] Separate MD simulations were carried out for the different charging methods of the solute molecules. Also, separate MD simulations were carried out with unconstrained (flexible) or constrained (rigid) solute molecule during the 2 ns production phase. In the latter case, the harmonic potential of 10 kcal mol^{-1} Å^{-2} on the solute atoms was maintained during the course of the entire MD protocol. Similar MD protocols were applied to carry out gas-phase simulations for all molecules with full solute flexibility and for all charge models using a 2 fs time step. All averages were calculated for 100 snapshots at 10 ps intervals from the last nanosecond of the 2 ns trajectories.

Bootstrapped statistical analyses were carried out for 5000 samples using the boot library within the R software.[54] In order to help identify systematic prediction errors, functional group assignment for the compounds in the testing subset was carried out with CHECKMOL.[55] Since CHECKMOL only identifies if a group is present, further manual groupings were created for molecules that only contain one type of a functional group (see Table S2, Supporting Information).

Results and Discussion

We explored hydration free energy calculations within the LIE formalism based on MD simulations in explicit solvent on three data sets with measured water-to-vacuum transfer free energies: the more traditional training and testing data sets[16,29] and the challenging drug-like SAMPL1 data set.[11] We will start by calibrating an empirical term for the cost of cavity formation in water, which is not directly provided by the force-field energy terms in the LIE approach. The LIE results for the case of flexible solutes and the AM1BCC-SP charge model will be described first vis-à-vis experimental data. We will separately analyze the effects of (1) removing the bulk-solvent continuum correction to infinity, (2) changing the solute charging method, (3) using a single-conformation representation of the solute, and (4) incorporating the internal energy difference between the two media on the accuracy of LIE predictions of hydration free energies. Functional group analysis will then be carried out to single out systematic outliers and problematic groups and try to identify potential sources of errors. The LIE prediction accuracy and functional group analysis of errors will be compared with published FEP alchemical calculations that are more computationally demanding but are thought to be more accurate, or at least more rigorous. Values for LIE components and predicted hydration free energies for all molecules can found in Table S3 (Supporting Information).

Deriving the Cavity Contribution. The dispersive (van der Waals) and electrostatic (Coulombic) solute−solvent interactions are well described by linear response theory, which implies Gaussian distributions of the energy fluctuations associated with these interactions.[26] On the other hand, the formation of a cavity the size of a molecule in water is not a linear process, which means that cavity formation fluctuations are significantly non-Gaussian. The cost of cavity formation in water, mainly of entropic nature, should in principle be directly related to the size of the cavity. Since in the LIE approach this term is not captured by the force-field-based interaction energy terms, we sought a linear relationship to the molecular surface area (MSA) of the solute in order to calibrate the cavity term. We started with the LIE model described in eq 1 and MD simulations with fully flexible solutes and AM1BCC-SP atomic charges. We defined a pseudoexperimental cavity contribution as the residual between the experimental hydration free energy and the calculated electrostatic and van der Waals solute−solvent interaction energies. Using the training set of 200 compounds, we obtained a robust linear correlation between the pseudoexperimental cavity cost and the MSA (Figure 1A), with a bootstrapped squared correlation coefficient of 0.923 ± 0.015 and a slope and intercept (γ and C, respectively, in eq 1) of 0.108 ± 0.002 kcal mol^{-1} Å^{-2} and −3.298 ± 0.283 kcal/mol, respectively (Table 1).

The derived area coefficient, γ, is surprisingly close to the macroscopic surface tension of water, which is 0.105 kcal mol^{-1} Å^{-2} (converted from 72.75 dyn/cm at 20 °C).[56] Other simulations have also yielded microscopic surface tensions of water close to the macroscopic one. Postma et al. used MD simulations with explicit water to investigate the dependence of free energy of cavity formation on the cavity size.[57] The calculated free energies of formation of various sizes of spherical cavities were then related to the cavity radius using a quadratic polynomial. The coefficient of the squared term is then the calculated surface tension for cavities with radii not close to zero. They obtained a surface tension of 0.067 N/m or 0.096 kcal mol^{-1} Å^{-2}. Prevost et al.[58] again using explicit water molecules and free energy perturbation methods gave a coefficient of 0.095 kcal mol^{-1} Å^{-2}.

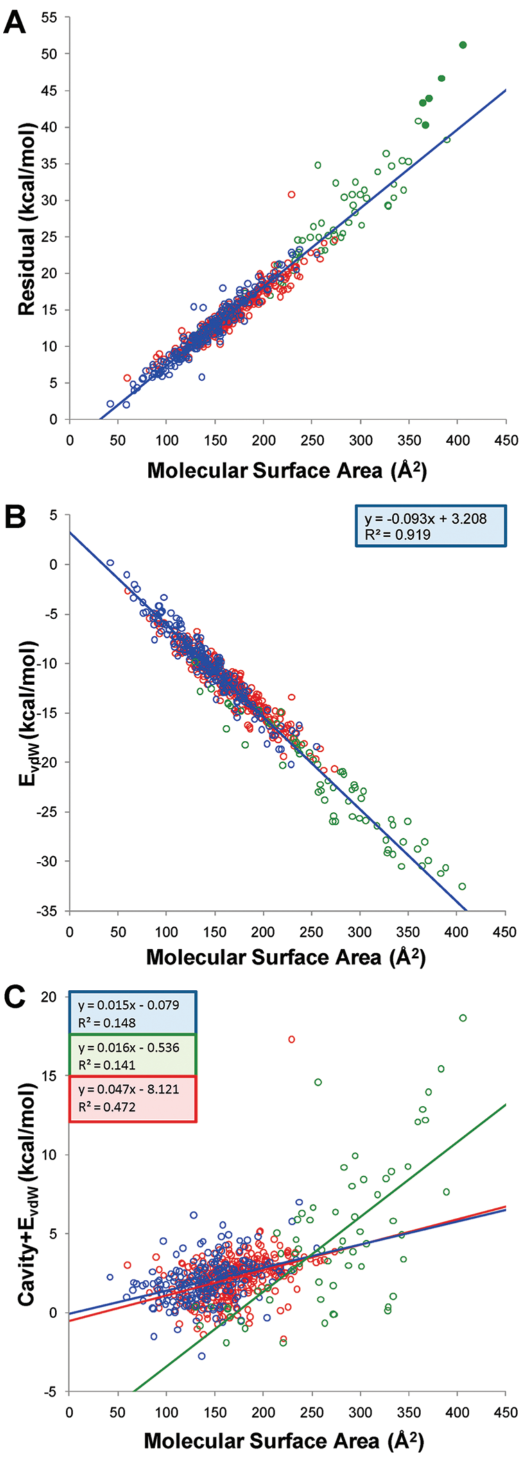


Figure 1. Deriving the cavity contribution. (A) Linear relationship between pseudoexperimental (residual) cavity contribution and the MSA for the training (blue symbols) and testing (red symbols) data sets and for the SAMPL1 (green symbols) data set. The plotted data correspond to MD simulations with flexible solute and AM1BCC-SP charges. Only the regression line for the training data set is shown, since this is used to predict the cavity contribution for the testing and SAMPL1 data sets. Filled circle points correspond to 5 sulfoneurea analogs from the SAMPL1 data set. (B) Correlation between the solute−solvent van der Waals interaction energy and the molecular surface area of the solute for the same LIE model. (C) Scatter plot of the total nonpolar term to solvation and the molecular surface area of the solute for the same LIE model. Regression lines are shown for each data set and are colored as the corresponding symbols.

Table 1. Parameters for the Cavity Cost That Can Be Derived from Linear Relationships between the Pseudo-Experimental (Residual) Cavity *versus* the Solute Molecular Surface Area, for the Indicated Hydration Data Sets[a]

set	slope (γ)	intercept (C)	R^2
training	0.108 ± 0.002	−3.298 ± 0.283	0.923 ± 0.015
testing	0.097 ± 0.002	−1.869 ± 0.372	0.909 ± 0.022
SAMPL1	0.126 ± 0.006	−7.118 ± 1.488	0.896 ± 0.026

[a] γ is in kcal mol^{-1} Å^{-2} and C is in kcal mol^{-1} units. Data are for AM1BCC-SP charges.

We used the two cavity parameters, γ and C, calibrated on the training data set, to predict the cavity contributions for testing data sets. The validity of this extrapolation can be judged from the relationship between pseudoexperimental cavity contribution and the MSA for the training, testing, and SAMPL1 data sets totalling 564 compounds (Figure 1A). The linear relationship calibrated on the training data set extends very well to the testing data set (γ = 0.097 kcal mol^{-1} Å^{-2}, C = −1.869 kcal/mol) and also to most compounds of the SAMPL1 data set (γ = 0.126 kcal mol^{-1} Å^{-2}, C = −7.118 kcal/mol), up to an MSA of about 350 Å^2. For the larger compounds from the SAMPL1 data set, the calibrated cavity parameters appear to underestimate their pseudoexperimental (residual) cavity cost, which is also reflected by the larger slope (γ) value obtained by fitting directly to SAMPL1 data set (Table 1). Although this behavior may suggest the existence of nonlinear components in the dependence of the cavity term to the MSA, we also note that these larger compounds are sulfoneurea analogs and thus belong to the same functional class (filled circles in Figure 1). It is possible that the apparent steeper surface area dependence of the cavity term for these compounds is only compensating for other factors, e.g., limitations of the molecular mechanics force-fields to accurately describe the interaction energy terms for this functional group, as previously suggested.[17] Indeed, most of the current state-of-the-art solvation methods when applied to the SAMPL1 data set failed miserably on these sulfoneurea analogs,[17,31−33] although some of the QM-based methods seem to provide a better agreement to experiment for some of the sulfoneurea analogs.[30,34]

Total Nonpolar Component. The solute−solvent van der Waals interaction energy also correlates well with the MSA (Figure 1B). However, the total nonpolar contribution to solvation, comprising the cavity cost and the solute−solvent van der Waals interaction energy, does not correlate with the solute surface area (Figure 1C), due to the strong anticorrelation between these terms. This lack of correlation is particularly pronounced for compounds from the training and testing data sets, which mirrors the results reported for these types of compounds based on FEP simulations in explicit water.[16,18] Our analysis of the drug-like SAMPL1 compounds, which have larger surface areas, shows only a moderate correlation (Figure 1C) between the total nonpolar solvation and the MSA for this data set (R^2 of 0.472). Clearly, a single linear surface-area-dependent term cannot describe the total nonpolar component of solvation.

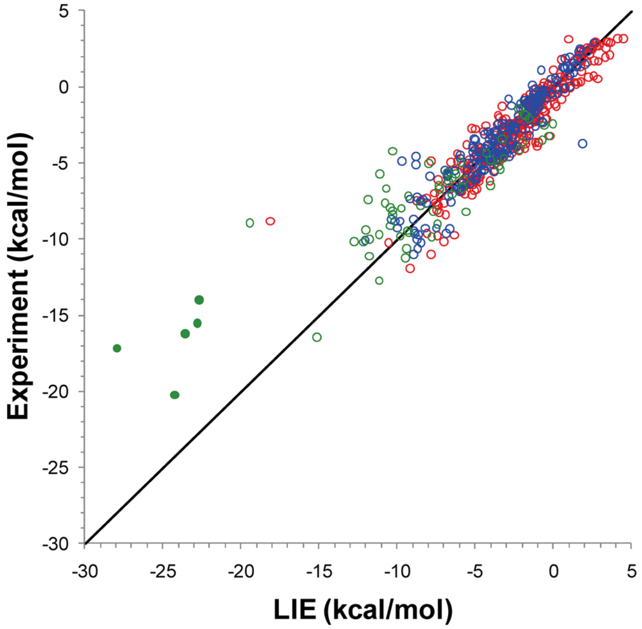


Figure 2. Correlation between LIE predictions of hydration free energy and experimental data for the training (blue symbols) and testing (red symbols) sets and for the SAMPL1 (green symbols) data set. Filled circle points correspond to 5 sulfoneurea analogs from the SAMPL1 data set. The plotted data correspond to MD simulations with flexible solute and AM1BCC-SP charges, and cavity parameters derived on the training data set. The diagonal line indicates ideal correlation and a unit slope.

Correlation with Experimental Hydration Data. The correlation between the experimental hydration free energies and the values calculated with the LIE approach (eq 1) using full flexibility of the solute and AM1BCC-SP partial charges is plotted in Figure 2. As described earlier, LIE application to hydration free energy prediction requires only two fitted parameters (for the cavity cost) derived from a training data set. The excellent performance of the LIE approximation on the testing set of 301 molecules is similar to that on the training set, i.e., mean-unsigned errors (MUE) below 0.9 kcal/mol, slopes close to unity, and high values of squared correlation coefficients (Table 2). Testing on the SAMPL1 data set of 63 drug-like compounds is less accurate but still very acceptable for such a challenging data set: MUE slightly above 2 kcal/mol and an R^2 of about 0.8. Somewhat worrisome is the correlation slope of about 0.6.

Excluding Continuum Correction Terms. We tested the effect of ignoring the electrostatic and van der Waals correction terms beyond the 12 Å shell of explicit water (eqs 2 and 3, respectively). Hence, the removal of these terms from eq 1 and recalibration of the two cavity parameters resulted in only a negligible change in the prediction performance (Table 2). This is due to small values for the calculated correction terms for neutral molecules beyond the 12 Å shell of explicit water and would suggest that these terms need not be calculated in this case. Nevertheless, thinner shells of explicit water would benefit from the corrections. Also, our calculations indicate that the electrostatic correction becomes significant in the case of charged compounds; for example, it represents about 15% of the solute−solvent electrostatic interaction within the 12 Å explicit water shell around a monatomic monovalent ion.[35] We will continue to present the rest of the data in the paper by including these corrections due to the completeness of the approach.

Comparison of Charge Models. Results presented up to this point were obtained with the single-point (SP) version of the AM1BCC charging method based on semiempirical AM1 determination of the charge followed by bond charge correction. In AM1BCC-SP, charges are obtained without AM1 geometry optimization of the molecule beyond its already force-field-minimized conformation. This version was inspired by the work of Nicholls et al.[32] who observed improved throughput and results without AM1 optimization, as measured both by consistency of transfer energy predictions using Poisson−Boltzmann and by comparison to experimental dipoles, presumably due to overpolarization after AM1 geometry optimization. We also tested here the more typical AM1BCC charges with AM1 geometry optimization (OPT), as well as RESP charges that are fitted to the electrostatic potential calculated from *ab initio* quantum mechanics. For each new charging method, the cavity parameters were recalibrated on the same training set. These cavity parameters varied only slightly between the various charging methods employed (Table S4, Supporting Information).

The change in the accuracy of LIE predictions with various charging methods is modest (Table 3). In terms of MUEs (for the training, testing, and SAMPL1 sets), the AM1BCC-SP charging performed best overall (0.830, 0.849, and 2.245 kcal/mol), followed by AM1BCC-OPT (0.792, 0.903, and 2.400 kcal/mol) and RESP (0.943, 0.911, and 2.333 kcal/mol). In correlative terms, the two AM1BCC versions produced comparable results, with R^2 close to 0.9 for the training and testing and around 0.8 for SAMPL1, but RESP charges gave lower R^2 values for all data sets, with an R^2 of just above 0.8 for the training and testing and around 0.7 for SAMPL1. The only improvement seen with the RESP charges was a small increase of the correlation slope for the SAMPL1 data set (0.65) relative to the other methods (below 0.6). These modest differences most likely arise from the

Table 2. Effect of Continuum Correction Terms beyond the 12-Å Explicit Water Shell for LIE Predictions of Experimental Hydration Free Energy[a]

	with correction to ∞			without correction to ∞		
set	MUE	slope	R^2	MUE	slope	R^2
training	0.830 ± 0.055	0.940 ± 0.037	0.864 ± 0.025	0.838 ± 0.053	0.948 ± 0.037	0.863 ± 0.024
test	0.849 ± 0.049	0.927 ± 0.051	0.867 ± 0.025	0.838 ± 0.048	0.937 ± 0.050	0.868 ± 0.025
SAMPL1	2.245 ± 0.342	0.583 ± 0.045	0.793 ± 0.056	2.228 ± 0.292	0.589 ± 0.046	0.795 ± 0.054

[a] Data are for AM1BCC-SP charges. Errors are in kcal mol^{-1} units.

Table 3. Effect of the Partial Charge Model for LIE Predictions of Experimental Hydration Free Energy[a]

	AM1BCC-SP		
set	MUE	slope	R^2
training	0.830 ± 0.055	0.940 ± 0.037	0.864 ± 0.025
testing	0.849 ± 0.049	0.927 ± 0.051	0.867 ± 0.025
SAMPL1	2.245 ± 0.342	0.583 ± 0.045	0.793 ± 0.056

	AM1BCC-OPT		
set	MUE	slope	R^2
training	0.792 ± 0.054	0.941 ± 0.035	0.870 ± 0.025
testing	0.903 ± 0.047	0.909 ± 0.046	0.858 ± 0.223
SAMPL1	2.400 ± 0.354	0.568 ± 0.046	0.786 ± 0.051

	RESP		
set	MUE	slope	R^2
training	0.943 ± 0.067	0.931 ± 0.044	0.810 ± 0.033
testing	0.914 ± 0.051	0.953 ± 0.051	0.828 ± 0.026
SAMPL1	2.333 ± 0.31	0.652 ± 0.046	0.670 ± 0.051

[a] Errors are in kcal mol^{-1} units.

training of the AM1BCC charges to reproduce RESP charges.[43,44] When using other less accurate charging methods (such as MMFF[42] or Gasteiger−Marselli[59]), larger variation can be expected. These results indicate that, while there is a relatively minor influence of the partial charge set on the accuracy of the LIE predictions using our charge selection, AM1BCC-SP charges are favored. These results agree with the results of Roux and co-workers,[18] and thus, given its throughput and accuracy, AM1BCC-SP appears as the charging method most applicable to screening of large databases of small molecules, at least in the case of neutral compounds, instead of the more expensive RESP method. Therefore, other dependencies are examined with all charge sets, but only AM1BCC-SP will be discussed unless another charging model yields a significant improvement.

Flexible versus Single-Conformation Solute. We were interested to examine whether the explicit solvation model deteriorates considerably if it is based on a single conformation of the solute. This is important when developing continuum solvation models based on explicit models. Thus, we have carried out MD simulations in TIP3P water at 300 K, but with the solute constrained to its starting conformation, and then applied LIE calculations (after recalibrating the cavity parameters on the training set). The results listed in Table 4 show that the LIE predictions with respect to MUE (training, testing) with the rigid solute (0.847, 0.906 kcal/mol) are not significantly worse than those with full solute flexibility (0.830, 0.849 kcal/mol) for the training and testing data sets, and somewhat to our surprise, these predictions actually even improve in the case of the SAMPL1 data set where the rigid model gave an MUE of 1.924 kcal/mol compared to an MUE of 2.245 kcal/mol. At the moment, we do not have an explanation for this latter behavior, which may be fortuitous, but nevertheless is encouraging in terms of our ultimate goal of developing a solvation model that is physics-based as in an explicit model but fast as in a continuum model. The change in the accuracy of LIE predictions between the rigid and flexible solute models does not appear to be related to the number of solute rotatable bonds (as examined for the testing set, see Figure S1, Supporting Information).

In a study by Mobley et al.,[29] it was shown that the single-conformer solvation free energy computed with an implicit model can vary significantly depending on the conformation used for the calculation. Interestingly, however, using the single-conformation approach based on the lowest potential energy in a vacuum (the “BestVac” scheme) gave predictions similar to the solvation free energies calculated from the flexible-solute implicit solvation model (RMS deviation of 0.34 kcal/mol between the models, with less than 0.1 kcal/mol difference between the RMS errors of these models relative to experiment). In the present study, which in effect uses the same “traditional” hydration data set and the BestVac approach for the single-conformation model, we obtain a similar deviation between the explicit-solvent LIE predictions based on rigid and flexible solutes (RMS deviation of 0.52 kcal/mol between the models, with less than 0.1 kcal/mol difference between the RMS errors of these models relative to the experiment). Nonetheless, although there is good agreement between the LIE data for the flexible-solute and rigid-solute models for most compounds in this data set, differences of 1−2 kcal/mol are obtained in some cases (Figure 3).

Inclusion of Internal Energy Terms. We explored further the case of flexible solute and carried out additional MD simulations in the gas phase in order to take into account the difference in the internal energy of the solute between the solution phase and the gas phase (eq 4). As with all tests presented earlier, the cavity parameters had to be recalibrated on the training subset for each model being examined. We calculated all molecular mechanics intramolecular energy terms, i.e., bond stretching, angle bending, torsional (including improper corrections), 1−4 and 1−5 van der Waals, and 1−4 and 1−5 Coulombic electrostatic energies. Inclusion of all intramolecular terms resulted in marginal changes of the LIE predictions of hydration free energies in terms of MUE (training, testing, SAMPL1) for the data sets investigated (1.090, 0.926, 2.174 kcal/mol; Table 5). We noticed large fluctuations of the bond stretching, angle bending, and torsional energies along the MD trajectories in both phases (data not shown), which prompted us to exclude these terms

Table 4. Effect of Flexible versus Single-Conformation Solute for LIE Predictions of Experimental Hydration Free Energy[a]

	flexible solute			rigid solute		
set	MUE	slope	R^2	MUE	slope	R^2
training	0.830 ± 0.055	0.940 ± 0.037	0.864 ± 0.025	0.847 ± 0.054	0.962 ± 0.038	0.858 ± 0.025
testing	0.849 ± 0.049	0.927 ± 0.051	0.867 ± 0.025	0.906 ± 0.043	0.974 ± 0.025	0.864 ± 0.013
SAMPL1	2.245 ± 0.342	0.583 ± 0.045	0.793 ± 0.056	1.924 ± 0.246	0.635 ± 0.050	0.808 ± 0.057

[a] Data are for AM1BCC-SP charges. Errors are in kcal mol^{-1} units.

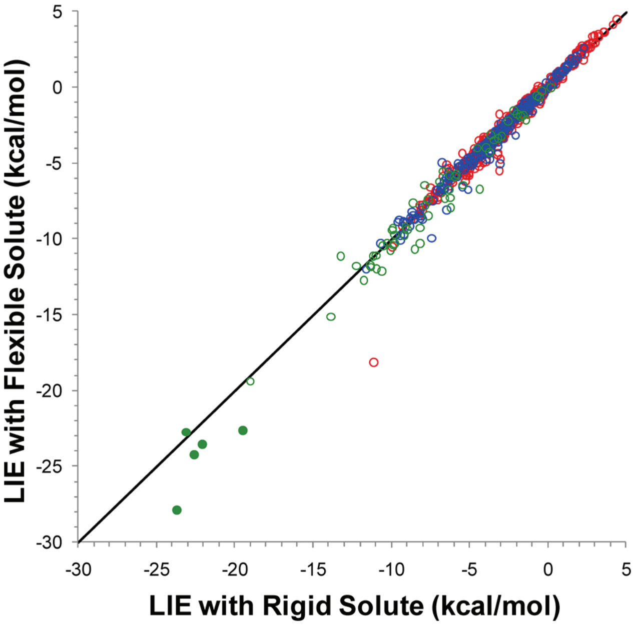


Figure 3. Correlation between LIE predictions of hydration free energy based on MD simulations with flexible solute versus single-conformation (rigid) solute, for the training (blue symbols) and testing (red symbols) sets and for the SAMPL1 (green symbols) data set. Filled circle points correspond to 5 sulfoneurea analogs from the SAMPL1 data set. The plotted data correspond to MD simulations with AM1BCC-SP charges and cavity parameters derived from the training data set. The diagonal line indicates ideal correlation and unit slope.

Table 5. Effect of Including Internal Energy Terms for LIE Predictions of Experimental Hydration Free Energy[a]

	no intramolecular terms		
set	MUE	slope	R^2
training	0.830 ± 0.055	0.940 ± 0.037	0.864 ± 0.025
testing	0.849 ± 0.049	0.927 ± 0.051	0.867 ± 0.025
SAMPL1	2.245 ± 0.342	0.583 ± 0.045	0.793 ± 0.056
	with 1−4 and 1−5 intramolecular terms		
set	MUE	slope	R^2
training	0.918 ± 0.058	0.967 ± 0.039	0.838 ± 0.026
testing	0.925 ± 0.046	0.986 ± 0.025	0.925 ± 0.016
SAMPL1	1.836 ± 0.193	0.711 ± 0.066	0.750 ± 0.068
	with all intramolecular terms		
set	MUE	slope	R^2
training	1.090 ± 0.296	0.828 ± 0.086	0.715 ± 0.066
testing	0.926 ± 0.047	0.987 ± 0.025	0.854 ± 0.016
SAMPL1	2.147 ± 0.434	0.624 ± 0.076	0.712 ± 0.074

[a] Data are for AM1BCC-SP charges and flexible solutes. Errors are in kcal mol^{-1} units.

and retain only the 1−4 and 1−5 intramolecular terms that have smaller fluctuations. As seen in Table 5, inclusion of 1−4 and 1−5 intramolecular terms deteriorates only slightly the predictions for the training and testing sets in terms of MUE (by 0.1 kcal/mol compared to no intramolecular terms) while bringing the correlation slopes closer to unity. Inclusion of these terms improves predictions for the challenging SAMPL1 data set both in terms of MUE (by over 0.4 kcal/mol) and slope. Thus, inclusion of the difference in nonbonded intramolecular van der Waals and electrostatic energies alongside the LIE terms appears to be a promising approach that deserves further attention.

Table 6. Listing of Function Groups Used for Error Analysis

functional group	# of members
other	81
alkane	20
alkene	13
alkyne	3
aromatic	18
halogen	57
F	3
Cl	31
Br	12
I	4
OH	27
1° OH	10
2° OH	4
3° OH	2
phenyl OH	11
amine	10
alkyl amine	7
aryl amine	3
carboxylic acid	2
ester	30
amide	2
ether	8
ketone	12
aldehyde	8
nitro	1
cyano	3
hypervalent S	1
thiol	5

Functional Group Analysis. In order to identify problematic functional groups for the LIE solvation models, we have separately examined several classes of compounds which contain only one functional group (may contain multiple of the same function group) from the testing set of 301 compounds (Table 6). The error analysis was carried out for flexible solute molecules (without inclusion of internal energies) with various charge sets (Figure 4). There are a multitude of ways these data sets can be split into functional classes. We were particularly interested to compare the performance on saturated alkanes, alkenes, alkynes, aromatic hydrocarbons, halogenated compounds, alcohols, phenols, aryl-amines, amides, esters, ketones, aldehydes, nitro, cyano, hypervalent S, and thiol derivates. We also examined aliphatic amines and carboxylic acids, although the hydration free energy of these groups in neutral form as included in these data sets is less relevant for studies in biomolecular systems at the physiological pH.

We see in Figure 4 (see Table S5, Supporting Information, for the raw data used in Figure 4) in terms of MUEs (range for various charge sets) that the alkanes (0.40−0.58 kcal/mol), alkene (0.36−0.71 kcal/mol), and aromatic hydrocarbons (0.29−0.42 kcal/mol) are predicted well, with about half the MUE of the entire testing set (0.85−0.91 kcal/mol), but the RESP charges overestimate the hydration of alkanes, giving a mean-signed error (MSE) of −0.55 kcal/mol. Alkynes are not predicted well in terms of MUE with AM1BCC-SP (1.55 kcal/mol) and AM1BCC-OPT (1.67

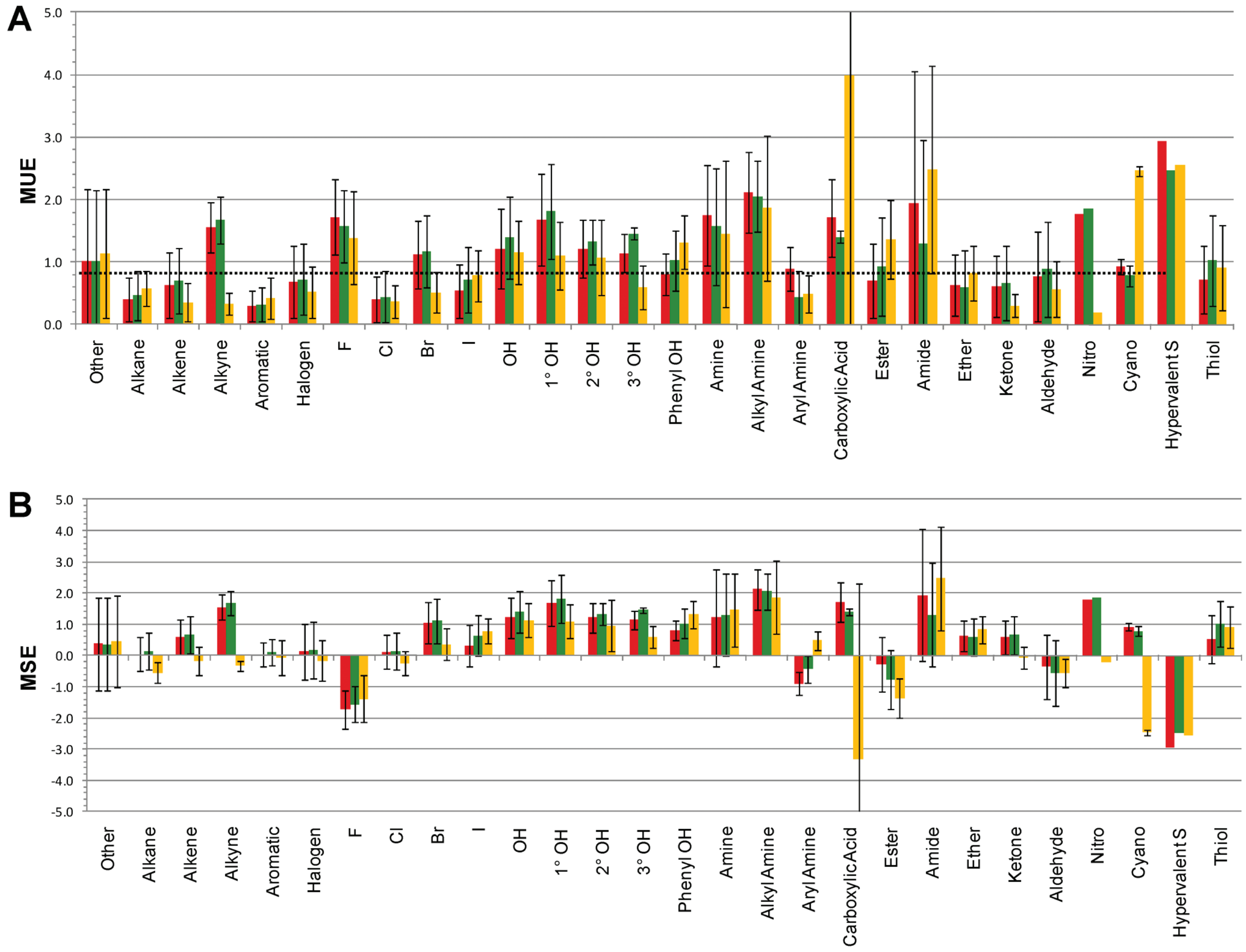


Figure 4. Functional group analysis of the testing set in terms of LIE prediction errors for various partial charge sets: AM1BCC-SP (red bars), AM1BCC-OPT (green bars), and RESP (yellow bars). (A) MUE ± SD values. The data correspond to MD simulations with flexible solute and cavity parameters derived on the training data set for each charge model. The dotted line corresponds to the MUE value of 0.85 kcal/mol for the entire data set with AM1BCC-SP partial charges (see also Table 3). (B) MSE ± SD values.

kcal/mol) and are underestimated, giving an MSE of 1.55 and 1.67 kcal/mol, respectively, yet changing to RESP charges decreases the MUE (0.33 kcal/mol) to well below the average for the set. Other studies have attributed this inaccuracy to the GAFF force field parameters for alkynes,[16] yet this shows that charging may also play a role. The subset of halogenated compounds as a whole also provides good predictions with all charge sets (0.52−0.72 kcal/mol), with an MUE below that of the entire data set and an MSE close to zero (between −0.16 and 0.18 kcal/mol). Further decomposition into various halogens highlights problems with the fluorinated and brominated compounds, but not with the chlorinated and iodinated ones. The hydration of fluorinated compounds is overestimated (MSE between −1.57 and −1.72 kcal/mol), and that of brominated compounds is underestimated (MSE between 1.05 and 1.10 kcal/mol), leading to MUEs in the 1−1.5 kcal/mol range with the AM1BCC charges, but these errors can be partially corrected by employing RESP partial charges. Fluorinated compounds are still overestimated but less so (MSE of −0.68 kcal/mol) with an MUE around that of the testing set (0.88 kcal/mol), with brominated compounds seeing a greater improvement with an MUE of 0.51 kcal/mol and an MSE of 0.35 kcal/mol. Using AM1BCC charges, alcohols of all types gave larger MUE values (above 1.2 kcal/mol) with all alcohol types having underestimated hydration free energies with the MSE ranging between 1.21 and 1.40 kcal/mol. Alcohols are particularly problematic with the AM1BCC partial charge sets, which is demonstrated when broken into primary (MUE range of 1.68−1.81 kcal/mol), secondary (1.21−1.32 kcal/mol), and tertiary alcohols (1.14−1.46 kcal/mol) and phenols (0.80−1.03 kcal/mol), whereas RESP partial charges improve LIE predictions for all alcohol classes (MUE for 1° OH, 1.10 kcal/mol; 2° OH, 1.07 kcal/mol; 3° OH, 0.60 kcal/mol) but worsen the LIE prediction for phenols (1.32 kcal/mol). The amides were predicted worse than the average for the entire data set (MUE of 1.3−2.5 kcal/mol depending on the charge model), whereas esters and aryl-amines better than the average (MUE range of 0.36−0.89 kcal/mol) with AM1-BCC charge sets, and with RESP faring well with aryl-amines (0.48 kcal/mol) yet worse for esters (1.36 kcal/mol). The aliphatic amines and carboxylic acids were also problematic and with all charge models (MUE within 1.41−3.98 kcal/mol range depending on charge set); however, these

Table 7. Comparison of LIE and FEP Predictions of Experimental Hydration Free Energy[a]

	LIE			FEP		
set	MUE	slope	R^2	MUE	slope	R^2
training	0.792 ± 0.054	0.941 ± 0.035	0.870 ± 0.025	1.095 ± 0.055	0.866 ± 0.027	0.873 ± 0.020
testing	0.903 ± 0.047	0.909 ± 0.046	0.858 ± 0.223	0.997 ± 0.040	0.936 ± 0.026	0.897 ± 0.012
SAMPL1[b]	2.260 ± 0.325	0.591 ± 0.047	0.821 ± 0.045	2.594 ± 0.380	0.567 ± 0.039	0.826 ± 0.051

[a] LIE data are for AM1BCC-OPT charges and flexible solutes. FEP data are taken from Mobley et al.[16,17] Errors are in kcal mol^{-1} units. [b] LIE and FEP data are for 56 out of the 63 compounds in the SAMPL1 data set, as in Mobley et al.[17]

functional groups are less important in the neutral state studied here. Other functional classes of compounds, ethers, ketones, aldehydes, thiols, and cyano derivatives, were predicted close to the set average or better (MUE in the range 0.57−1.02 kcal/mol), irrespective of the charging method. One exception was the cyano derivatives that had severely overestimated hydration free energies by using RESP charges (MSE of −2.46 kcal/mol, MUE of 2.46 kcal/mol). There is only one hypervalent S-containing compound and one nitro compound in the testing set, so we could not assess the prediction errors for these two chemical classes.

In summary, we see that some functional groups like primary alcohols, neutral alkyl amines, and amides remain relatively problematic, on average having hydration free energy prediction errors underestimated by about 2 kcal/mol. For a few functional classes of compounds average prediction errors are between 1 and 2 kcal/mol, either underestimated or overestimated. The choice of partial charge set does not appear to be a consensus source for such deviations.

Comparison between LIE and FEP Predictions. The results of the LIE study presented here can be directly compared with those from alchemical FEP calculations of hydration free energies carried out by Mobley et al. on the same data sets and with the same force-field and charging method.[16,17] One notable difference between the FEP and LIE approaches is that much shorter MD simulations in explicit solvent are required with the latter, which for the current comparison translate into at least a 1 order of magnitude speedup. This is an advantage if the increased efficiency does not compromise prediction accuracy. In Table 7, we compare the FEP predictions with LIE predictions carried out with AM1BCC-OPT partial charges and flexible solutes. We see that the LIE approach yields comparable, slightly improved predictions relative to FEP predictions in terms of MUE on the training (0.792 vs 1.095 kcal/mol), testing (0.903 vs 0.997 kcal/mol), and SAMPL1 (2.260 vs 2.594 kcal/mol) data set (SAMPL1 data set is only for 56 compounds analyzed by Mobley et al.[17]). A direct scatter plot between the LIE and FEP predictions is shown in Figure 5. We see that, in the case of the combined training and testing data set compounds, the LIE predictions are more negative that the FEP predictions. This change is in the correct direction since the FEP predictions for this data set were shown to be too positive relative to the experimental data.[16] Indeed, for the combined training and testing data set, we obtain an MSE of 0.21 kcal/mol with the LIE method, compared to 0.68 kcal/mol afforded by the FEP data (MSEs of 0.35 kcal/mol versus 0.69 kcal/mol with the LIE versus FEP methods on the testing subset).[16] In the case of the SAMPL1 data set, the LIE predictions are less negative than the FEP predictions, whereas the latter were found to overestimate the experimental data (MSE of −1.88 kcal/mol). Thus, the LIE approach appears to at least be as accurate as the FEP approach for the prediction of hydration free energies, at a fraction of computing time. While this observation may seem counterintuitive, one possible explanation is that the terminal step in the alchemical transformation when the solute "disappears" ($\lambda = 1.0$ in turning off the Lennard-Jones solute-water interactions) may introduce some noise in the FEP calculations. Also, our LIE model has two fitted parameters for the cavity term, while FEP has no fitted parameters at all.

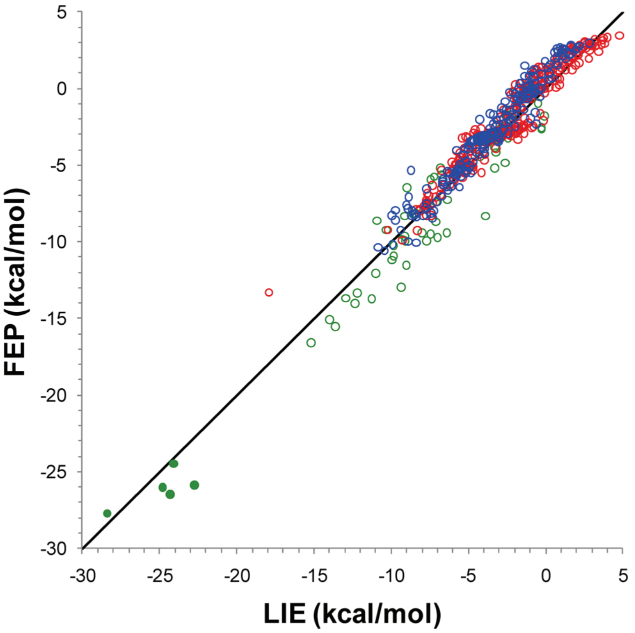


Figure 5. Comparison between LIE predictions (this study) and FEP predictions (Mobley et al.[16,17]) for the training (blue symbols) and testing (red symbols) sets and for the SAMPL1 (green symbols) data set. Filled circles correspond to 5 sulfoneurea analogs from the SAMPL1 data set. The diagonal line indicates ideal correlation. The data correspond to MD simulations with flexible solute and AM1BCC-OPT charges, and cavity parameters derived from the training data set.

We further extended this comparison to examine whether these LIE and FEP models also perform similarly on the same functional groups. The analysis shown in Figure 6 (for raw data see Table S5), although limited to a relatively few functional groups represented by compounds containing only one type of functional group from the testing set, highlights that the LIE methods performed better on alkenes; chlorinated, brominated, and iodinated compounds; ethers; and nitriles, whereas the FEP method was superior on fluorinated

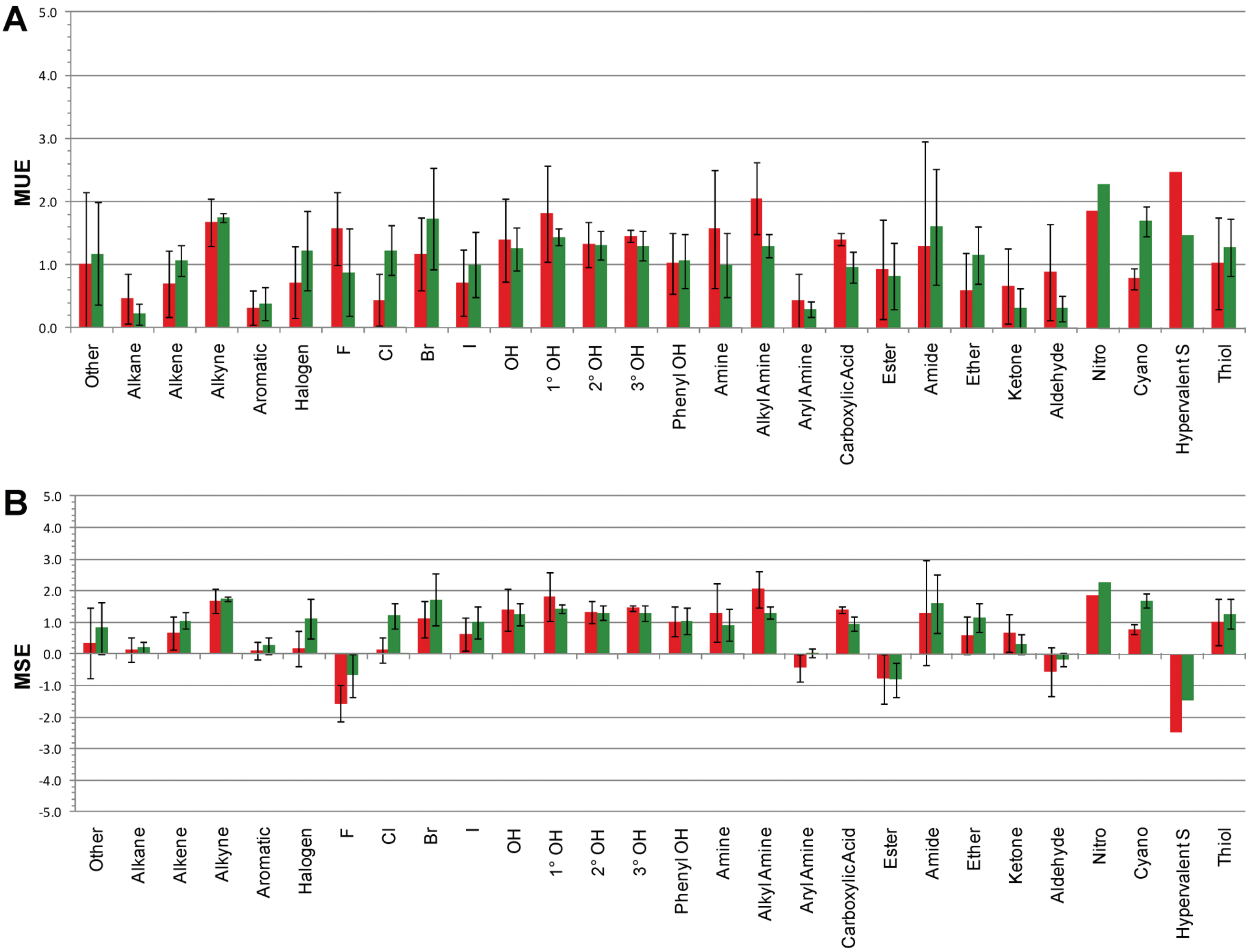


Figure 6. Comparison between LIE predictions (red bars, this study) and FEP predictions (green bars, Mobley et al.[16,17]) for functional classes represented by monofunctional compounds from the testing subset. The LIE data correspond to MD simulations with flexible solute and AM1BCC-OPT charges, and cavity parameters derived from the training data set. (A) MUE ± SD values. (B) MSE ± SD values.

compounds, primary alcohols, neutral aliphatic amines and carboxylic acids, aldehydes, and ketones.

Decomposition into Electrostatic and Nonpolar Contributions. Free energies are often decomposed into contributions from various components. A common decomposition is into electrostatic and nonpolar contributions. In the FEP approach, this is obtained through a charge decoupling calculation. The free energy change resulting from alchemically turning off all solute partial charges is assigned to the electrostatic contribution to hydration free energy. The nonpolar component is then obtained as the free energy change from subsequently turning off the solute−solvent Lennard-Jones interactions. However, it can be argued that this decoupling scheme does not yield a purely electrostatic contribution to the free energy because turning off the solute partial charges also results in a change in the solute−solvent van der Waals interaction energy. On the other hand, in the LIE calculation, the decomposition is somewhat cleaner. The electrostatic contribution is directly calculated from the solute−solvent Coulomb interaction energy in the presence of but not including the van der Waals interaction energy. By that we mean the configurations in the trajectory are determined by both the electrostatics and van der Waals interactions, but the electrostatic and van der Waals contributions can be formally completely separately accounted for. This distinction is important because part of the motivation for our carrying out LIE simulations on the training and testing data sets is to use the results to calibrate a new continuum solvation model. In continuum electrostatics theory, the electrostatic hydration free energy obtained from a solution of the Poisson equation (i.e., the reaction field energy) is more closely related to the LIE electrostatic decomposition than the FEP one. Hence, the LIE electrostatic decomposition is more appropriate for comparison with continuum electrostatics. Similarly, the LIE solute−solvent van der Waals energy can be used directly to calibrate a continuum van der Waals function.

A direct comparison of the electrostatic and nonpolar contributions to hydration calculated with the FEP and LIE approaches on the compounds in the training and testing data sets is given in Figure 7. While there is a good correlation between the electrostatic components from the LIE and FEP approaches, the FEP electrostatic component is systematically more positive than the LIE electrostatic component (Figure 7A). This is possibly due to the net positive change in the solute−solvent van der Waals interaction energy upon solute

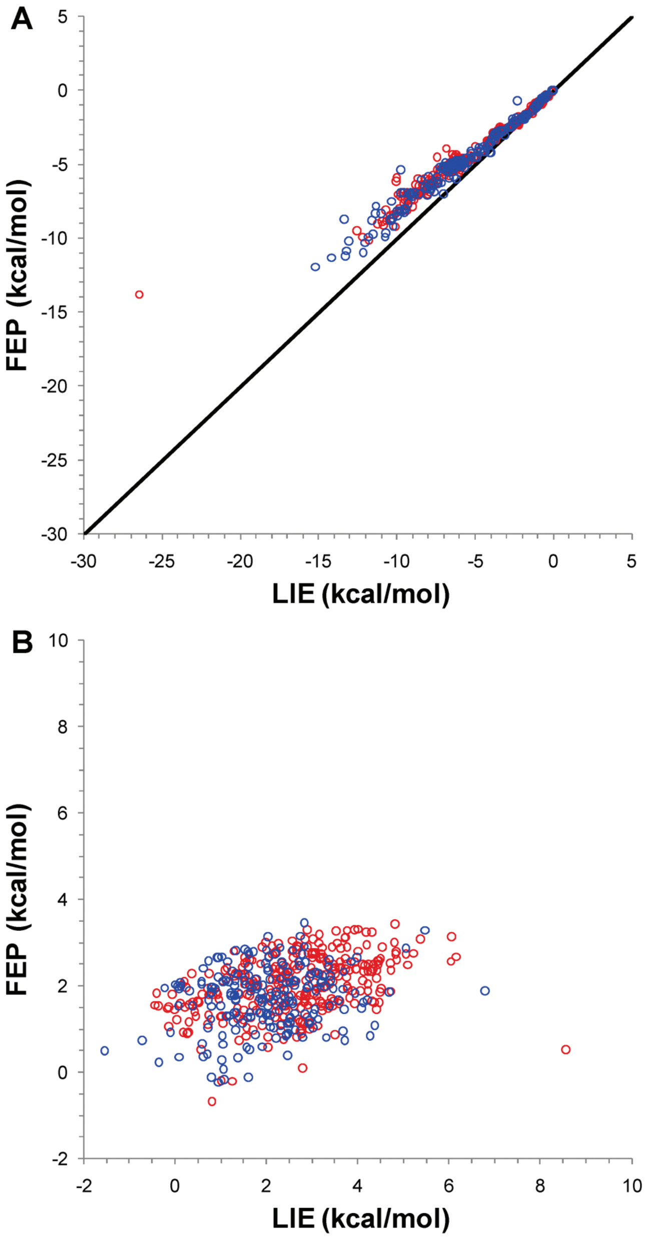


Figure 7. Correlation between the LIE and FEP contributions to hydration free energy. FEP data are taken from Mobley et al.[16] Data points for the training and testing sets are shown with blue and red symbols, respectively. (A) Electrostatic contributions. Line indicates ideal correlation. (B) Total nonpolar contributions. In the case of LIE data, the total nonpolar contribution includes solute−solvent van der Waals interactions and the cavitation cost.

charging, which is embedded into the FEP "electrostatic" component, although deviations from the linear response approximation may also be claimed. We have performed additional end-point calculations to quantify these effects. For 20 molecules with the largest deviations between the FEP and LIE electrostatic components (Figure 7A), we carried out MD simulations with zeroed solute charges, then recharged the solute and computed the solute electrostatic and van der Waals interactions with the nonpolarized solvent. The additional data are summarized in Table S6 (Supporting Information). Two observations stem out of these calculations. First, there is very little residual electrostatic interaction with the nonpolarized solvent, which strengthens the LIE approximation and supports the ideal theoretical value of 0.5 for the scaling of the LIE electrostatic interaction energy. Second, the difference in van der Waals interaction energy of the solute with the polarized and nonpolarized solvent compares well with the difference between the LIE and FEP electrostatic components. This supports the view that an important part of the deviation seen in Figure 7A is due to a "contamination" of the FEP electrostatic component with a positive van der Waals term incurred upon turning off the van der Waals potentials, as previously suggested by others.[60,61] There is little correlation between the total nonpolar components calculated with the LIE and FEP methods (Figure 7B), due to the aforementioned formal redistribution of contributions and the narrower range of values relative to the electrostatic component.

Conclusions

The present study provides a comprehensive systematic analysis on the applicability of the LIE approach to the prediction of gas-to-water transfer free energy of small-molecule organic solutes. While the approach presented here is not new, to the best of our knowledge, this is the first theoretical analysis of hydration free energy that unifies such an extensive and diverse hydration data set comprising 564 neutral compounds with measured hydration free energies, including both "traditional" simpler compounds (the training and testing data sets) as well as more complex, drug-like compounds (the SAMPL1 data set).

Application of the LIE approach to solvation requires no empirical scaling of the solute−solvent interaction energy terms. However, a term describing the cost of cavity formation in water needs to be added to the force-field-based interaction energy terms. Using a diverse training subset that includes both polar and nonpolar solutes, we calibrated a robust linear relationship to the molecular surface area of the solute to describe the cavitation cost. On the basis of this relationship, we find that the microscopic surface tension of water is surprisingly close to the macroscopic one of 0.105 kcal mol^{-1} Å^{-2}. The calibrated parameters of the cavity term extend well to the compounds in the testing data sets. In agreement with other studies of solvation based on MD simulation in explicit solvent, the total nonpolar contribution to solvation calculated with the LIE method does not correlate with the solute surface area, due to a strong anticorrelation between its two main contributing factors, the cavity cost and the solute−solvent van der Waals interaction energy.

Excellent LIE models could be obtained with AM1BCC partial charges on flexible solutes in explicit water shells of 12 Å thickness and continuum models extending to infinity. These LIE models were highly correlative for the training, testing, and SAMPL1 data sets and are particularly accurate for the "traditional" compounds (MUE below 0.9 kcal/mol) and of acceptable accuracy in the case of the challenging drug-like compounds (MUE slightly above 2 kcal/mol). In the latter case, a group of sulfoneurea derivatives remain the major outliers largely responsible for the deterioration in performance of the LIE model, as recently reported with most solvation methods.

We have systematically analyzed the dependence of the LIE predictions to several parameters and models, namely, continuum corrections to infinity, partial charge set, solute flexibility, and internal energy terms. Excluding the continuum correction terms applied outside the explicit water shell has no impact on the LIE performance for these data sets of neutral compounds but will likely be important for solvation calculations on charged molecules. The change in the accuracy of LIE predictions with various partial charge sets is modest, with the AM1BCC-SP (without AM1 geometry optimization) charges favored over AM1BCC-OPT and RESP charges. Given its throughput and accuracy, AM1BCC-SP appears as a charging method well suited for important molecular discovery applications, such as virtual screening. The LIE predictions obtained on single-conformation solutes are not much worse than those with full solute flexibility for the training and testing data sets, and somewhat to our surprise, these predictions actually improve in the case of the SAMPL1 data set. This result is extremely important for developing continuum solvation models based on explicit models. In examining the effect of including the difference in the internal energy of the solute between the solution and gas phases, we found that the inclusion of all intramolecular energy terms in the LIE model appears to be a promising approach that can lead to improved prediction accuracies. The exclusion of covalent terms yielded slightly better results over using all terms, probably due to the larger fluctuations observed for the bonded terms during the MD simulations.

In an analysis of errors for a selection of functional groups represented by compounds only containing one type of functional group, we did not find any particular functional class to be a systematic major outlier. Various charge models impacted differently on the accuracy of prediction for different functional groups. For example, primary alcohols and neutral aliphatic amines had consistently underestimated hydration free energies by the LIE models with the AM1BCC partial charges, with the RESP charges having a larger improving effect for alcohols than for amines. In contrast, esters for example were only slightly overestimated with the AM1BCC charges, but RESP charges led to larger errors.

A direct comparison of the LIE and alchemical FEP approaches was possible given that they were applied on the same data sets using MD simulations in explicit water with the same force field and same charging method. One notable difference between the FEP and LIE approaches is that much shorter MD simulations in explicit solvent are required with LIE over FEP, which for the current comparison translate into at least a 1 order of magnitude speedup. This speedup is a real advantage since the increased efficiency of LIE relative to FEP does not compromise prediction accuracy, and we noticed even slightly improved LIE predictions relative to FEP on both the more “traditional” training and testing data sets and the challenging drug-like SAMPL1 data set. Thus, the LIE approach appears at least as accurate as the FEP approach for predicting hydration free energies, and this at a fraction of the computing time. Finally, different free energy decomposition paths are adopted in the FEP approach and the LIE approximation. It appears that LIE provides a simplified method for formally decomposing the solvation components that are calculated in the presence of each other. The LIE decomposition is also more compatible with the contributions to solvation that are typically calculated with continuum models including Poisson electrostatics, continuum van der Waals integrals, and surface-area-based cavitation. Together with its accuracy and speed, these attributes make LIE a suitable method for calibrating a continuum solvent model that will capture the physics of the explicit-solvent model and have the required speed for accurate high-throughput applications, as we have attempted in the companion paper.[35]

Acknowledgment. This is National Research Council of Canada publication number 50690.

Supporting Information Available: Composition of the hydration data sets with experimental transfer free energies (Table S1). Composition of groups used in the functional group analysis (Table S2). LIE components to hydration free energy for all models (Table S3). Cavity parameters calibrated on the training subset for various LIE models (Table S4). Raw data from Figure 4 and 6 (Table S5). Comparison of FEP data with LIE data for polarized and nonpolarized solvent for 20 molecules from the training and testing sets (Table S6). Performance of the LIE models with flexible and single-conformation solute as a function of the number of rotatable bonds (Figure S1). This material is available free of charge via Internet at http://pubs.acs.org.

References

(1) Rashin, A. A. *Prog. Biophys. Mol. Biol.* **1993**, *60*, 73–200.

(2) Honig, B.; Sharp, K.; Yang, A. S. *J. Phys. Chem.* **1993**, *97*, 1101–1109.

(3) Gilson, M. K.; Zhou, H. X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

(4) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199–203.

(5) Kang, Y. K.; Némethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1987**, *91*, 4109–4117.

(6) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978–1988.

(7) Chambers, C. C.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 16385–16398.

(8) Marten, B.; Kim, K.; Cortis, C.; Friesner, R. A.; Murphy, R. B.; Ringnalda, M. N.; Sitkoff, D.; Honig, B. *J. Phys. Chem.* **1996**, *100*, 11775–11788.

(9) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517–529.

(10) Tan, C.; Yang, L.; Luo, R. *J. Phys. Chem. B* **2006**, *110*, 18680–18687.

(11) Guthrie, J. P. *J. Phys. Chem. B* **2009**, *113*, 4501–4507.

(12) Mobley, D. L.; Barber, A. E.; Fennell, C. J.; Dill, K. A. *J. Phys. Chem. B* **2008**, *112*, 2405–2414.

(13) Chorny, I.; Dill, K. A.; Jacobson, M. P. *J. Phys. Chem. B* **2005**, *109*, 24056–24060.

(14) Purisima, E. O.; Sulea, T. *J. Phys. Chem. B* **2009**, *113*, 8206–8209.

(15) Reddy, M. R.; Erion, M. D. *Free Energy Calculations in Rational Drug Design*; Springer-Verlag: New York, 2001.

(16) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.

(17) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Dill, K. A. *J. Phys. Chem. B* **2009**, *113*, 4533–4537.

(18) Shivakumar, D.; Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 919–930.

(19) Lee, F. S.; Chu, Z. T.; Bolger, M. B.; Warshel, A. *Protein Eng.* **1992**, *5*, 215–228.

(20) Aqvist, J.; Medina, C.; Samuelsson, J. E. *Protein Eng.* **1994**, *7*, 385–391.

(21) Aqvist, J.; Luzhkov, V. B.; Brandsdal, B. O. *Acc. Chem. Res.* **2002**, *35*, 358–365.

(22) Aqvist, J.; Marelius, J. *Comb. Chem. High Throughput Screen.* **2001**, *4*, 613–626.

(23) Jones-Hertzog, D. K.; Jorgensen, W. L. *J. Med. Chem.* **1997**, *40*, 1539–1549.

(24) Su, Y.; Gallicchio, E.; Das, K.; Arnold, E.; Levy, R. M. *J. Chem. Theory Comput.* **2006**, *3*, 256–277.

(25) Wang, W.; Wang, J.; Kollman, P. A. *Proteins* **1999**, *34*, 395–402.

(26) Ben-Amotz, D.; Underwood, R. *Acc. Chem. Res.* **2008**, *41*, 957–967.

(27) Carlson, H. A.; Jorgensen, W. L. *J. Phys. Chem.* **1995**, *99*, 10667–10673.

(28) Almlof, M.; Carlsson, J.; Aqvist, J. *J. Chem. Theory Comput.* **2007**, *3*, 2162–2175.

(29) Mobley, D. L.; Dill, K. A.; Chodera, J. D. *J. Phys. Chem. B* **2008**, *112*, 938–946.

(30) Klamt, A.; Eckert, F.; Diedenhofen, M. *J. Phys. Chem. B* **2009**, *113*, 4508–4510.

(31) Sulea, T.; Wanapun, D.; Dennis, S.; Purisima, E. O. *J. Phys. Chem. B* **2009**, *113*, 4511–4520.

(32) Nicholls, A.; Wlodek, S.; Grant, J. A. *J. Phys. Chem. B* **2009**, *113*, 4521–4532.

(33) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 4538–4543.

(34) Soteras, I.; Forti, F.; Orozco, M.; Luque, F. J. *J. Phys. Chem. B* **2009**, *113*, 9330–9334.

(35) Corbeil, C. R.; Sulea, T.; Purisima, E. O. *J. Chem. Theory Comput.* **2010**; DOI: 10.1021/ct9006037.

(36) Purisima, E. O.; Nilar, S. H. *J. Comput. Chem.* **1995**, *16*, 681–689.

(37) Purisima, E. O. *J. Comput. Chem.* **1998**, *19*, 1494–1504.

(38) Bhat, S.; Purisima, E. O. *Proteins* **2006**, *62*, 244–261.

(39) Floris, F.; Tomasi, J. *J. Comput. Chem.* **1989**, *10*, 616–627.

(40) Floris, F. M.; Tomasi, J.; Pascual-Ahuir, J. L. *J. Comput. Chem.* **1991**, *12*, 784–791.

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

(42) Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 730–748.

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

(44) Jakalian, A.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

(45) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, 10269–10280.

(46) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *J. Am. Chem. Soc.* **1993**, 9620–9631.

(47) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

(48) 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.

(49) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, *25*, 247–260.

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

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

(52) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(53) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.

(54) *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2005.

(55) Haider, N. Checkmol. http://merian.pch.univie.ac.at/~nhaider/cheminf/cmmm.html (accessed March 4, 2010).

(56) Castellan, G. W. *Physical Chemistry*; 2nd ed.; Addison-Wesley: Reading, MA, 1971.

(57) Postma, J. P. M.; Berendsen, H. J. C.; Haak, J. R. *Faraday Symp. Chem. Soc.* **1982**, *17*, 55–67.

(58) Prevost, M.; Oliveira, I. T.; Kocher, J. P.; Wodak, S. J. *J. Phys. Chem.* **1996**, *100*, 2738–2743.

(59) Gasteiger, J.; Marsili, M. *Tetrahedron Lett.* **1978**, *19*, 3181–3184.

(60) Orozco, M.; Luque, F. J. *Chem. Phys. Lett.* **1997**, *265*, 473–480.

(61) Westergren, J.; Lindfors, L.; Höglund, T.; Lüder, K.; Nordholm, S.; Kjellander, R. *J. Phys. Chem. B* **2007**, *111*, 1872–1882.

CT9006025

JCTC Journal of Chemical Theory and Computation

Rapid Prediction of Solvation Free Energy. 2. The First-Shell Hydration (FiSH) Continuum Model

Christopher R. Corbeil, Traian Sulea, and Enrico O. Purisima*

Biotechnology Research Institute, National Research Council Canada,
6100 Royalmount Avenue, Montreal, Quebec H4P 2R2, Canada

Received November 17, 2009

Abstract: Local ordering of water in the first hydration shell around a solute is different from isotropic bulk water. This leads to effects that are captured by explicit solvation models and missed by continuum solvation models which replace the explicit waters with a continuous medium. In this paper, we introduce the First-Shell Hydration (FiSH) model as a first attempt to introduce first-shell effects within a continuum solvation framework. One such effect is charge asymmetry, which is captured by a modified electrostatic term within the FiSH model by introducing a nonlinear correction of atomic Born radii based on the induced surface charge density. A hybrid van der Waals formulation consisting of two continuum zones has been implemented. A shell of water restricted to and uniformly distributed over the solvent-accessible surface (SAS) represents the first solvation shell. A second region starting one solvent diameter away from the SAS is treated as bulk water with a uniform density function. Both the electrostatic and van der Waals terms of the FiSH model have been calibrated against linear interaction energy (LIE) data from molecular dynamics simulations. Extensive testing of the FiSH model was carried out on large hydration data sets including both simple compounds and drug-like molecules. The FiSH model accurately reproduces contributing terms, absolute predictions relative to experimental hydration free energies, and functional class trends of LIE MD simulations. Overall, the implementation of the FiSH model achieves a very acceptable performance and transferability improving over previously developed solvation models, while being complemented by a sound physical foundation.

Introduction

Molecular recognition in biological systems usually takes place in aqueous solution and is accompanied by the desolvation of the interacting surfaces and reorganization of the solvent around the ensuing complex. Hence, theoretical prediction of protein−ligand binding modes (i.e., docking) and binding affinities (i.e., scoring) require an accurate description of the change in hydration that accompanies solute binding.[1] With the advent of faster computers over the past decade, large-scale *in silico* docking-scoring (aka virtual screening) of small-molecule libraries has become appealing due to its speed and cost efficiency.[2] Unquestionably, the success (or failure) of virtual screening relies mostly on the quality of the underlying docking and scoring function(s). The challenge in virtual screening is augmented by the fact that, in order to provide a useful hit-enrichment level, accurate docking-scoring has to be achieved under the constraint of fast computing. To this end, a fast yet accurate solvation model is of paramount importance in the early stages of the drug discovery process.

Over the past years, much research has been dedicated to developing and parametrizing solvation models at various levels of theory.[3−9] Explicit-solvent models of hydration, including rigorous pathway methods such as free energy perturbation (FEP) and thermodynamic integration (TI),[1,10] or approximate end-point methods such as linear interaction energy (LIE),[11−14] address the discrete nature of water around the solute. This treatment results

* To whom correspondence should be addressed. Phone: 514-496-6343. Fax: 514-496-5143. E-mail: enrico.purisima@cnrc-nrc.gc.ca.

10.1021/ct9006037 Published 2010 by the American Chemical Society
Published on Web 04/02/2010

in transferability across a wide chemical space which is dependent on the underlying force-field. However, explicit models require molecular dynamics (MD) or Monte Carlo simulations and are therefore not practical for high-throughput applications. Implicit models of hydration (i.e., continuum models) have been precisely developed to address the speed issue, and they excel in this regard. However, this speed increase associated with continuum models has a cost, an impact on accuracy.[15−18] The current focus in the field of continuum solvation is for the development of models which can capture the underlying physics of solvation, while retaining the speed achieved by the current generation of continuum solvation models.

The local ordering of water in the first hydration shell around a solute is different from isotropic bulk water and varies depending on solute polarity. Around a hydrophobic solute surface, interactions within the first hydration shell itself are favored over interactions with the solute or with bulk solvent.[19] Around polar solute surfaces, water molecules interact strongly with the solute but orient differently around positively and negatively charged atoms, a phenomenon known as the charge asymmetry of water.[20] It is imperative for implicit solvation models to be able to capture the effects of first shell water ordering.

The philosophy adopted in this study is to develop a continuum solvation model that emulates physics-based explicit solvent models. In this way, the transferability should increase in comparison with empirical models that incorporate a large number of parameters fitted directly to experimental data.[21,22] The physical meaning of the tunable descriptors in empirical models is also often times lost. We propose here the First-Shell Hydration (FiSH) model, a continuum solvation model that reformulates the usual continuum electrostatics and van der Waals treatments in order to capture features present in the all important first shell of hydration.[23] The FiSH continuum model is designed to mimic an explicit solvent LIE model of hydration. As in the companion study using explicit solvent LIE simulation,[24] the FiSH model is applied on a large hydration data set encompassing 501 “traditional” compounds[25,26] and 63 neutral drug-like compounds from the more challenging SAMPL1 data set.[27] In the Theory and Implementation, we present improvements to the original continuum electrostatics-dispersion (CED) model of solvation[21] which have led to the development of the FiSH continuum solvation model. In the Results and Discussion section, we assess the main objective of the FiSH model, its ability to reproduce hydration free energies of the explicit-solvent LIE model. This is followed with a comparison to experimental hydration free energies and an assessment of its transferability compared to our previously developed solvation models.

Theory and Implementation

Continuum Electrostatics-Dispersion (CED) Solvation Model. Our previous attempts at formulating a continuum solvation model led to the development of the CED solvation model, which has the following functional form:[21]

$$\Delta G_{\mathrm{hyd}}^{\mathrm{CED}}(D_{\mathrm{in}}, \rho, \gamma_{\mathrm{cav}}, \{B_i\}) = \Delta G_{\mathrm{hyd}}^{\mathrm{R}}(D_{\mathrm{in}}, \rho) + \gamma_{\mathrm{cav}}\mathrm{MSA} + \sum_i U_i^{\mathrm{cvdW}}(B_i) + C \quad (1)$$

where D_{in} is the solute dielectric constant, ρ is the block-scaling factor for the AMBER van der Waals radii, γ_{cav} is the cavity surface coefficient, and $\{B_i\}$ represents the set of atom-type-dependent continuum van der Waals coefficients. These coefficients were trained to fit experimental hydration free energies of a set of 129 neutral solutes. The electrostatic contribution, $\Delta G_{\mathrm{hyd}}^{\mathrm{R}}$, was calculated using the BRI-BEM program,[28,29] which solves the Poisson equation using a boundary element method. The cavity contribution is proportional to the total molecular surface area, MSA, which was calculated using a variable surface probe.[30,31] The dispersion-repulsion term, U_i^{cvdW}, was calculated by integrating the 6−12 Lennard-Jones potential over the molecular surface[21,32,33] for a set of defined atom types, each with its own van der Waals coefficients trained to fit experimental hydration free energy. This model yielded very good results on a test set of traditional solutes similar to those used for its training. Application to the more challenging drug-like SAMPL1 data set[27] demonstrated limited transferability to more drug-like molecules. These results prompted us to change our strategy and develop a model trained primarily on explicit-water simulations instead of on experimental hydration free energies.

First-Shell Hydration (FiSH) Continuum Model Formulation. As with its CED solvation model predecessor, the FiSH model includes electrostatic, van der Waals and cavity contributions to solvation as formulated in eq 2:

$$\Delta G_{\mathrm{hyd}}^{\mathrm{FiSH}}(\{r_i^{\mathrm{Born}}\}, \gamma_{\mathrm{cav}}) = \Delta G_{\mathrm{hyd}}^{\mathrm{R}}(\{r_i^{\mathrm{Born}}\}) + U^{\mathrm{vdw}} + \gamma_{\mathrm{cav}}\mathrm{MSA} + C \quad (2)$$

The electrostatic contribution is the change in the solute reaction field energy, $\Delta G_{\mathrm{hyd}}^{\mathrm{R}}$, calculated by solving the Poisson equation in water and in vacuum dielectrics. The nonpolar hydration effects are described by the solute−solvent van der Waals interaction energy, U^{vdw}, and by the cost of cavity formation in water that is proportional to the solute molecular surface area, MSA. Even though the components of the FiSH and CED models are similar, they are obtained in different ways.

FiSH Born Radii. The FiSH continuum electrostatic term uses atomic Born radii $\{r_i^{\mathrm{Born}}\}$, derived from general corrections to the van der Waals radii of atoms in a molecule that restore the asymmetric response of water to solutes of different polarities.[34−37] The charge asymmetry phenomenon is dominated by first solvation shell effects.[20] Water molecules orient differently around positively and negatively charged atoms, resulting in changes in the dielectric boundaries (Figure 1). This leads to different effective Born radii and an asymmetric dependence of the reaction field energies on solute charge. Charge asymmetries are captured by explicit water simulations, but the usual continuum electrostatics calculations fail miserably in capturing this phenomenon.[20,38] We have recently presented a proof of concept, in which charge asymmetry of solvation can be handled in

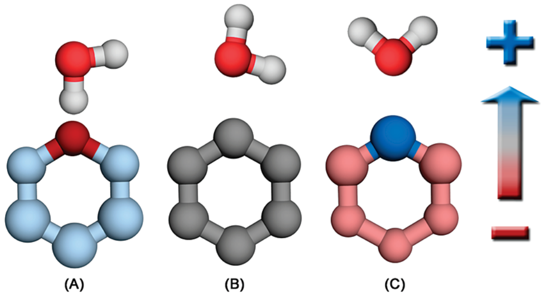


Figure 1. Schematic showing the dependence of the dielectric boundary on orientation of a water molecule around the dominant charge of a neutral hexagonal bracelet model.[20] (A) Bracelet with negative dominant charge (red = −1.0*e* charge, blue = +0.2*e* charge). (B) Uncharged bracelet (gray = 0.0*e* charge). (C) Bracelet with positive dominant charge (blue = +1.0*e* charge, red = −0.2*e* charge).

a simple, systematic, and transferable way within a purely continuum electrostatics framework.[23] In this approach, we used the average induced surface charge density (ISCD), σ_i, obtained from a boundary element solution of the Poisson equation to derive a simple linear correction to the van der Waals radius to obtain the Born radius, r_i^{Born}, for each atom, *i,* of a molecule.[23]

$$r_i^{\mathrm{Born}} = \begin{cases} r_i^0 - c_+\sigma_i \text{ if } \sigma_i \geq 0 \\ r_i^0 - c_-\sigma_i \text{ if } \sigma_i < 0 \end{cases} \tag{3}$$

To obtain the σ_i for eq 3, all atoms are initially assigned Born radii equal to the General AMBER Force Field (GAFF)[39] van der Waals radii r_i^0. From the boundary element solution to the Poisson equation, the average ISCDs for each atom are then calculated by assigning the surface patches and their associated charge density to the nearest atom. Only atoms with solvent exposure have their Born radii modified from the initial van der Waals radii since only these atoms define the molecular surface. However, it should be noted that the correction embodied in eq 3 is based on a molecular property, the ISCD, and not just on a local effect. Thus, even buried atoms, whose Born radii remain unchanged, can affect the Born radii of surface atoms because of their influence on the molecule's ISCD. The two coefficients, c_+ and c_-, used for positive and negative σ_i were previously trained on the electrostatic free energy from FEP simulations for a set of model systems consisting of pairs of neutral hexagonal bracelets with mirrored charge distribution (Figure 1).[20] Tests on pairs of model systems with different geometries indicated the generality of the approach and the transferability of the calibrated coefficients.[23] However, the c_+ and c_- coefficients derived previously were for highly simplified model systems made of a single atom type. Thus, in this work, we retrained the continuum electrostatic coefficients, c_+ and c_-, to the explicit-solvent LIE electrostatic component for the training data set of 200 neutral molecules and obtained slightly different values of 16.222 Å^3/*e* and 11.843 Å^3/*e* for c_+ and c_-, respectively, compared to the previous values of 15.5 Å^3/*e* and 11.5 Å^3/*e*.[23]

The relatively small change in c_+ and c_- coefficients in going from simple model systems to a 200-molecule training set suggests that the coefficients are not overly sensitive to the atom types, at least as far as neutral molecules go. It also suggests that the linear correction in eq 3 may perform relatively well for the normal range of partial charges observed in neutral real molecules. However, that approximation has its limitations. We expect the linear dependence to level off at some point or even reverse in the case of a negative σ_i. At moderately large negative σ_i, the Born radius is larger than the Lennard-Jones radius because this reflects the orientation of the first solvation shell water molecule with the water hydrogen atoms pointing away from the surface. However, as the ISCD becomes even more negative (i.e., the solute electrostatic potential in that region becomes more positive), the water molecule will be drawn closer to the solute molecular surface, and the effective Born radius should decrease. For positive σ_i, increases in the value of σ_i are associated with a decrease in the Born radius as the hydrogen of the water molecule is pulled closer to the solute, effectively decreasing the Born radius. As these decreases become larger, a leveling off should occur since the van der Waals repulsion will start to become significant. Also, we expect the coefficients to depend upon the well depth of the Lennard-Jones potential. These limitations of the linear functional form in eq 3 motivated an exploration of a nonlinear dependence of the Born radii on the ISCD, as discussed below.

Nonlinear Dependence on ISCD. The dependence of the Born radii on the ISCD and van der Waals parameters can be examined using simple spherical solutes of varying partial charges, *q*, Lennard-Jones well depths, *ε*, and van der Waals radii, r^0. Spherical solutes are ideal to investigate the shape of the nonlinearity with respect to ISCD since effective Born radii can be obtained directly from the Born equation.[40] In Figure 2, we plot the difference between the effective Born radii and the original van der Waals radii of these model spherical solutes versus ISCD. Born radii were obtained by fitting the analytical expression for the reaction field energy of a spherical ion to that calculated with the LIE approach based on MD simulations in explicit water (see Materials and Methods). The results shown in Figure 2 indicate that the nonlinear dependence on the ISCD follows the expected behavior described above. Figure 2A shows the dependence of the radius correction on the Lennard-Jones well depth, *ε*, at a fixed van der Waals radius. The data points for each well depth define roughly parallel curves. For a given van der Waals radius, the radius correction becomes more negative for smaller well depths, *ε*, (Figure 2A) across the entire range of *σ*, which correspond to a series of partial charges from −1 to +1. This behavior is understandable due to the closer approach of water in the case of a "softer" solute sphere (i.e., smaller well depth). Figure 2B shows the dependence of the radius correction on the van der Waals radius at a fixed well depth. The variation among the different curves seems to be more pronounced for negative *σ* compared to positive ones. The radius correction is more negative for larger r^0, but only marginally so for positive *σ*. These changes in Born radii are size effects due to geometrical restrictions of accommodating discrete water molecules in the first solvation shell around very small solutes.

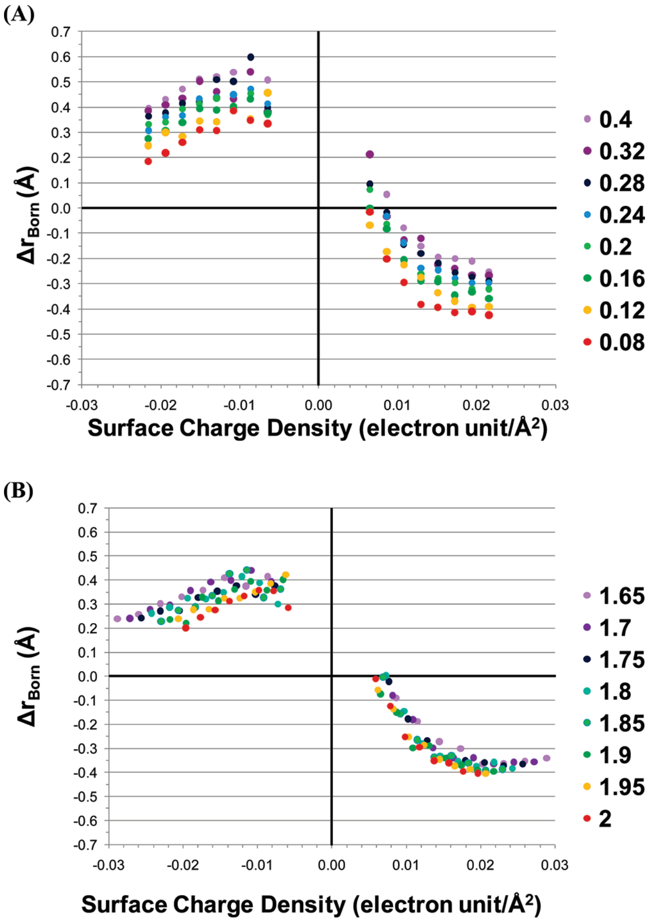


Figure 2. Calculated change in Born radius with induced surface charge density and its dependence on solute van der Waals parameters for singly charged spherical solutes. (A) Effect of the well depth of the 6−12 Lennard-Jones potential, ε (in kcal/mol), on the Born radius. (B) Effect of the equilibrium radius of the 6−12 Lennard-Jones potential, r^0 (in Å), on the Born radius. See the Theory and Implementation section for details.

This leads to a compensatory effect of maintaining a certain Born radius as the van der Waals radius decreases.

The nonlinear dependence on ISCD, the direct dependence on the well depth, and the inverse dependence on van der Waals radii led us to consider a functional form based on combining the arctan and Gaussian functions:

$$r_i^{\mathrm{Born}} = r_i^0 + A \arctan\,(B\sigma_i + C) + \frac{D}{r_i^0}\exp\left[\frac{\left(\sigma_i - \dfrac{E}{r_i^0}\right)^2}{2\left(\dfrac{F}{r_i^0}\right)^2}\right] + G\varepsilon_i + \frac{H}{r_i^0} \quad (4)$$

where A, B, C, D, E, F, G, and H are fitted parameters. This allows us to easily relate the shape of the correction function to the underlying physical interactions. The arctan dependence of the Born radii on the ISCD relates to the water hydrogen orientation around a positively or negatively charged solute atom assuming the water oxygen atom is at a fixed contact distance to the solute atom. The shifted arctan is a more sophisticated version of the linear functions in eq 3. The Gaussian dependence of the Born radii on the ISCD emulates the attraction of the entire water molecule as the partial charge of the solute atom increases (irrespective of sign) and the limitation of the solute−solvent approach due to the Lennard-Jones repulsion. We noticed that the arctan component is fairly constant on the negative ISCD, allowing the nonlinear correction to be mostly controlled by the Gaussian component. The reverse is true for positive ISCD. These features should enable this function to capture all the aspects seen in Figure 2, including the hump at small negative σ. In terms of the dependence on Lennard-Jones parameters of the solute, the arctan−Gaussian function shifts the Born radius up with increasing well depth (ε) and decreasing size (r^0), arising mainly from the last two terms in eq 4. The inverse dependence on size is also included in the Gaussian component, critical at negative ISCD.

Even though the form of the arctan and Gaussian function allows for an interpretation in terms of the underlying physics, there is a danger of overfitting due to the large number of parameters. Hence, an alternate simpler functional form, a rational function, was also explored:

$$r_i^{\mathrm{Born}} = r_i^0 + \frac{A\sigma_i + B\left(\dfrac{\varepsilon_i}{r_i^0}\right)}{D\sigma_i^2 + E\sigma_i + 1} \quad (5)$$

where A, B, D, and E are fitted parameters. Regarding the Born radius dependence on Lennard-Jones parameters, this rational function essentially shifts the Born radius correction up and down with the solute softness and size, respectively. The advantages of the rational function are the good quality of the fit with a lower number of fitting parameters compared to eq 4.

We note that both nonlinear correction functions report a Born radius for an uncharged spherical solute that is larger than its van der Waals radius, r_i^0. There is no physical reason why these radii should be identical in the case of uncharged solutes. In fact, for a convex solute, a slightly larger Born radius than the van der Waals radius may be expected because the hydrogen density from the first shell of waters would be located at a slightly greater distance from the surface than the oxygen density. This effect is especially pronounced for small spherical solutes, but still present around an uncharged protein using long MD simulations in SPC/E water[41] where hydrogen densities were 0.1 Å further away from the protein than peak oxygen densities for the first solvation shell. Also, interpolation of our data on varying the size of spherical solutes (Figure 2B) suggests slightly larger increases of Born radii for the uncharged spheres with smaller van der Waals radii (i.e., more convex).

The purpose of the calculations on the spherical model solutes was simply to guide the selection of the nonlinear functional form to use. All parameters in the two nonlinear functions were later retrained on real molecules from the training set against the electrostatic component of solute−solvent interaction energy from explicit water MD simulations using the LIE approximation.

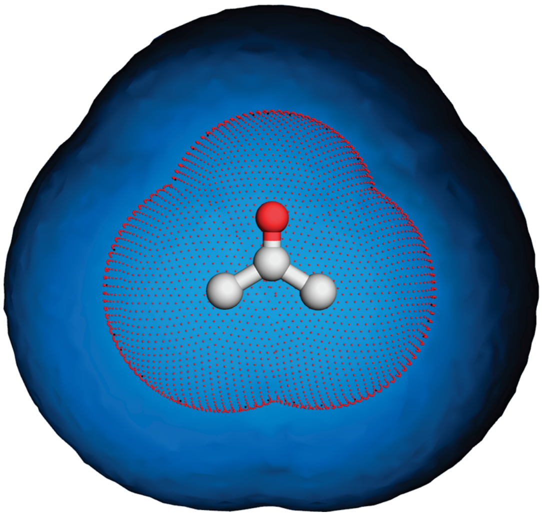


Figure 3. Illustration of the two regions defined for the hybrid van der Waals component of the FiSH continuum model, using acetone as an example. Red dots represent the first shell of water oxygen atoms uniformly dispersed over the solvent-accessible surface (SAS). The blue surface represents where the outer region of continuum solvent starts (SAS + 2.8 Å).

Continuum van der Waals Model. In the usual continuum van der Waals model, the solute−solvent van der Waals interactions are obtained from the integral of the solute-continuum interactions over all of space, with the volume integral transformed into a surface integral at the solute−solvent boundary represented by solvent-accessible surface or molecular surface.[21,32,33] In this approach, a uniform density function for the solvent is assumed. Clearly, this is a gross approximation for the first hydration shell. Partly due to this, scaling coefficients are typically required to adjust the continuum van der Waals energy to the magnitude of explicit solute−solvent interactions or experimental data.[42] Here, in order to avoid empirical scaling of atom-type-based van der Waals parameters as in the previous CED model and to mimic more closely the important first solvation shell interactions from explicit solvent simulations, we devised a two-region solvent model for the calculation of Lennard-Jones interactions with the solute (Figure 3).

$$U^{\mathrm{vdW}} = U_1^{\mathrm{vdW}} + U_2^{\mathrm{vdW}} \qquad (6)$$

The basic idea is that, for the first shell, which is represented by the solvent-accessible surface (SAS), we assume that the water oxygen is completely restricted to lie on the SAS but is uniformly distributed along the surface. We take the second and succeeding shells to start 2.8 Å (a water diameter) away from the SAS and to be uniformly distributed from that point out to infinity. The contribution of the first shell, U_1^{vdw}, is then calculated as a discretized integral between the solute atoms i and the surface distribution of the TIP3P water oxygen atoms.

$$U_1^{\mathrm{vdW}} = \rho_{\mathrm{S}} \sum_i^{\mathrm{atoms}} \sum_j^{\mathrm{patches}} \left(\frac{A_{iw}}{r_{ij}^{12}} - \frac{B_{iw}}{r_{ij}^{6}} \right) \mathrm{SA}_j \qquad (7)$$

ρ_{S} is the number density of water along the surface. The A_{iw} and B_{iw} coefficients are from TIP3P and the general AMBER force field (GAFF).[39] SA_j is the area of the triangulated surface patch j.

The second region in our continuum van der Waals model is constructed by extending the SAS by 2.8 Å, i.e., one water diameter (Figure 3). The solvent density is assumed to be approximately uniform from this point onward, allowing the dispersion (attractive) component to be computed as a discretized surface integral in the usual way:

$$U_2^{\mathrm{vdW}} = -\rho_{\mathrm{N}} \sum_i^{\mathrm{atoms}} \sum_j^{\mathrm{patches}} \frac{1}{3} \frac{B_{iw}}{r_{ij}^{6}} \mathbf{r}_{ij} \cdot \mathbf{n}_j SA_j \qquad (8)$$

where $\mathbf{r}_{ij}$ is the vector from solute atom i to boundary surface patch j, $\mathbf{n}_j$ is the unit surface normal at j, SA_j is the area of patch j, and ρ_{N} is the solvent number density of bulk water. The atomic dispersion parameters B_{iw} are taken from the GAFF force field and TIP3P. Ignoring the repulsive r^{-12} contribution saves some computation time without introducing much error. It should be noted that no scaling or fitting of the $U_2{}^{\mathrm{vdw}}$ term is carried out.

A key component of this approach is that the SAS is constructed using solute atom-specific solvent probe radii. The starting point for defining the atom-specific solvent probe radii is the location of the first peak for various atom types in the radial distribution function (RDF) determined from MD simulations in explicit water for the training set. Specifically, we determine average first RDF peak distances between the water oxygen and the GAFF atom types and use that to define the atom-specific solvent probe radii. The SAS is then generated by operationally inflating the force field van der Waals atomic radii by the appropriate probe radii (Table 1). Additional manual fine-tuning of the radii is carried out in order to improve the agreement between U_1^{vdw} and the average solute−solvent van der Waals interaction energy with an effective first hydration shell defined as all water molecules with oxygen centers not farther than the SAS + 2.8 Å (a water diameter), calculated with the LIE approach based on MD simulations in explicit water.

Materials and Methods

Hydration Data Sets. A data set consisting of experimental hydration free energies for 501 neutral organic small molecules compiled from the published literature[25] was used as prepared in a previous study.[24] As in the previous study, the conformations used for implicit solvation predictions correspond to the conformation with the best potential energy in a vacuum, which have been shown to reproduce well hydration free energies in explicit-solvent models.[21,26] This data set was split into a training set of 200 compounds and a testing data set of 301 compounds. The training data set was used for calibrating the cost of cavity formation in water against experimental hydration free energy data, and for calibrating the electrostatic and van der Waals components of the FiSH continuum model against calculated explicit-solvent LIE data. In the training set, we included mostly rigid representatives of the various chemical classes, with the majority of compounds being monofunctional, and only a

Table 1. First RDF Peak for Various GAFF Atom Types[a]

atom type	RDF peak	r^o	solvent probe radius	atom type	RDF peak	r^o	solvent probe radius
c	3.20	1.908	1.292	f	3.15	1.750	1.400
c1	3.25	1.908	1.342	cl	3.40	1.948	1.452
c2	3.20	1.908	1.292	br	3.75	2.220	1.530
c3	3.30	1.908	1.392	i	3.85	2.350	1.500
ca	3.25	1.908	1.342	n	3.15	1.824	1.326
cp	3.25	1.908	1.342	n1	3.25	1.824	1.426
cq	3.25	1.908	1.342	n2	2.80	1.824	0.976
cc	3.25	1.908	1.342	n3	3.03	1.824	1.206
cd	3.25	1.908	1.342	na	2.95	1.824	1.126
ce	3.25	1.908	1.342	nb	2.95	1.824	1.126
cf	3.25	1.908	1.342	nc	2.95	1.824	1.126
cg	3.55	1.908	1.642	nd	2.95	1.824	1.126
ch	3.55	1.908	1.642	ne	2.80	1.824	0.976
cx	3.30	1.908	1.392	nf	2.80	1.824	0.976
cy	3.30	1.908	1.392	nh	3.10	1.824	1.276
cu	3.60	1.908	1.692	no	3.95	1.824	2.126
cv	3.60	1.908	1.692	o	2.92	1.661	1.259
h1	2.75	1.387	1.363	oh	2.92	1.721	1.199
h2	1.00	1.287	−0.287	os	2.95	1.684	1.266
h3	1.00	1.187	−0.187	ow	2.75	1.768	0.982
h4	2.75	1.409	1.341	p5	3.20	2.100	1.100
h5	2.75	1.359	1.391	s	3.65	2.000	1.650
ha	2.82	1.459	1.361	s4	3.30	2.000	1.300
hc	2.82	1.487	1.333	s6	3.30	2.000	1.300
hn	1.80	0.600	1.200	sh	3.40	2.000	1.400
ho	1.00	0.000	0.000	ss	3.40	2.000	1.400
hs	1.00	0.600	0.400	sx	3.65	2.000	1.650
hx	1.00	0.000	1.000	sy	4.05	2.000	2.050

[a] All values are in Ångstroms. The SAS is constructed using r^0 + solvent probe radius. For hydrogens with small or even negative solvent probe radii, this simply means that the SAS is entirely determined by the heavy atom to which it is connected and the RDF value is a dummy one.

few polyfunctional compounds were included to increase coverage of some functional groups. The testing data set mirrors the training data set in terms of chemical class representation for monofunctional compounds but differs from the training analogs by having increased flexibility and containing a larger collection of polyfunctional compounds. We also used the more challenging SAMPL1 data set[27] consisting of 63 drug-like, diverse, polyfunctional, neutral polar compounds, which spans wider ranges of transfer free energies and molecular weights in comparison to the training and testing data sets. The SAMPL1 data set was also used as prepared in our previous study.[24] Details on the composition of the training and testing and SAMPL1 data sets are provided as Supporting Information (Table S1). A functional group analysis was carried out using the testing set. We used the definitions of groups used in the previous companion paper.[24]

LIE Data and MD Simulations. The LIE data for the 564 compounds in the training, testing, and SAMPL1 data sets were taken from the companion study,[24] in which the following implementation of the LIE approximation was used:

$$\Delta G_{\text{hyd}}^{\text{LIE}} = \underbrace{\alpha(\langle E_{\text{S-W}}^{\text{Coul}}\rangle_{\le 12\text{Å}} + \langle G_{\text{S}}^{\text{RF}}\rangle_{12\text{Å-}\infty})}_{electrostatic} + \underbrace{\beta(\langle E_{\text{S-W}}^{\text{vdW}}\rangle_{\le 12\text{Å}} + \langle E_{\text{S}}^{\text{cvdW}}\rangle_{12\text{Å-}\infty})}_{van\ der\ Waals} + \underbrace{\gamma_{\text{cav}}\langle MSA\rangle + C}_{cavity} \tag{9}$$

From the various LIE models derived and described in the companion paper,[24] for this study, we have taken LIE data generated for rigid solute geometries at various partial charge models (primarily AM1BCC-SP, but also AM1BCC-OPT and RESP), with continuum corrections beyond the explicit water shell as described.[24] These data were considered most suitable for the calibration of a continuum solvation model described in this study. LIE simulations were favored over FEP-like methods for training due to their simpler decomposition of the electrostatic and van der Waals component along with a slightly improved accuracy on the testing and SAMPL1 data set.[24] It can be argued that the electrostatic component in the FEP method is possibly contaminated with a net positive change in the solute−solvent van der Waals interaction energy upon solute charging that is embedded into the FEP "electrostatic" component,[24,43,44] although this interpretation is a matter of some debate.[45] The LIE data based on rigid-solute geometries were selected for FiSH continuum model training since the rigid paradigm is often used by implicit solvation models. Hydration free energy predictions can be greatly affected by the choice of solute conformation and the degree of flexibility of the investigated molecules. Therefore, in principle, rigid-solute simulation data should streamline the training and the comparison of an implicit solvation model against a more rigorous explicit-water model.

We also generated LIE data for spherical model solutes that were used to elucidate the nonlinear dependence of atomic Born radii on the ISCD. Spherical model systems were created by varying their van der Waals radius, r^0, from 1.65 Å up to 2.00 Å with 0.05 Å increments while keeping the Lennard-Jones potential well depth, ε, at 0.1094 kcal/

mol (corresponding to the GAFF atom type c3), and by varying ε from 0.08 kcal/mol up to 0.32 kcal/mol with 0.04 kcal/mol increments and including 0.40 kcal/mol while keeping r^0 at 1.908 Å (for GAFF atom type c3). The spherical model solutes had a single atom-centered charge which systematically varied between $-1e$ and $+1e$ with $0.1e$ increments.

MD simulations were carried out with the AMBER 9 software[46] applying the systematically varied parameters described above for the spherical solute. The spherical model systems were solvated in an octahedron of TIP3P water[47,48] extending 13 Å around the solute. The system was energy-minimized first, followed by heating from 100 K to 300 K over 25 ps in the canonical ensemble (NVT), and equilibrating to adjust the solvent density under 1 atm of pressure over 25 ps in the isothermal−isobaric ensemble (NPT) simulation. A 1 ns production NPT run was obtained with snapshots collected every 10 ps, using a 2 fs time-step and 9 Å nonbonded cutoff. The Particle Mesh Ewald (PME) method[49] was used to treat long-range electrostatic interactions, and bond lengths involving bonds to hydrogen atoms were constrained by SHAKE.[50]

Continuum Electrostatic Calculations. Reaction field energies were calculated for a single conformer of each solute molecule using the BRI BEM program, which solves the Poisson equation using a boundary element method.[28,29] The solute and solvent dielectric constants were taken to be 1.0 and 78.5, respectively. The dielectric boundary was taken to be the solvent-excluded surface (also known as the molecular surface) as generated and triangulated using a marching tetrahedra algorithm and a solvent probe radius of 1.4 Å.[30,31] The induced surface charge density (ISCD) distribution at the dielectric boundary was automatically obtained as part of the solution of the Poisson equation. The atom-based ISCD was determined by assigning surface patches to the nearest atom and averaging the ISCDs of the patches associated with a particular atom. All calculations of the ISCD-based Born radii (eqs 3−5) used GAFF[39] van der Waals radii as the initial value, r^0.

Parameter Fitting. Optimization of parameters in the linear and nonlinear correction functions (eqs 3−5) was carried out in order to minimize the mean unsigned error (MUE) of the electrostatic component of solvation calculated with a continuum model from that calculated with an explicit-solvent LIE model, for a given set of compounds.

In the case of spherical model solutes, parameter optimization for the nonlinear models (eqs 4 and 5) was carried out with the Solver plug-in in Microsoft Excel. These values were then used as starting points for parameter optimization against the training data set of real molecules, for which the Nelder−Meade (aka downhill simplex) algorithm using the TCL8.4 math::optimize library was employed. Optimized parameters in eqs 3−5 are given as Supporting Information (Table S2). Bootstrapped statistical analyses were carried out for 5000 samples using the boot library within the R software.[51] In the case of the linear function in eq 3, initial values for the c_+ and c_- parameters were taken from our previous fitting to hexagonal neutral bracelets as model compounds.[23]

Other Continuum Solvation Models. The transferability of the FiSH model will be assessed by comparison to previously developed continuum solvation models, a continuum electrostatics-dispersion (CED) model and a continuum model consisting of only reaction field electrostatics (RF), both of which have been developed and used previously.[21] The CED model consists of continuum reaction field electrostatics, continuum solute−solvent van der Waals interactions, and surface-area-based cavity cost. Unlike in FiSH, where the parameters were trained on explicit water simulation, the parameters were calibrated against the experimental hydration free energy data.[21] The CED model uses a solute dielectric of 1, Born radii that were 0.9 of the AMBER van der Waals radii, and a continuum van der Waals model with 25 atom types, all of which were taken from the previous study. Since the CED parametrization lacked continuum van der Waals parameters for the iodine atom, CED predictions were not obtained for all molecules from the current data sets containing iodine. In the RF model, the scanning of the scalar for the AMBER van der Waals radii, used as the Born radii, and solute dielectric parameters in a previous study suggest optimal values of 1.1 and 1, respectively, for acceptable and transferable prediction of small-molecule hydration.[21] These values are then used for reaction field calculations by solving the Poisson equation using a boundary element method. Both models were used as implemented within the BRI-BEM program.

Results and Discussion

We will begin by presenting the data obtained for new formulations of the electrostatic and van der Waals components of the FiSH model, which were calibrated and tested against explicit-water LIE data. We will then analyze the cavity cost and the total nonpolar contribution to solvation vis-à-vis the solute surface area. The performance of the generated FiSH models versus LIE explicit models, experimental data, and earlier continuum models will be presented in detail. Functional group analysis of errors will be used to detect trends in the FiSH model predictions.

Electrostatic Component of the FiSH Model. As seen in Table 2, even the parametrization of charge-asymmetry-corrected continuum electrostatics on the spherical model alone improved significantly the agreement with the explicit-solvent electrostatic solvation data across the three molecule sets—training, testing, and SAMPL1 relative to using the GAFF radii. For example, in the linear model, the MUEs go from 1.676, 1.557, and 3.506 kcal/mol to 0.513, 0.479, and 0.731 kcal/mol, respectively, for the three sets. Similar improvements can also be seen for the slope and squared correlation coefficient (R^2).

Parametrization on real molecules further improves significantly the agreement between the explicit and continuum models of electrostatic hydration with nonlinear correction functions and marginally with the linear correction function. Results on the testing and SAMPL1 data sets indicate similar performances for the three correction functions: MUE values below 0.5 kcal/mol on the testing data set and slightly above 0.7 kcal/mol for the SAMPL1 data set, in all cases highly

Table 2. Comparing the Electrostatic Component of FiSH Models with the Electrostatic Component of the LIE Explicit-Solvent Model[a]

	r^{Born}	trained on spherical solutes			trained on molecules		
set	correction	MUE	slope	R^2	MUE	slope	R^2
training	original[b]	1.676 ± 0.090	1.348 ± 0.038	0.889 ± 0.016			
	linear[c]	0.513 ± 0.033	0.980 ± 0.015	0.957 ± 0.007	0.506 ± 0.033	0.982 ± 0.014	0.958 ± 0.007
	Atan+Gauss[d]	1.196 ± 0.078	1.149 ± 0.028	0.900 ± 0.013	0.480 ± 0.031	0.978 ± 0.015	0.962 ± 0.006
	rational [e]	0.859 ± 0.065	1.153 ± 0.020	0.935 ± 0.010	0.531 ± 0.035	0.977 ± 0.008	0.953 ± 0.008
testing	original	1.557 ± 0.071	1.434 ± 0.029	0.918 ± 0.011			
	linear	0.479 ± 0.023	1.022 ± 0.012	0.964 ± 0.004	0.468 ± 0.022	1.018 ± 0.012	0.965 ± 0.004
	Atan+Gauss	1.043 ± 0.060	1.211 ± 0.021	0.920 ± 0.009	0.414 ± 0.021	1.001 ± 0.014	0.970 ± 0.003
	rational	0.758 ± 0.046	1.175 ± 0.015	0.953 ± 0.006	0.444 ± 0.024	0.988 ± 0.014	0.963 ± 0.005
SAMPL1	original	3.506 ± 0.319	1.490 ± 0.025	0.967 ± 0.011			
	linear	0.731 ± 0.077	1.029 ± 0.020	0.982 ± 0.006	0.704 ± 0.072	1.027 ± 0.019	0.983 ± 0.005
	Atan+Gauss	2.732 ± 0.222	1.315 ± 0.027	0.962 ± 0.011	0.732 ± 0.074	1.055 ± 0.014	0.985 ± 0.004
	rational	2.219 ± 0.237	1.281 ± 0.024	0.974 ± 0.008	0.734 ± 0.077	1.005 ± 0.019	0.981 ± 0.006

[a] Statistics are given as averages ± standard deviation for 5000 bootstrapped samples. Errors are in kcal mol^{-1} units. [b] GAFF vdW radii. [c] Equation 3. [d] Equation 4. [e] Equation 5.

correlative and with slopes very close to unity. By comparison, the original Born-radius uncorrected continuum electrostatic model differed from the LIE explicit-solvent electrostatic model by MUEs larger than 1.5 kcal/mol for the training and testing data set and 3.5 kcal/mol for the SAMPL1 data set. Throughout the rest of the paper, all results discussed or presented will be with the parameters that have been trained on molecules.

These results highlight the benefits of ISCD-dependent Born radii to account for charge asymmetry effects, as well as the improvements afforded by training on real molecules for the nonlinear model. The linear correction function appears to provide robust and competitive results when compared with the nonlinear correction functions. However, as presented in the Theory and Implementation section, our computational experiments on spheres clearly show that correction should be nonlinear with respect to the induced surface charge density. The linear function most likely gives good results since the ISCD range explored by the neutral molecules in our data sets is rather narrow (between −0.01 and +0.01 $e/\text{Å}^2$, see Figure S2, Supporting Information), for which the linear approximation is still applicable (see Figure 2). With charged molecules, the range of ISCD will be expanded and the linear correction will most likely fail. For example, the nitrogen of a terminal alkyl ammonium would have an ISCD of around −0.025 $e/\text{Å}^2$, which falls outside of the linear region seen in Figure 2 and justifies the use of a nonlinear function. A more complete study (outside the scope of this work) is needed for charged molecules, but for now, the nonlinear model seems most appropriate because of its greater generality. Given the comparable performances obtained with the two nonlinear correction functions, the rational function (eq 5) is preferred over the arctan + Gaussian function (eq 4) due to a lower number of fitted parameters and will be featured for the rest of the paper. The correlation between the FiSH continuum electrostatic component and the explicit-solvent LIE electrostatic term for the training, testing, and SAMPL1 data sets is shown in Figure 4.

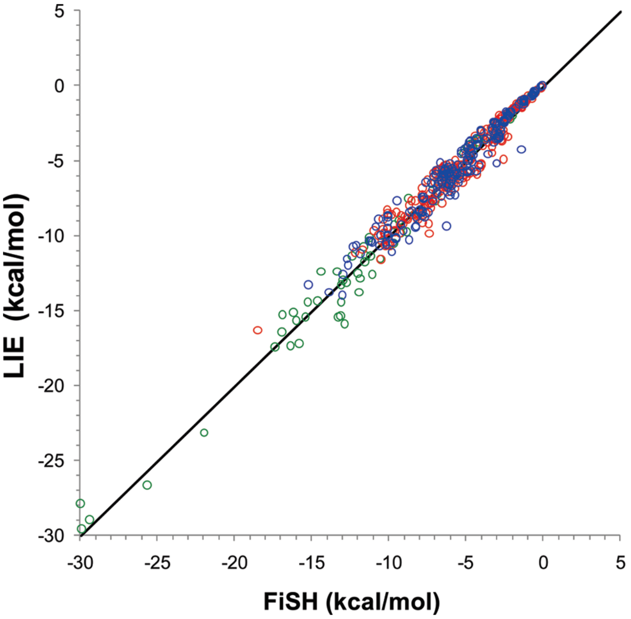


Figure 4. Comparison between the electrostatic component of the FiSH model and the electrostatic component of the LIE explicit-solvent model on the training (blue symbols), testing (red symbols), and SAMPL1 data sets (green symbols). The FiSH continuum electrostatic component is calculated with the optimized rational function (eq 5) for the Born radii. LIE data on single-conformation solutes are taken from a separate study.[24] Both models are derived using AM1BCC-SP partial charges. The diagonal line indicates ideal correlation.

van der Waals Component of the FiSH Model. Comparison with LIE data from explicit-solvent MD simulations with AM1BCC-SP solute charges and rigid solute geometries indicates that our two-zone continuum van der Waals model reproduces the explicit-solvent van der Waals contribution to solvation with MUEs below 0.6 kcal/mol for the testing set and about 1.4 kcal/mol for the SAMPL1 data set (Table 3). The model slightly overestimates the explicit-solvent van der Waals interactions, especially for the SAMPL1 molecules (Figure 5). This may reflect some additivity problems in the continuum model for highly polyfunctional molecules. The FiSH model addresses some of the nonhomogeneity of the solvent distribution function in directions radially away from the solute surface. However, it still assumes a uniform distribution tangential to the solute surface in the first shell.

Table 3. Comparing the van der Waals Component of the FiSH Model against the van der Waals Component of the LIE Explicit-Solvent Model

set	MUE[a]	slope	R^2
training	0.519 ± 0.034	0.900 ± 0.012	0.946 ± 0.004
testing	0.584 ± 0.025	0.882 ± 0.010	0.966 ± 0.004
SAMPL1	1.403 ± 0.110	0.884 ± 0.020	0.925 ± 0.110

[a] Errors are in kcal mol^{-1}.

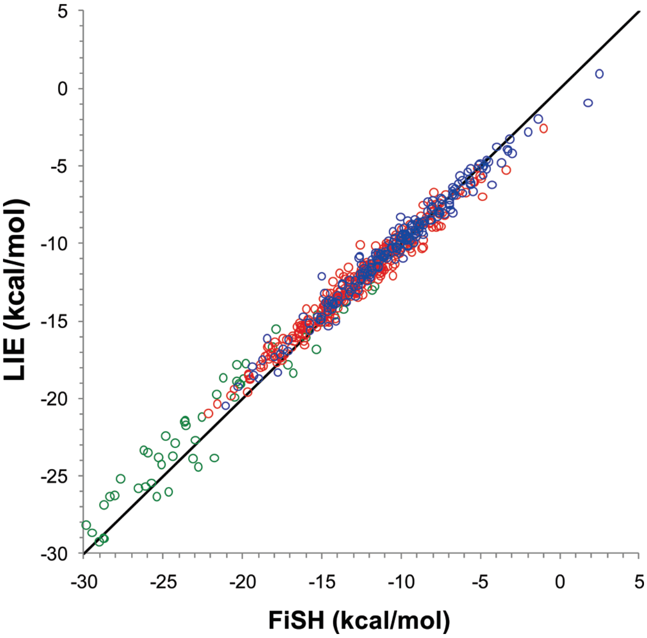


Figure 5. Comparison between the van der Waals component of the FiSH model and the van der Waals component of the LIE explicit-solvent model on the training (blue symbols), testing (red symbols), and SAMPL1 data sets (green symbols). LIE data on single-conformation solutes and AM1BCC-SP partial charges are taken from a separate study.[24] The diagonal line indicates ideal correlation.

This can be a gross approximation with ordered water molecules that may be present in highly polyfunctional molecules.

Cavity and Total Nonpolar Contributions of the FiSH Model. The cost of cavity formation in water was treated as a linear dependence on the molecular surface area, MSA, of the solute (eq 2) and fitted to a pseudoexperimental cavity cost for the training data set. This cost was obtained by subtracting the FiSH continuum electrostatic and van der Waals contributions, described earlier, from the experimental hydration free energy. A robust linear relationship was obtained (Figure 6), characterized by a bootstrapped correlation coefficient of 0.906 ± 0.016 and a slope and intercept (γ and C, respectively, in eq 2) of 0.115 ± 0.003 kcal mol^{-1} Å^{-2} and −4.276 ± 0.386 kcal/mol, respectively (Table 4). We note that the microscopic surface tension of water, γ, is close to the macroscopic one, reiterating the findings obtained with the LIE explicit-solvent model.[24] This linear relationship extends very well to the testing set and SAMPL1 data set (Figure 6), which is further supported by similar cavity parameters γ and C derived by fitting directly to these data sets. The slightly larger slope (γ) obtained in the case of the SAMPL1 data set is partially due to a few sulfoneurea analogs with larger MSA values. Overall, the data obtained for the cavity component of the FiSH continuum model mirror closely those obtained in a study of LIE models of hydration.[24] A direct comparison between the cavity parameters with the FiSH continuum model and the corresponding LIE explicit model is also given in Table 4, where the presented LIE data were derived using AM1BCC-SP partial charges and single-conformation geometries for the solutes. We see that the macroscopic surface tension is consistently slightly larger for the FiSH continuum model relative to the LIE explicit model. This compensates for the modest underestimation of the explicit solute−solvent van der Waals interactions by the FiSH continuum model (Figure 5, Table 3). Intercepts are also consistently more negative in the case of the FiSH continuum model relative to the LIE explicit model.

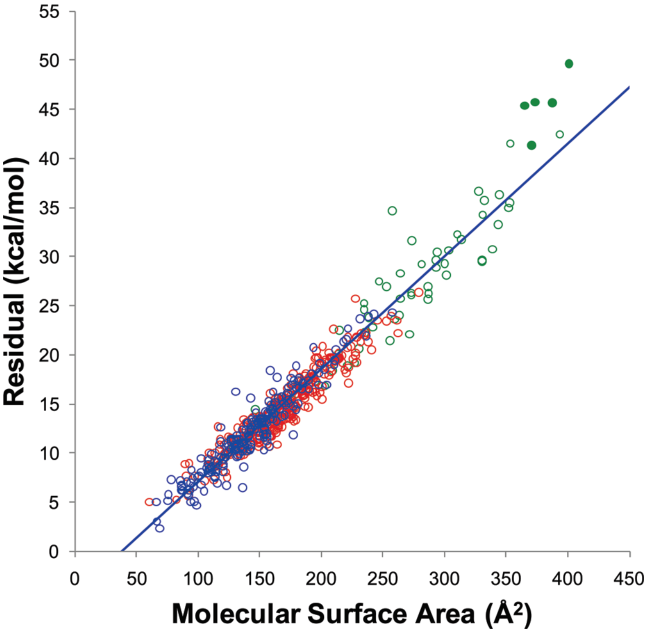


Figure 6. Deriving the cavity contribution for the FiSH model. Linear relationship between pseudoexperimental (residual) cavity contribution and the MSA for the training (blue symbols) and testing (red symbols) data sets and for the SAMPL1 (green symbols) data set. Only the regression line for the training data set is shown, since this is used to predict the cavity contribution for the testing and SAMPL1 data sets. Filled circle points correspond to 5 sulfoneurea analogs from the SAMPL1 data set.

The total nonpolar solvation component, i.e., the van der Waals contribution plus cavity cost, does not correlate with the solute MSA, due to strong anticorrelation between these contributions leading to cancellation of large opposing numbers (Figure S3, Supporting Information). A weak correlation is seen only in the case of the SAMPL1 data set (Figure S3B). These results mirror LIE data from a previous study demonstrating the FiSH continuum solvation model's ability to mimic an explicit-solvent model.[24] Together with earlier reports from FEP calculations,[25,52] these results stress the requirement for separate accounting of van der Waals and cavity terms.

The FiSH model draws its roots from a continuum electrostatics-dispersion (CED) solvation model,[21] which we

Table 4. Parameters for the Cavity Cost That Can Be Derived from Linear Relationships between the Pseudo-Experimental (Residual) Cavity *versus* the Solute Molecular Surface Area, for the Indicated Hydration Data Sets[a]

	FiSH			LIE		
set	slope (γ)	intercept (C)	R^2	slope (γ)	intercept (C)	R^2
training	0.115 ± 0.003	−4.276 ± 0.386	0.906 ± 0.016	0.108 ± 0.002	−3.488 ± 0.282	0.923 ± 0.015
testing	0.103 ± 0.002	−2.739 ± 0.346	0.903 ± 0.011	0.095 ± 0.002	−1.674 ± 0.291	0.913 ± 0.009
SAMPL1	0.127 ± 0.006	−7.204 ± 1.378	0.902 ± 0.025	0.118 ± 0.005	−5.625 ± 1.231	0.904 ± 0.026

[a] γ is in kcal mol^{-1} Å^{-2} and C is in kcal mol^{-1} units.

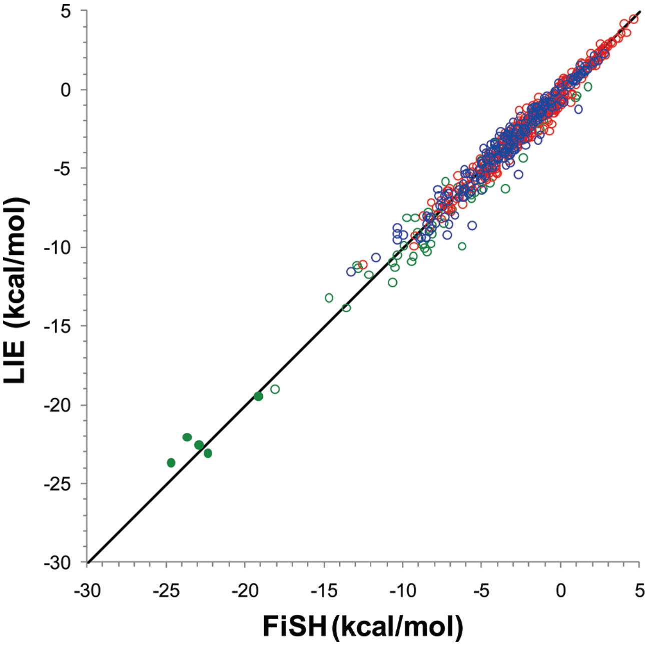


Figure 7. Comparison between hydration free energy predictions with the FiSH model (this study) and with the LIE explicit-solvent model[24] for the training (blue symbols) and testing (red symbols) data sets and for the SAMPL1 (green symbols) data set. Filled circles correspond to 5 sulfoneurea analogs from the SAMPL1 data set. The plotted data correspond to the FiSH model with AM1BCC-SP partial charges, and cavity parameters derived from the training data set. The LIE data are for AM1BCC-SP partial charges and single-conformation solutes and are taken from a separate study.[24] The diagonal line indicates ideal correlation.

have previously employed in the SAMPL1 prospective challenge.[27] An important aspect that differentiates the FiSH model from that earlier model is the reduction of parameters fitted to experimental hydration data in an attempt to improve model transferability. In the FiSH model, only the two cavity parameters require fitting to the experiment, the microscopic surface tension of water, γ_{cav}, and a constant, C. The van der Waals and electrostatic components were calibrated against the corresponding components derived from explicit-solvent simulations using the linear interaction energy (LIE) approach.[11−14,24] The philosophy adopted here is that the transferability of continuum solvation models can be increased by emulating the physics captured by explicit solvation models.

Performance of FiSH Model versus LIE Explicit Model. The primary objective of this study is to develop a continuum model that mimics as closely as possible an explicit solvation model. Performance testing was carried out on the 301 compounds of the testing set and 63 compounds from the SAMPL1 data set. In Figure 7, we plot the hydration free energies predicted with the FiSH continuum model versus those calculated with the LIE explicit-solvent model (based on AM1BCC-SP partial charges and single-conformation representations of the solutes). It is apparent that the continuum model developed here reproduces closely the explicit model al the level of hydration free energies. In quantitative terms, the FiSH continuum model predicts the explicit model data with MUE values of ~0.5 kcal/mol and slightly below 1 kcal/mol for all three data sets, respectively, with correlation slopes and coefficients close to unity (Table 5). There are no major outliers even for the SAMPL1 data set that include more complex, drug-like compounds (Figure 7). We have seen in the previous sections that the excellent agreement carries on to the hydration component terms as well.

Performance of FiSH Continuum Model versus Experimental Data. The absolute performance of the developed FiSH continuum solvation model is tested against the experimental hydration free energies for the testing set and the SAMPL1 drug-like data set. As seen in Figure 8, the FiSH continuum model predictions achieve a fairly good correlation with the experiment. In the case of the testing set, MUE is close to 1 kcal/mol, with a slope close to unity and R^2 above 0.8 (Table 5). We note that the MUE obtained with the FISH model is only 0.1 kcal/mol higher than that obtained with the corresponding LIE model (0.906 kcal/mol).[24] For testing on SAMPL1, MUE is slightly larger than 2 kcal/mol, with a slope and R^2 around 0.6 and of 0.8, respectively. These results are slightly better than those

Table 5. Comparing the Hydration Free Energy Predictions of the FiSH Model with Predictions from the Explicit-Solvent LIE Model and with Experimental Hydration Free Energies[a]

	FiSH versus LIE			FiSH versus experiment		
set	MUE	slope	R^2	MUE	slope	R^2
training	0.524 ± 0.033	0.953 ± 0.017	0.946 ± 0.009	0.985 ± 0.066	0.914 ± 0.042	0.806 ± 0.028
testing	0.469 ± 0.020	0.968 ± 0.010	0.968 ± 0.004	1.075 ± 0.052	0.938 ± 0.030	0.826 ± 0.018
SAMPL1	0.958 ± 0.084	0.930 ± 0.018	0.968 ± 0.011	2.173 ± 0.250	0.599 ± 0.043	0.805 ± 0.056

[a] LIE data are for AM1BCC-SP partial charges and rigid solutes.[24] Errors are in kcal mol^{-1} units.

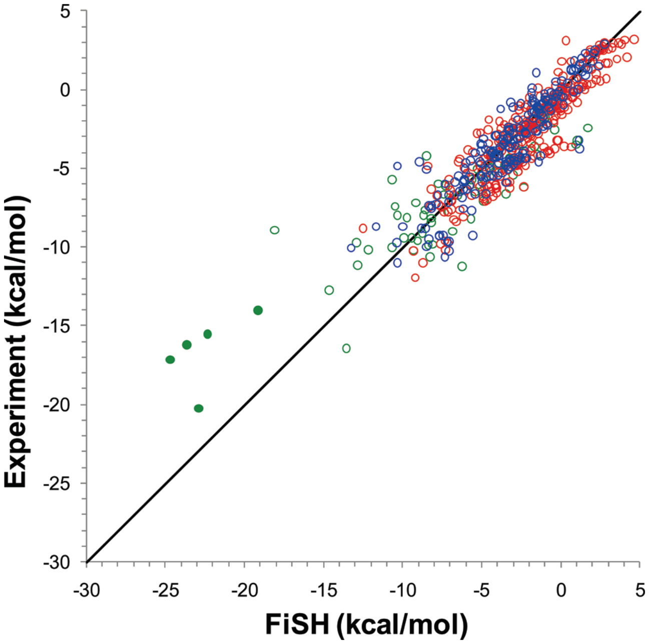


Figure 8. Correlation between hydration free energy predictions with the FiSH model and experimental hydration free energies for the training (blue symbols) and testing (red symbols) data sets, and for the SAMPL1 (green symbols) data set. Filled circles correspond to 5 sulfoneurea analogs from the SAMPL1 data set. The plotted data correspond to the FiSH model with AM1BCC-SP partial charges, and cavity parameters derived from the training data set. The diagonal line indicates ideal correlation.

obtained from the explicit-solvent LIE model with full solute flexibility (MUE of 2.25 kcal/mol), and a little worse than from the LIE model with rigid solute (MUE of 1.92 kcal/mol), for the same partial charge set.[24]

While these data correspond to AM1BCC-SP partial charges, we also calibrated the FiSH model with different partial charge sets, AM1BCC-OPT and RESP. We retrained the cavity component on the training set each time we changed the charge set. The parameters for the cavity term do not vary too much depending on charge (Table S4, Supporting Information). The overall performance of the FiSH models does not depend much on these different charge sets, similar to what was observed with the explicit-solvent LIE models (see Table S5, Supporting Information).[24] Comparable performances in terms of MUEs (training, testing, SAMPL1) were obtained by employing the AM1BCC-SP (0.995, 1.075, 2.173 kcal/mol) or RESP (1.173, 1.068, 2.156 kcal/mol) partial charges in the FiSH models in terms of MUEs. In terms of correlation coefficients and slopes, RESP charges yielded improved slopes (0.793−0.997) compared to those of the AM1BCC-SP charges (0.599−0.938), yet with smaller correlation coefficients (0.660−0.801 for RESP vs 0.805−0.826 for AM1BCC-SP). This decrease in correlation coefficient going from RESP to AM1BCC-SP charges may be partly because the RDF peaks used to define the SAS for the continuum van der Waals term were originally obtained from MD simulations using AM1BCC charges.

Table 6. Listing of Function Groups Used for Error Analysis

functional group	# of members
other	81
alkane	20
alkene	13
alkyne	3
aromatic	18
halogen	57
F	3
Cl	31
Br	12
I	4
OH	27
1° OH	10
2° OH	4
3° OH	2
phenyl OH	11
amine	10
alkyl amine	7
aryl amine	3
carboxylic acid	2
ester	30
amide	2
ether	8
ketone	12
aldehyde	8
nitro	1
cyano	3
hypervalent S	1
thiol	5

Functional Group Analysis of FiSH Continuum Model Predictions. We have separately examined the performance of the derived continuum model on specific chemical classes, on the basis of monofunctional compounds that could be found in the testing set. A majority of functional groups that are commonly encountered in typical drug-like compounds were assessed (see Table 6) in this analysis of FiSH model prediction errors. A similar analysis was previously carried out on the explicit-solvent LIE model of hydration.[24] As seen in Figure 9 (data tabulated in Table S6, Supporting Information), the functional group based error profile of the FiSH continuum model mirrors closely that of the LIE explicit model. The changes in prediction errors between these two models are within 1 kcal/mol for all functional groups investigated. This further emphasizes that the FiSH continuum model succeeded in its primary objective, that is, to mimic a physics-based explicit-solvent hydration model.

In terms of mean-unsigned errors to experimental data, the FiSH continuum model performs well on alkanes (0.394 kcal/mol), alkenes (0.846 kcal/mol), aromatic hydrocarbons (0.691 kcal/mol), chlorinated (0.541 kcal/mol) and iodinated compounds (0.644 kcal/mol), aryl amines (0.410 kcal/mol), esters (0.747 kcal/mol), ethers (0.825 kcal/mol), and ketones (0.500 kcal/mol), with MUE values well below the average for the entire testing set of 1.08 kcal/mol (for AM1BCC-SP partial charges). Interestingly, even though aromatic compounds are predicted well by both the FiSH continuum model and the LIE explicit model when compared to the experiment, the continuum model predictions underestimate LIE predictions by a considerable margin (0.76 kcal/mol). For brominated compounds, neutral carboxylic acids, aldehydes,

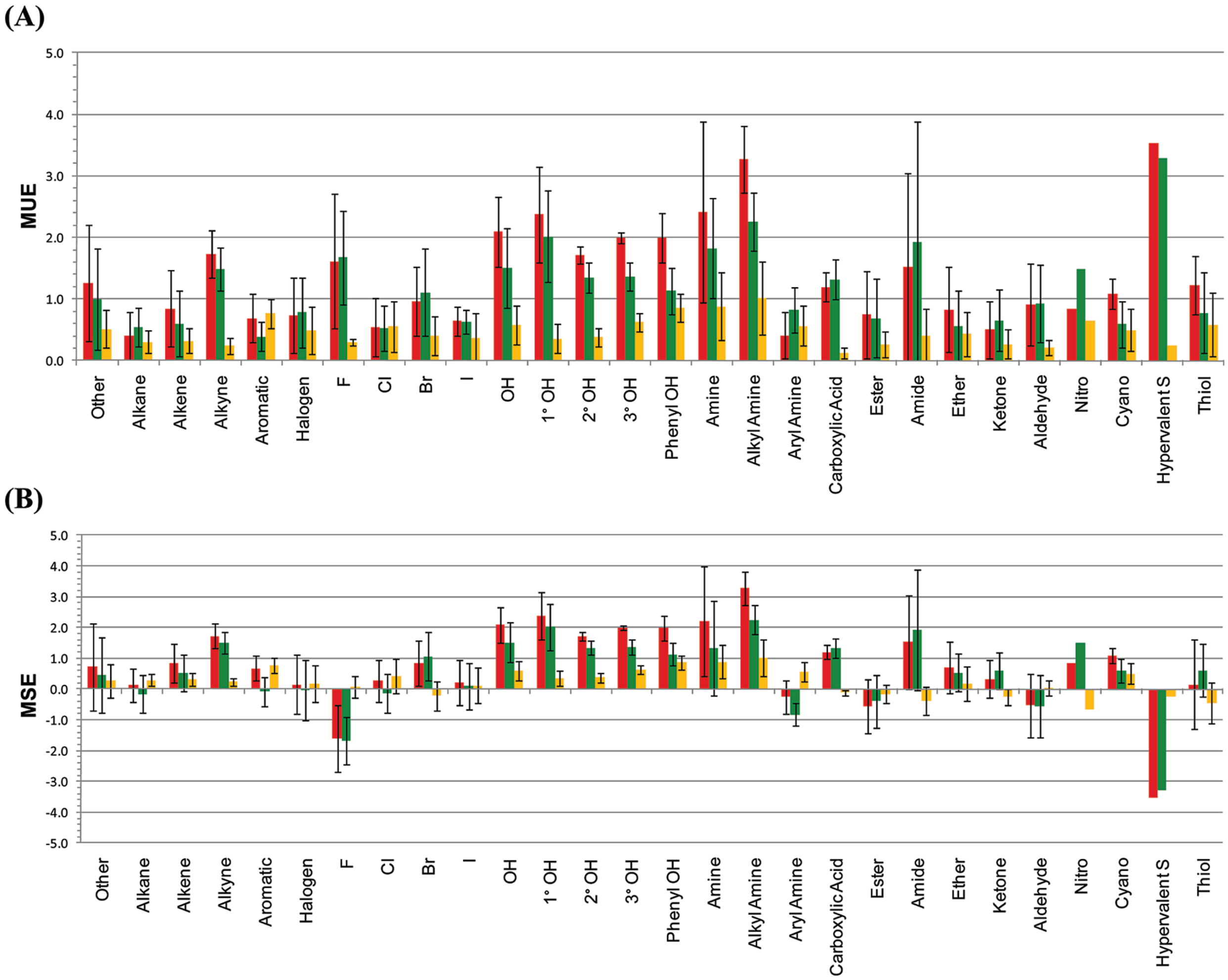


Figure 9. Comparative functional group analysis of prediction errors for the FiSH model and the LIE explicit-solvent model. FiSH model versus experiment (red bars); LIE explicit model versus experiment (green bars); FiSH model versus LIE explicit-solvent model (orange bars). The LIE data are for AM1BCC-SP partial charges and single-conformation solutes and are taken from a separate study.[24] (A) MUE ± SD values. (B) MSE ± SD values.

cyano derivatives, and thiols, the predictions are close to the MUE value of the entire data set, with either the continuum or explicit model having a marginal advantage.

Problematic functional classes for the FiSH continuum model, having MUE values larger than 1.5 kcal/mol, include alkynes (1.730 kcal/mol), fluorinated compounds (1.611 kcal/mol), alcohols (2.090 kcal/mol) and phenols (1.986 kcal/mol), neutral aliphatic amines (3.270 kcal/mol) and amides (1.525 kcal/mol). As seen in Figure 9, these are the same chemical classes that are problematic with the corresponding explicit-solvent LIE model. Similar mean signed errors (FiSH, LIE) were obtained with the two models in the case of alkynes (1.730, 1.491 kcal/mol), fluorinated compounds (−1.611, −1.675 kcal/mol), and amides (1.525, 1.920 kcal/mol). Also, similarly to what was observed with the LIE method, the hydration free energy predictions for alkynes can be significantly improved by employing a FiSH continuum model based on RESP charges (MUE reduced from 1.73 to 0.61 kcal/mol, see Table S6, Supporting Information). Fluorinated compounds were among the few functional classes that were overestimated (Figure 9B). In our FiSH continuum model, the Born radius for the F atom is about 1.72 Å, which is typically less than often used, but which we find is appropriate to mimic well the explicit solvent based data. Hence, nonempirical improvements in the hydration free energy prediction of fluorinated compounds (e.g., not simply based on ad-hoc adjustment of F radius) have to be sought in force field modifications like Lennard-Jones potential parameters or atomic partial charges. Indeed, the FiSH continuum model based on RESP charges does provide some relief in the case of fluorinated compounds, with MUE being reduced from 1.61 to 1.04 kcal/mol. For some chemical classes, particularly the alcohols, phenols, and aliphatic amines, the FiSH continuum model performs poorly (MUEs of 1.72 to 3.27 kcal/mol), and the underestimation of experimental data is accentuated (by 0.3 to 0.9 kcal/mol) with the FiSH continuum model compared to that with the LIE explicit solvent model. In these cases, again, significant improvements can be obtained by employing RESP charges (Table S6). Unfortunately, this does not extend generally to other functional classes, as RESP partial charges degrade the overall predictions obtained with AM1BCC partial charges for both the FiSH and LIE models (Table S6).

Comparison of FiSH Model with Other Continuum Models. The FiSH model is compared with the CED and RF models in Table 7 on the testing set and the SAMPL1 data set. On the training and testing sets, there was little

Table 7. Comparing the Hydration Free Energy Predictions of the FiSH Model with Those from Previously Developed Continuum Solvation Models, Continuum Electrostatics-Dispersion (CED) Solvation Model and Reaction Field (RF) Electrostatics-Only Model[a]

	FiSH		
set	MUE	slope	R^2
training	0.985 ± 0.066	0.914 ± 0.042	0.806 ± 0.028
testing	1.075 ± 0.052	0.938 ± 0.030	0.826 ± 0.018
SAMPL1	2.173 ± 0.250	0.599 ± 0.043	0.805 ± 0.056

	CED[a]		
set	MUE	slope	R^2
training	0.762 ± 0.063	0.881 ± 0.040	0.877 ± 0.034
testing	0.874 ± 0.046	0.872 ± 0.044	0.879 ± 0.023
SAMPL1	2.729 ± 0.331	0.542 ± 0.041	0.818 ± 0.051

	RF[b]		
set	MUE	slope	R^2
training	1.140 ± 0.064	1.150 ± 0.054	0.789 ± 0.035
testing	1.186 ± 0.048	1.309 ± 0.041	0.848 ± 0.017
SAMPL1	1.631 ± 0.188	0.751 ± 0.059	0.780 ± 0.064

[a] CED model at $D_{\mathrm{in}} = 1$, $\rho = 0.9$, and 25-atom-type c-vdW parameters, as described previously.[21] [b] RF model at $D_{\mathrm{in}} = 1$, $\rho = 1.1$, as described previously.[21] [a] All iodinated molecules have been removed. AM1BCC-SP charges used throughout. Errors are in kcal mol^{-1} units.

variation in the performance of these models in terms of MUE (training, testing), with FiSH model predictions (0.995, 1.075 kcal/mol) marginally outperformed by those of the CED model (0.762, 0.874 kcal/mol). We note that the FiSH model had the correlation slope closest to 1 among the three models. Functional group analysis of mean-unsigned-errors (FiSH vs CED) shows that the difficult functional groups for the FiSH continuum model (alkynes, 1.730 vs 1.127 kcal/mol; fluorinated compounds, 1.611 vs 0.617 kcal/mol; alcohols, 2.090 vs 0.680 kcal/mol; phenols, 1.986 vs 1.083 kcal/mol; and neutral aliphatic amines, 3.270 vs 1.397 kcal/mol) receive better predictions with the CED model (Figure 10, Table S6). The CED model also improves the predictions of brominated compounds (0.985 vs 0.436 kcal/mol), neutral carboxylic acids (1.197 vs 0.140 kcal/mol), and thiols (1.229 vs 0.482 kcal/mol), whereas FiSH continuum model predictions were better on chlorinated compounds (0.541 vs 0.911 kcal/mol), aryl-amines (0.410 vs 1.183 kcal/mol), and cyano derivatives (1.087 vs 1.603 kcal/mol). The slightly better performance of the CED solvation model is understandable since it has many more parameters and was fitted on experimental data for monofunctional compounds from the traning and testing data sets, whereas the FiSH continuum model was trained on LIE data from explicit solvent simulations and therefore mimics LIE's shortcomings.

The CED solvation model, however, has serious transferability problems for the SAMPL1 data set, noted previously,[21] and highlighted in Table 7, that are outside its applicability domain. The predictions on the SAMPL1 data set are improved with the FiSH model (2.173 kcal/mol) relative to the CED solvation model (2.729 kcal/mol), with a MUE decrease of 0.55 kcal/mol and a slightly larger correlation slope (0.599 vs 0.542) to experimental data. In terms of transferability, the increase in MUE from testing set to SAMPL1 data set is 1.1 kcal/mol in the case of the FiSH model, and 1.9 kcal/mol in the case of the CED solvation model. Hence, the FiSH continuum model is more transferable. This supports the hypothesis that the transferability of continuum solvation models can be increased by fitting to physics-based explicit solvation models rather than directly to experimental data. Although performing worst on traditional compounds, the simple RF continuum solvation model outperforms the complex FiSH continuum model on the SAMPL1 data set by a further decrease of 0.55 kcal/mol in MUE and an increase in correlation slope, with a good transferability reflected by only a 0.5 kcal/mol decrease in MUE from the testing set to the SAMPL1 data set. However, the RF continuum model is not practical because it fails on all types of hydrocarbons and primarily on alkanes, noting that hydrocarbon moieties are ubiquitous in organic molecules. Alkanes, alkenes, aromatics, and alkynes all are overestimated (MSE of 1.0 to 2.7 kcal/mol) with the RF model, while large errors are also obtained for fluorinated and iodinated compounds, and for amides (Figure 10, Table S6).

Conclusions

In this paper, we propose a novel continuum solvation model, the First-Shell Hydration (FiSH) model, as an attempt to capture the physics of an explicit solvation model by focusing on the first shell of water around the solute while maintaining the speed provided by the continuum approach. The FiSH continuum model consists of an electrostatic, van der Waals, and cavity contribution to solvation, with only the latter fitted to experimental data. Changes have been introduced to the definition of the continuum electrostatic and van der Waals components, which have been calibrated against explicit-solvent MD simulations *via* the linear interaction energy (LIE) method. The central premise of this study is that the transferability of the continuum model can be increased by reducing the number of parameters fitted directly to the experiment, and by emulating the physics captured by an explicit solvation model. A continuum model designed to mimic an explicit solvent force field model will inherit the transferability and generality of the force field model, for better or for worse.

To capture first hydration shell effects with the FiSH model, we first incorporated charge asymmetry[20,34−37] into the continuum electrostatics model. This was achieved through a modification of our earlier approach of defining the Born radii of atoms as a function of the ISCD.[23] Multiple functional forms were explored and trained on explicit solvent simulations. A nonlinear function with four parameters yielded optimal correlations to explicit water simulations and gave drastic improvements over the initial continuum electrostatic model. A hybrid continuum van der Waals model introduced in this paper creates a first shell of solvent restricted to and distributed uniformly over the SAS. A second region, starting one solvent diameter away from the SAS and extending to infinity, is treated

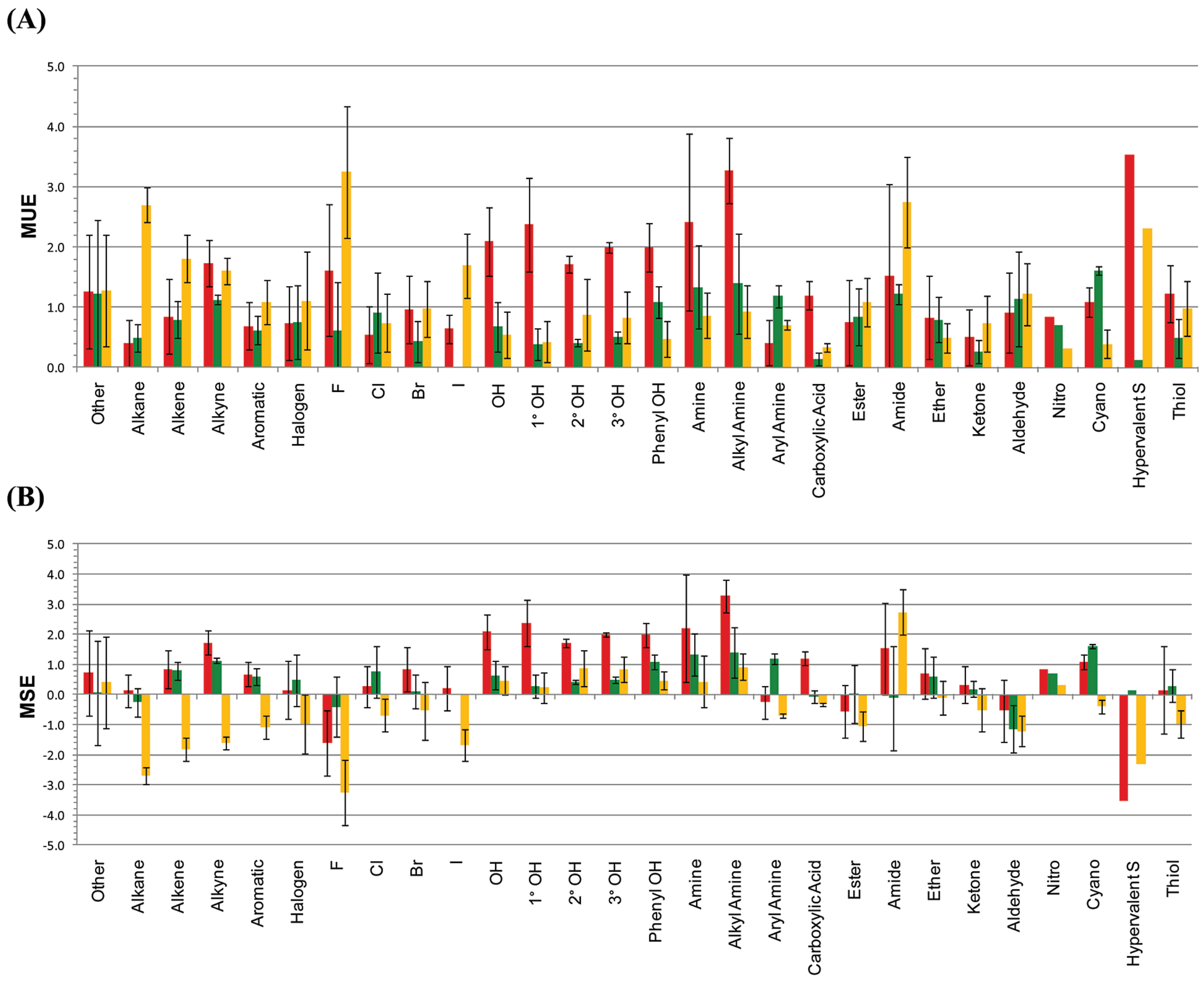


Figure 10. Comparative functional group analysis of prediction errors for the FiSH model, the continuum electrostatics-dispersion (CED) solvation model, and the reaction field (RF) electrostatics-only model. FiSH model versus experiment (red bars); CED model versus experiment (green bars); RF model versus experiment (orange bars). The CED model was employed here with $D_{in} = 1$, $\rho = 0.9$, and 25-atom-type c-vdW parameters, as described previosuly.[21] The RF model was employed here with $D_{in} = 1$, $\rho = 1.1$, as described previously.[21] (A) MUE ± SD values. (B) MSE ± SD values.

as a uniform continuum. This model does not require the large number of fitted parameters used in the previous CED model,[21] instead relying entirely on force field parameters for the Lennard-Jones potentials. Testing of the FiSH van der Waals continuum model against the explicit-solvent van der Waals data showed an excellent performance on simple compounds and moderate performance on more complex, drug-like molecules.

The primary objective of the FiSH continuum model, to mimic the hydration free energies from an explicit-solvent model, has been achieved. It predicts the explicit-solvent LIE data with MUEs of about 0.5 kcal/mol for the training and testing data sets and slightly below 1 kcal/mol for the drug-like SAMPL1 data set, with correlation slopes and coefficients close to unity for all three data sets. The excellent agreement carries on to the hydration component terms, as well as to various chemical functional groups commonly present in small organic molecules. The absolute performance against experimental data obtained with the FiSH continuum model is as good as that afforded by the explicit-solvent LIE model, i.e., MUEs of about 1 kcal/mol for the training and testing sets and slightly above 2 kcal/mol for the SAMPL1 data set. Another similarity to the explicit-solvent model is the weak dependence of the overall performance of the FiSH continuum model on the tested partial charge sets. There is, however, an uneven impact of the charging method across functional classes, with RESP charges providing better prediction than AM1BCC charges on certain chemical classes that are poorly predicted (e.g., alkynes, fluorinated compounds, alcohols, phenols, and aliphatic amines), but worse predictions on others.

Another objective that has been achieved with the FiSH continuum model is the improvement of transferability relative to previously developed CED solvation model that has been (over)fitted against experimental data. Comparatively, the transferability of the FiSH continuum model is improved by about 0.8 kcal/mol between simple compounds from the training and testing data sets over the more complex molecules found in the SAMPL1 data set when compared to the CED solvation model. On the basis of a very acceptable performance, the sound physical foundation of the FiSH continuum model is an important attribute that should not be overlooked when compared to other models in terms of global fitness measures.

Acknowledgment. This is National Research Council of Canada publication number 50689.

Supporting Information Available: Composition of the hydration data sets with experimental transfer free energies (Table S1). Optimized parameters for Born radii correction functions (Table S2). LIE data for spherical solutes (Table S3). Cavity parameters calibrated on the training subset for various FiSH models (Table S4). Errors for various charging methods with FiSH continuum model (Table S5). Raw data for Figures 9 and 10 (Table S6). Plots of Born radii correction functions trained on spheres or bracelets (Figure S1). Distribution of atomic ISCD within the training, testing, and SAMPL1 data sets (Figure S2). Correlation plots for the van der Waals and total nonpolar components of the FiSH continuum model against the solute molecular surface area (Figure S3). This material is available free of charge via Internet at http://pubs.acs.org.

References

(1) Gilson, M. K.; Zhou, H. X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

(2) McInnes, C. *Curr. Opin. Chem. Biol.* **2007**, *11*, 494–502.

(3) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199–203.

(4) Kang, Y. K.; Némethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1987**, *91*, 4109–4117.

(5) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978–1988.

(6) Chambers, C. C.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 16385–16398.

(7) Marten, B.; Kim, K.; Cortis, C.; Friesner, R. A.; Murphy, R. B.; Ringnalda, M. N.; Sitkoff, D.; Honig, B. *J. Phys. Chem.* **1996**, *100*, 11775–11788.

(8) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517–529.

(9) Tan, C.; Yang, L.; Luo, R. *J. Phys. Chem. B* **2006**, *110*, 18680–18687.

(10) Reddy, M. R.; Erion, M. D. *Free Energy Calculations in Rational Drug Design*; Springer-Verlag: New York, 2001.

(11) Lee, F. S.; Chu, Z. T.; Bolger, M. B.; Warshel, A. *Protein Eng.* **1992**, *5*, 215–228.

(12) Aqvist, J.; Medina, C.; Samuelsson, J. E. *Protein Eng.* **1994**, *7*, 385–391.

(13) Carlson, H. A.; Jorgensen, W. L. *J. Phys. Chem.* **1995**, *99*, 10667–10673.

(14) Su, Y.; Gallicchio, E.; Das, K.; Arnold, E.; Levy, R. M. *J. Chem. Theory Comput.* **2006**, *3*, 256–277.

(15) Chen, J.; Brooks, C. L., III; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140–148.

(16) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.

(17) Simonson, T. *Curr. Opin. Struct. Biol.* **2001**, *11*, 243–252.

(18) Baker, N. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137–143.

(19) Raschke, T. M.; Levitt, M. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6777–6782.

(20) Mobley, D. L.; Barber, A. E.; Fennell, C. J.; Dill, K. A. *J. Phys. Chem. B* **2008**, *112*, 2405–2414.

(21) Sulea, T.; Wanapun, D.; Dennis, S.; Purisima, E. O. *J. Phys. Chem. B* **2009**, *113*, 4511–4520.

(22) Nicholls, A.; Wlodek, S.; Grant, J. A. *J. Phys. Chem. B* **2009**, *113*, 4521–4532.

(23) Purisima, E. O.; Sulea, T. *J. Phys. Chem. B* **2009**, *113*, 8206–8209.

(24) Sulea, T.; Corbeil, C. R.; Purisima, E. O. *J. Chem. Theory Comput.* **2009**, DOI: 10.1021/ct9006025.

(25) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.

(26) Mobley, D. L.; Dill, K. A.; Chodera, J. D. *J. Phys. Chem. B* **2008**, *112*, 938–946.

(27) Guthrie, J. P. *J. Phys. Chem. B* **2009**, *113*, 4501–4507.

(28) Purisima, E. O.; Nilar, S. H. *J. Comput. Chem.* **1995**, *16*, 681–689.

(29) Purisima, E. O. *J. Comput. Chem.* **1998**, *19*, 1494–1504.

(30) Chan, S. L.; Purisima, E. O. *J. Comput. Chem.* **1998**, 1268–1277.

(31) Bhat, S.; Purisima, E. O. *Proteins* **2006**, *62*, 244–261.

(32) Floris, F.; Tomasi, J. *J. Comput. Chem.* **1989**, *10*, 616–627.

(33) Floris, F. M.; Tomasi, J.; Pascual-Ahuir, J. L. *J. Comput. Chem.* **1991**, *12*, 784–791.

(34) Latimer, W. M.; Pitzer, K. S.; Slansky, C. M. *J. Chem. Phys.* **1939**, *7*, 108–111.

(35) Rashin, A. A.; Honig, B. *J. Phys. Chem.* **1985**, *89*, 5588–5593.

(36) Roux, B.; Yu, H. A.; Karplus, M. *J. Phys. Chem.* **1990**, *94*, 4683–4688.

(37) Babu, C. S.; Lim, C. *J. Phys. Chem. B* **1999**, *103*, 7958–7968.

(38) Chorny, I.; Dill, K. A.; Jacobson, M. P. *J. Phys. Chem. B* **2005**, *109*, 24056–24060.

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

(40) Born, M. *Z. Phys.* **1920**, *1*, 45–48.

(41) Cerutti, D. S.; Baker, N. A.; McCammon, J. A. *J. Chem. Phys.* **2007**, *127*, 155101–155112.

(42) Tan, C.; Tan, Y. H.; Luo, R. *J. Phys. Chem. B* **2007**, *111*, 12263–12274.

(43) Orozco, M.; Luque, F. J. *Chem. Phys. Lett.* **1997**, *265*, 473–480.

(44) Westergren, J.; Lindfors, L.; Höglund, T.; Lüder, K.; Nordholm, S.; Kjellander, R. *J. Phys. Chem. B* **2007**, *111*, 1872–1882.

(45) Almlof, M.; Carlsson, J.; Aqvist, J. *J. Chem. Theory Comput.* **2007**, *3*, 2162–2175.

(46) 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.

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

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

(49) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(50) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.

(51) *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2005.

(52) Shivakumar, D.; Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 919–930.

CT9006037

JCTC Journal of Chemical Theory and Computation

Mechanical Properties of Coarse-Grained Bilayers Formed by Cardiolipin and Zwitterionic Lipids

Martin Dahlberg* and Arnold Maliniak

Division of Physical Chemistry, Arrhenius Laboratory, Stockholm University, S-106 91 Stockholm, Sweden

Received December 4, 2009

Abstract: Lipid shape and charge are connected with the physical properties and the biological function of membranes. Cardiolipin, a double phospholipid with four chains and the potential of changing its charge with pH, is crucially connected with mitochondrial inner membrane shape, and recent experiments suggest that local pH changes allow highly curved local geometries. Here, we use a coarse-grained molecular dynamics model to investigate the mechanical properties of cardiolipin bilayers, systematically varying the headgroup charge and the composition in mixtures with zwitterionic 1,2-dioleoyl-glycero-3-phosphatidylcholine (DOPC) or 1,2-dioleoyl-glycero-3-phosphatidylethanolamine (DOPE). Low cardiolipin charge, corresponding to low pH, was found to induce bending moduli on the order of k_BT and curved microdomains. On the length scale investigated, in contrast to continuum theoretical models, we found the area modulus and bending modulus to be inversely correlated for mixtures of cardiolipin and DOPC/DOPE, explainable by changes in the effective headgroup volume.

Introduction

The physical and mechanical properties of lipid bilayer membranes are central for understanding their shapes and functions.[1–3] The strong connection between biological function and membrane properties is highlighted in the mitochondrion, which plays an important role in energy production in eukaryotic cells and has the ability to drastically change its morphology.[4–7] The understanding of the mitochondrial architecture has been expanded recently due to detailed electron microscopy images, and the highly convoluted inner mitochondrial membrane (IMM) is now believed to be composed of distinct, but dynamic, regions. The emerging picture is that proteins and lipids dynamically optimize the mitochondrial topology to adjust performance.[8] It is known that proteins are involved in the organization of mitochondrial and cristae structure[5,9] and that lipid type and composition affect membrane proteins.[10] Recent experiments on model lipid vesicles with a composition close to that of the IMM show that small amounts of locally applied acid can give deformations of the membrane in the shape of tubes with an approximate radius of 40 nm, highly reminiscent of the native mitochondrial tubules and cristae junctions.[11] A deeper understanding for the lipid components of mitochondria is needed to explain the connection between topology and function, both of which are affected by the lipid composition.[12] We thus focus on the lipid components of the IMM. The major IMM lipid constituents are the zwitterionic phosphatidylcholines (PC) and phosphatidylethanolamines (PE) and the negatively charged cardiolipins (CL).[13] The membrane composition varies with species and cell type, but in eukaryotes, the typical ratios of PC:PE:CL are 2:2:1−6:3:1. In the eukaryotes, CLs are specific to mitochondria and typically show a distinct saturation/length pattern in their four acyl chains.[14] The CL headgroup is negatively charged at physiological pH, but different experiments have shown either a net −1 or −2 charge,[15,16] labeled below as CL-1 and CL-2, respectively. The high second pK_a (7.5−9.5) needed to explain the −1 charge at neutral pH has been attributed to intramolecular hydrogen bonding in the CL headgroup. It has also been hypothesized that the ability to trap and conduct protons is important to the proton transport in mitochondrial adenosine triphosphate (ATP) production. Additionally, low pH has been shown to induce negative curvature in CL aggregates (such as the inverse

* Corresponding author e-mail: martind@physc.su.se.

10.1021/ct900654e © 2010 American Chemical Society
Published on Web 04/08/2010

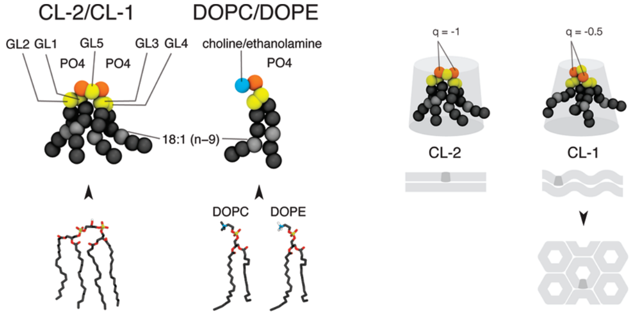


Figure 1. CG models for CL and DOPC/DOPE with 18:1 oleoyl chains. Phosphate charges in CL-2 and CL-1 are −1/−1 and −0.5/−0.5, respectively. The effective geometry of the lipids determines phase behavior. On the right, cylindrical CL-2 (bilayers) and inverse cone-shaped CL-1 (undulating bilayers and the inverse hexagonal phase, H_{II}).

hexagonal phase), and charge neutralization is strongly connected to the propensity to form nonlamellar aggregates with negative curvature.[17] This behavior, and the experimental evidence for tubule formation upon local pH reduction, is not understood mechanistically, and the link between lipid charge and membrane mechanical properties deserves more attention.

Here we present coarse-grained (CG) molecular dynamics simulations of the CL−DOPC and CL−DOPE systems (Figure 1), based on the interaction model proposed by Marrink et al.[18] Our aims are to construct a minimal IMM representation and to examine its mechanical properties. By adding DOPC or DOPE to CL bilayers, we study how the headgroup size and the bilayer charge affect the bending and area modulus of the bilayers. The bending modulus is especially important for understanding the folds of the inner mitochondrial membrane. To further investigate the properties of CL bilayers under stress, we calculate the critical tension for porated bilayers and show how the bilayer line tension is dependent on the composition. We discuss previously reported experiments that suggest the critical tension (or lysis tension) is lowered by introducing small amounts of CL.[19] In a recently developed CG model of CL, the headgroup charge was shown to strongly influence the phase behavior of the lipid aggregates and was shown to reproduce the CL phase polymorphism upon changes in the number of acyl chains and the headgroup charge.[20] Reducing the phosphate charge per molecule from −2 to −1 (in that study scaled down to −1.4 and −0.7) led to adoption of the inverse hexagonal phase (H_{II}), in agreement with experimental observations where the lamellar, to inverse hexagonal phase transition, occurred at pH 2.8.[21] The total bilayer charge was varied in two ways: by changing the concentration of the charged lipid and by changing the protonation state of the charged lipid. Because we chose the acyl chains of all lipids to be identical, we avoided the problem of confounding changes in properties originating from different headgroup characteristics and from those due to chain composition. Generally, lipids with small headgroups and/or large chain volume will form aggregates with negative spontaneous curvature, i.e. inverted phases. The CG models of DOPC and DOPE differ in the interaction parameters of one particle in the headgroup. Effectively, this gives two different headgroup volumes, which affects the spontaneous curvature (see Figure 1). With these two lipids, and the charged CL, we can systematically vary the spontaneous curvature and the effective membrane charge. It is unclear at what composition the lamellar to H_{II} phase transition ought to occur in mixtures of reduced charge CL (−1) and DOPC, but it has been observed experimentally that PC lipids stabilize CL in the lamellar phase.[22] The H_{II} phase is not expected in systems composed of DOPC and CL with full (−2) charges because both components alone form lamellar phases. Aggregates formed by DOPE and CL with reduced (−1) charge have negative spontaneous curvatures and under certain conditions show preference for inverted phases, due to relatively small headgroups. We emphasize that the lamellar bilayer is not the equilibrium aggregate geometry for such lipids, but that for our initial conditions, i.e. intact bilayers, the barrier for phase transition is high. It is, however, instructive to examine the properties of membranes under increasing amounts of frustration due to increased (negative) spontaneous curvature of the monolayers, especially because these changes can be triggered locally with pH changes. The connection between curvature and composition is important in bacterial membranes, where it has also been observed that PE and CL are involved in forming microdomains and that these domains are coupled with the membrane curvature.[23] Interestingly, the syntheses of the two lipids are regulated together,[24] and the lipids can replace some of each others' cellular functions.[25,26]

Based on continuum models for membranes,[27,28] the effects of charge on membrane properties have been described in various limits of electrolyte composition and surface geometry.[29−31] Overall, the electrostatic interactions are suggested to increase membrane rigidity. For mixtures with the possibility of segregating, unstable solutions are found, and the rigidity can instead be lowered as segregation occurs. Experimentally, determining the mechanical properties of mixtures of charged and uncharged lipids have been difficult, showing little or no effect of charged lipids,[32] but recently, Rowat et al. described increases in the electrostatic bending rigidity on the order of 3−5 k_BT for ionic surfactants added to DMPC vesicles.[33]

Recent molecular dynamics (MD) modeling of CL bilayers has shown that CL tends to increase membrane order and decrease lipid mobility,[34−36] which is similar to the behavior found in simulations of other charged lipids.[37] The lipid compositions in the previous MD simulations have been at pure CL or near-physiological CL concentrations, and to our knowledge there are no reports on CL membrane properties over a broad composition range. Atomistic MD simulations give highly detailed information about the specific interactions that occur in these lipid mixtures but are relatively expensive for systematic exploration of mechanical properties. CG approaches have shown much progress in recent years and reproduce many of the important mechanical and structural properties of lipid membranes.[38]

Methods

Model. The CG model used in the previous work[20] (Figure 1) was modified to fit the updated MARTINI07[18] force field:

(i) an increased headgroup charge, to −1.0 per phosphate group, in line with the increased ion charge in MARTINI07. Note that the dielectric constant was decreased from 20 to 15; (ii) the particle type of the unsaturated part of the acyl chain was changed from C1 to C3; and (iii) the GL1−GL2 bond distance was reduced from 0.47 to 0.37 nm.

To better fit the CL geometries from atomistic simulations,[34] the headgroup potentials were optimized (details are given in the Supporting Information). For the CL-1 (−1 total charge) system, the phosphate charges were reduced from −1 to −0.5 each, to model proton equilibration between the two phosphate groups, which was assumed to be rapid compared to the simulation time scale based on recent density functional theory (DFT) calculations of the CL headgroup done in our group.[39] These calculations show that proton exchange between the two phosphate groups is possible on the nanosecond time scale. In the CG model, the reduced charge induced an intramolecular P−P distance shift from 0.60 to 0.58 nm. As a test of the charge partition model, we also studied the −1/0 case in a 1:1 mixture of DOPC and CL.

Models for DOPC, DOPE, water, “antifreeze”, and sodium counterions from MARTINI07 were used. For the line tension simulations, antifreeze particles were necessary to avoid crystallization of the solvent, and approximately 10% of the water particles were replaced by antifreeze. Electrostatic and Lennard-Jones interactions were cutoff at 1.2 nm, with electrostatics shifted from 0 to the cutoff and Lennard-Jones interactions shifted from 0.9 nm to the cutoff, as described in MARTINI07.[18] Additionally, a 2.0 nm electrostatic cutoff was tested for the DOPC and CL-2 systems.

Computational Details. Simulations were run employing GROMACS 4,[40] with a 20 fs time step. The time scale was multiplied by a factor of four after simulations to match the experimental diffusion coefficient of water. The temperature was kept at 310 K with a Berendsen thermostat with coupling constant 1 ps. Pressure coupling was carried out using Berendsen barostats set to 1 bar and coupling constant 4 ps. A Nosé-Hoover thermostat (1 ps) and Parrinello-Rahman barostats (semi-isotropic at 10 ps) were also used to test the effect of noncanonical sampling in the Berendsen weak-coupling simulations.

Different compressibilities were used to enable calculation of the mechanical properties. For the line tension system, we used semi-isotropic pressure coupling with zero compressibility in the bilayer slab direction and 5×10^{-5} bar^{-1} in the other directions. In the area compression modulus simulations, compressibilities were set to 5×10^{-5} bar^{-1}, and surface tensions 5, 10, and 15 mN/m were applied in the bilayer plane (NPγT). Critical tension simulations were run similarly but with higher surface tensions (ranging from 30 to 60 mN/m). For all other simulations, semi-isotropic barostats with compressibility 5×10^{-5} bar^{-1} was used. For the pressure profile simulations, we used the local pressure calculations with an Irving−Kirkwood contour, as implemented by Lindahl et al.,[41] in GROMACS 3.0.2 with local pressure extensions. Total simulation times were dependent on which physical property was investigated: (i) 8 μs for the area per lipid and bending modulus; (ii) 4.8 μs for line tensions; (iii) 1.2 μs for area moduli with applied surface tensions; (iv) variable between 10 ns and 1.6 μs depending on the point of collapse for the critical tension simulations; and (v) 300 ns for pressure profiles. Simulation times were chosen to allow proper averaging of the mechanical properties studied.

Systems. Bilayers with 376 CL molecules (188 in each leaflet), based on previous work,[20] were used as a starting point. Pairs of DOPC molecules were generated from CL by splitting the position of the GL5 particle, moving the new choline particles along the GL5−PO_4 bond vectors, and from energy minimization. The mole fractions (X_{CL}) of CL generated were 0, 0.10, 0.25, 0.33, 0.50, 0.67, 0.75, and 1. The substitution was symmetric with respect to the bilayer leaflets. Bilayers with DOPE were obtained by changing the identity of the choline particles in DOPC molecules. No other changes were necessary. Water content was approximately 47 water molecules per lipid, counting each water particle as four water molecules and each CL as two lipids. This is approximately 50% higher than the water content needed to saturate zwitterionic bilayers and corresponds to fully hydrated CL bilayers.[42] We additionally simulated the $X_{CL} = 1$ system with approximately 147 waters per lipid to assess the effect of hydration.

Bilayer pores for critical tension simulations were generated by introducing a new interaction site with strong repulsion only affecting the acyl chain particles. The Lennard-Jones parameter C12 from the interaction level “IX” (0.02581 kJ mol^{-1} nm)[12] was scaled up to give a pore of sufficient radius. A column of 13 such particles with a spacing of 0.7 nm was introduced across the bilayers. The bilayers were then energy minimized in two steps, first by scaling the repulsion parameter with a factor 10 and then with a factor 1000. The positions of the column particles were not updated, which left them stationary throughout the simulations. Line tension simulations were done with slabs of bilayers (376 CL through 752 DOPC) surrounded by water boxes in the *y*- and *z*-directions.[43]

Results

Lipid Segregation. CL charge was found to greatly influence the properties of the bilayers. This was expected because of the propensities to adopt different phases at equilibrium, but we found that even small amounts of CL produced significant qualitative differences between the CL-1 and CL-2 systems. For CL-1, we observed partial lipid segregation, which did not emerge for any concentration of CL-2. Qualitatively the domains can be seen in Figure 2B, where the tighter packing of CL-1 headgroups tended to expose more hydrocarbon chains to water. Two quantitative measures of lipid segregation were investigated: the difference number density, $\Delta\rho$, and the phosphate−phosphate radial distribution functions (rdf) in the bilayer plane. We constructed $\Delta\rho$ by binning phosphate particles of the two leaflets separately onto grids and by taking differences in the number densities along the bilayer normal. The mean absolute $\Delta\rho$, see Figure 2C, was significantly higher in the CL-1 systems for all concentrations except 10% CL. Grid

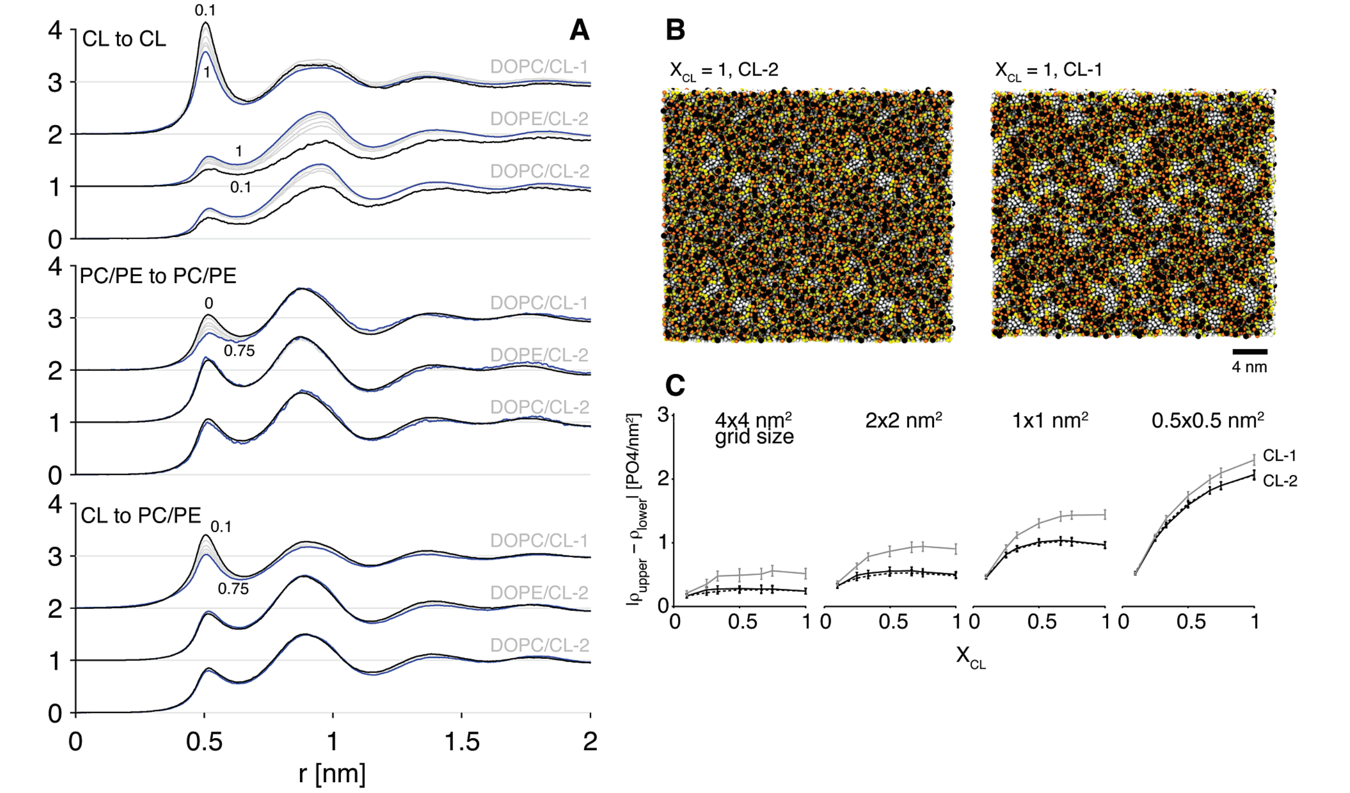


Figure 2. (A) Radial distribution functions two-dimensional (2D) of PO_4 particles for all compositions. Lowest X_{CL} is in black, highest in blue, intermediate in gray, and 10 point moving average; PC/PE represent DOPC or DOPE. (B) Snapshots of microdomain formation in X_{CL} = 1, CL-2 left, CL-1 right, and scale bar 4 nm. Hydrocarbon chains in white, and center glycerol (GL5) in black. (C) $\Delta\rho$ for decreasing grid size (4 × 4, 2 × 2, 1 × 1, and 0.5 × 0.5 nm^2). DOPC/CL-2 (black), DOPE/CL-2 (dotted black), DOPC/CL-1 (gray). Error bars are standard deviations.

sizes between approximately 0.5 × 0.5 to 4 × 4 nm^2 were tried, and the largest difference between $\Delta\rho_{CL\text{-}1}$ and $\Delta\rho_{CL\text{-}2}$ was found for grids of size 1 × 1 to 2 × 2 nm^2, which gives a semiquantitative view of the size of the domains. Radial distribution functions (Figure 2A) showed that the main effect of lowering CL charge was not primarily to induce segregation of CL-1 and DOPC into separate clusters but rather to bind either lipid more closely to CL-1. Such effects were larger for CL itself, resulting in a dramatic increase in the CL−CL first peak (Figure 2A) and a slightly increased first peak in the DOPC−CL and DOPE−CL systems. The increase in the first peak was compensated by a decrease in the second peak of the CL−DOPC/DOPE rdf.

Association of several CL-1 molecules into dynamic domains induced locally concave surfaces, consistent with negative spontaneous curvature, and was strong enough to expose more of the hydrocarbon chains to the aqueous phase (see Figure 2B). Because the bilayers were constructed with identical composition in the two leaflets, local negative curvature in one leaflet was limited by the hydrocarbon−water surface tension of the corresponding positive curvature in the opposite leaflet. To better understand these changes in the structure of the bilayers, we investigated the main properties that are commonly used to describe mechanical properties in bilayers: the area per lipid, the area compression and bending moduli, the line tension, and the critical tension.

Area Per Lipid. The area per lipid was calculated from the box area divided by the number of two-chained lipid equivalents per leaflet and is shown in Figure 3F. Pure CL had an area of 0.631 ± 0.003 nm^2, which was lower than the area of DOPC (0.682 ± 0.002 nm^2) and slightly lower than DOPE (0.647 ± 0.002 nm^2). It should be noted that, experimentally, the DOPE area per lipid at 310 K is not known, because DOPE is not stable in the lamellar phase at that temperature. At 271 K, where DOPE is in the lamellar phase, the area was 0.65 nm^2,[44] which compares favorably with the simulated value, but previous CG simulations with the first generation of the model used here, but at 273 K, showed a lower area of 0.61 nm^2.[45] The CL area was slightly lower with the updated CL model (0.631 vs 0.643 nm^2), which reflects the updated headgroup bond lengths and angles as well as the changed dielectric constant and the use of full ion charges. The DOPC area was in reasonable agreement with recent experiments (0.669 at 303 K)[46] and comparable to the CG model at 300 K (0.67 nm^2).[45]

The reduced charge model, CL-1, had a lower area per lipid, 0.596 ± 0.005 nm^2, consistent with reduced intra- and intermolecular electrostatic repulsion. It should be noted that the undulations for pure CL-1 were very strong and that undulations tend to reduce the area per lipid defined as the area projected onto the *xy* plane. The −1/0 charged CL model at X_{CL} = 0.5 gave an average area per lipid in close agreement (0.616 ± 0.006 nm^2) with the −0.5/−0.5 model (0.610 ± 0.007 nm^2), showing that the effects are not specific to the choice of charge partition.

For all three investigated systems, the area per lipid dependence of the composition was nonideal (as seen in Figure 3F), with a lower area per lipid than a linear

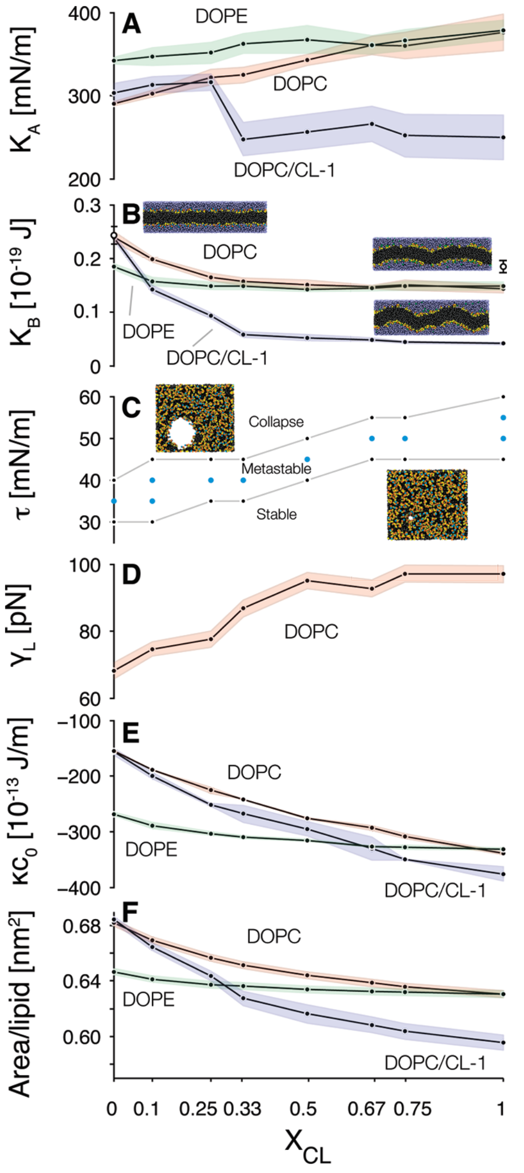


Figure 3. Mechanical properties as a function of mole fraction CL:DOPC and DOPE indicate DOPC/CL-2 and DOPE/CL-2, respectively. From top to bottom: (A) Elastic area compression modulus, K_A. Uncertainties (shaded) are standard deviations from a linear fit to the γ-A isotherms. (B) Bending modulus, K_B. Uncertainties (shaded) are standard deviations from five parts of the 8 μs trajectory. Inset images are snapshots showing bilayer undulations for mole fractions 0 and 1. Open symbols denote an electrostatic cutoff of 2.0 nm. (C) Critical tension, τ. Border between bottom (stable > 1 μs), metastable (40−1000 ns), and collapse (<40 ns) was determined by repeated simulations starting with a porated bilayer. Inset images are snapshots from X_{CL} 0.5 bilayers with a preformed pore at surface tension 40 and 45 mN/m. (D) Line tension, γ_L, for the DOPC/CL-2 bilayer ribbon systems. Uncertainties (shaded) are standard deviations. (E) First moment of lateral pressure profile, κc_0. Uncertainties (shaded) are standard deviations between upper and lower leaflets. (F) Area per lipid (CL counted as two lipids). Uncertainties (shaded) are standard deviations.

combination of the pure components. For a given mole fraction, the deviation from ideality was larger in DOPC/CL-2 than in DOPE/CL-2, which implies that the canonical area per lipid for CL-2 was smaller with DOPC than with DOPE.

Table 1. P−N Vector Tilt Angle in DOPC or DOPE Relative to the Bilayer Normal[a]

	DOPC/CL-2	DOPE/CL-2	DOPC/CL-1
X_{CL} = 0.75	74.4 ± 0.1	86.3 ± 0.1	71.62 ± 0.09
X_{CL} = 0.67	74.5 ± 0.1	86.1 ± 0.1	71.8 ± 0.1
X_{CL} = 0.5	74.52 ± 0.06	85.6 ± 0.1	72.4 ± 0.1
X_{CL} = 0.33	74.66 ± 0.05	84.92 ± 0.07	73.0 ± 0.1
X_{CL} = 0.25	74.68 ± 0.08	84.41 ± 0.05	73.12 ± 0.07
X_{CL} = 0.1	74.48 ± 0.03	83.08 ± 0.04	73.59 ± 0.03
X_{CL} = 0	73.83 ± 0.04	81.33 ± 0.04	73.75 ± 0.02

[a] Errors are standard errors of the mean (SEM) from two leaflets, and trajectory is split into five equal length parts.

The area per lipid of tetraoleoyl−CL is not known, but Goormaghtigh et al. estimated that saturated or unsaturated CL has a surface area of 1.2 nm^2,[47,48] which compares favorably to our results (1.26 and 1.19 nm^2 for CL-2 and CL-1, respectively). Previous all-atom simulations of pure CL bilayers gave an area per lipid of 0.99 nm^2,[34] which corresponds approximately to the experimental area of saturated tetramyristoyl−CL, estimated to 0.5 nm^2 (per two chains) in the fluid lamellar phase.[49] Ion binding in the carbonyl region reduced the area per lipid strongly in the MD simulations. In the present CG approach, such ion binding is much less specific and occurs instead at the level of the phosphate groups, resulting in a higher area per lipid. It should also be noted that the areas of the other components are only in semiquantitative agreement with experiments and that the balance of forces which determine the area will be dependent on all components of the mixture.

As a further characterization of the lipid behavior of the mixtures, we calculated the P−N tilt angle away from the bilayer normal for DOPE and DOPC (Table 1). The averaged P−N tilt was larger for DOPE than DOPC for all compositions, which is consistent with the smaller headgroup in DOPE. In contrast to the minor changes in the P−N tilt in DOPC, the DOPE headgroup tilted further away from the bilayer normal with increasing CL content, which is consistent with the “voltameter” model.[50] This behavior shows that the response of the P−N dipole to the charge of the bilayer is dependent on the size or interaction parameters of the choline/ethanolamine groups. Both effects, smaller volume and increased potential for hydrogen bonding to the phosphate groups of CL, enable larger P−N tilts in DOPE than in DOPC. For the DOPC/CL-1 system, there was even a small decrease in the tilt angle with increasing CL-1 content. It should be noted that increased local curvature in the DOPC/CL-1 systems makes the P−N definition, where the simulation box is used as the reference coordinate system, less useful for comparing with DOPC/CL-2 and DOPE/CL-2 cases where the undulations were more suppressed.

Compression Modulus. In Figure 3A, the elastic area compressibility modulus (elasticity), K_A, defined as $K_A = A \cdot d\gamma/dA$, for the mixed bilayer systems is shown. A series of simulations with increasing surface tension was performed, and the resulting response in the bilayer area was recorded. The uncertainties reported here are standard deviations from the linear fit to the area surface tension data. The overall trend shows an increase in K_A with X_{CL-2}, the effect being stronger with DOPC than with DOPE. At high

CL concentrations, changing the zwitterionic colipid had no significant effect on K_A. The opposite trend, i.e., decreasing K_A with X_{CL-1}, was observed in the CL-1 system above $X_{CL-1} = 0.25$, where K_A dropped significantly. The surface charge density, which is similar although not exactly the same in the DOPC and DOPE systems due to different mean areas, was not directly connected to K_A. We also note that if the net charge density was the main factor determining elasticity, we would expect the elasticity modulus at $X_{CL} = 1$ for CL-1 to correspond to the value at $X_{CL} = 0.5$ for CL-2, which is not the case. Continuum membrane models predict the area modulus to be affected most by the chains.[28] Because no changes were made in the acyl chain part, we would, therefore, expect similar K_A for CL-2 and CL-1. However, the packing parameter of the CL-1 lipids turned out to favor lipid segregation with locally increased curvature. The applied surface tension then not only elastically deforms the bilayer but also restricts undulations. This is related to the apparent area compressibility modulus measured in experiments, where the contribution from thermal undulations has to be disentangled from the elasticity and the system size effect seen in MD simulations of lipid bilayers.

Recently, micropipet and monolayer experiments[19] showed that the apparent area expansion modulus, K_A^{app}, was significantly reduced by introducing CL into SOPC vesicles and egg−PC monolayers, respectively. This runs contrary to the results found here for the CL-2 systems. It is necessary, in principle, to consider not only the concentration of the charged CL but also the change in concentration of different chains. Whereas our simulations were run with constant chain composition, the CL used for the experiments contains significantly more unsaturated chains than SOPC or egg−PC. Unsaturated chains tend not to affect K_A strongly, but can lower K_A^{app} due to a decreased bending modulus.[51,52] The effects of charged lipids on the mechanical properties of bilayers were better separated in a study of POPG/POPA in POPC bilayers,[53] where the area compressibility modulus, up to sensitivity of the experiment, was not changed by the inclusion of as much as 30% anionic lipid. A similar result was found for monolayers of pulmonary surfactant simulated with the same CG force field used here,[54] and no significant difference in the area compression modulus was found with charged lipid content.

Bending Modulus. The bending modulus can be calculated by quantifying the thermal undulations of the bilayer.[55] An alternative method based on pulling bilayer tethers, a process not dependent on thermal excitation of the long wavelength undulations, has been proposed recently.[56] Our main goal was to see the systematic changes in the bending modulus and thus opted for avoiding the added complexity of reliably equilibrating the inner and outer leaflets necessary in the tether method.

Spectral densities were calculated by interpolating the positions of all C2 chain particles to an upper and lower leaflet 200 × 200 grid (approximate grid spacing 0.08 nm) and by averaging over the z-position of the grids. Different grid spacings were tried with consistent results. The 2D-Fourier transformed grids for each trajectory frame were integrated in circles from the zero frequency and then averaged over the 8 μs trajectory. The bending modulus was calculated by fitting a q^{-4} function to the longest undulation modes (the three smallest $\mathbf{q}$ vectors were used) on a log−log scale and by setting the additive constant equal to log (k_BT/AK_B) and evaluating for K_B. Uncertainties in the bending moduli were estimated by splitting the trajectory into five equal length parts.

In Figure 3B, the bending modulus from bilayers at zero surface tension is shown. There was a general tendency toward decreasing K_B with the mole fraction of CL. Bilayers with DOPE had a slightly lower K_B than bilayers with DOPC, although above $X_{CL} = 0.33$ the difference was not significant, and K_B was essentially constant. Overall, the bending moduli for the pure zwitterionic bilayers were lower than the experimental value for DOPC (0.85×10^{-19} J)[52] but close to the range found previously in simulations with similar force fields.[45] We expect our bending moduli to be slightly lower than those found for DPPC, due to the unsaturated chains. This is in agreement with a decrease in the bending modulus with decreased saturation and charge for DPPC/POPG mixtures found by Baoukina et al.[54] Experimental DOPE bending moduli calculated from the H_{II} phase were 20% higher than for DOPC.[57]

A radical decrease in K_B was observed when CL-1 was added. Long undulation modes were increasingly excited, and the undulation spectrum also showed concentration dependence in the short-wavelength region (data not shown), absent in CL-2 simulations. Whereas DOPC and DOPE bilayers with CL-2 had bending moduli in the 3−6 k_BT range, CL-1 showed K_B close to 1 k_BT for all $X_{CL} > 0.25$, which is consistent with the strong thermal excitations of long undulation modes. This effect was also seen in the $X_{CL} = 0.5$ simulation (4 μs) with the −1/0 charge partition in the headgroup. Additionally, 4 μs runs of $X_{CL} = 0$ and 1 with the Nosé-Hoover/Parrinello-Rahman combination did not show significant differences in the bending moduli relative to the Berendsen method.

In recent experiments on DMPC vesicles with adsorbed surfactants, charged or uncharged, the charged surfaces exhibited higher bending rigidity.[33] The surface charge density was kept low in those experiments, with less than 5 mol % adsorbed surfactant. As a test of the dependency on electrostatic interactions, we also calculated the bending moduli for the pure CL and DOPC phases with an increased cutoff length on Coulomb interactions (open circles in Figure 3B). Increasing the cutoff from 1.2 to 2.0 nm increased the bending modulus by 27% for CL and 2% for DOPC, and it has been established previously that the surface concentration of counterions tends to increase with an increased cutoff.[20] Because long-ranged electrostatic interactions between adjacent bilayers can suppress undulations, we tested the effect of the amount of water on the bending modulus and found that K_B was lowered slightly: $0.13 \pm 0.01 \times 10^{-19}$ at 147 waters/lipid, and $0.15 \pm 0.01 \times 10^{-19}$ J at 47 waters/lipid. This corresponds to 55% (147 w/l) and 60% (47 w/l) of the bending modulus of DOPC and shows that the effect of including more water is very limited. The increase in ion binding caused by increasing the electrostatic cutoff radius is most likely part of the reason for the increased cohesive pressure (see Pressure Profiles below).

Critical Tension. As a measure of the stability of the CL bilayers, we subjected bilayers with preformed pores to surface tensions of increasing magnitude. The critical tension was defined as the tension where pore growth was dramatically increased. The time to bilayer collapse was defined, for simulations where collapse was observed, as the simulation time between the onset of applied surface tension and an increase of the projected pore area with a factor of two or more. This definition is arbitrary but gives a simple criterion for dramatic changes to the structure of the bilayer. The critical tension is, in general, dependent on the loading rate.[58] Repeated simulations with our protocol gave critical tensions that were reproducible to within 5 mN/m, which was enough to show the trend of increasing critical tension with CL concentration. When the target surface tension was above the threshold for pore expansion, the actual applied surface tension dropped slightly below the target value. This is expected because the system is not in mechanical equilibrium during pore expansion. In the simulations where a critical tension was applied, the actual surface tension was typically between 93 and 98% of the target tension. Because a large part of pore expansion is doing work against the line tension of the pore, we also measured the line tension with the protocol of Tolpekina et al.[43] The line tension was defined as $\gamma_L = A_{xy}[(P_{yy} + P_{xx})/2 - P_{zz}]/2$, where the bilayer slab is periodic in the z-direction, A_{xy} is the cross section area of the box in the $x-y$ plane, and P_{xx}, P_{yy}, and P_{zz} are the diagonal elements of the pressure tensor.

The increase in line tension with CL concentration is shown in Figure 3C, and the effect was largest for small amounts of CL and turning essentially constant above $X_{CL} = 0.5$. From a geometric point of view, lipids with a positive curvature, e.g., micellar surfactants, can stabilize the edge[59] and thereby allow pore growth. Conversely, inclusion of inverted cone shaped lipids, such as cholesterol, has been shown to increase line tension in DOPC vesicles.[60] Additionally, membranes with PE lipids and anionic PS lipids have been shown to have a higher line tension than neutral PC membranes.[61] By changing the spontaneous curvature through the addition of CL, we thus expected pores to be less likely to expand, which is what we observed in the critical tension simulations. The calculated line tension for DOPC, 68 ± 2 pN, is comparable to previous results with the MARTINI force field (62−64 pN for DPPC/POPG mixtures) but larger than the experimentally determined line tension, which is in the 7−25 pN range.[58,60,61]

Nichols-Smith et al.[19] measured the lysis tension (critical tension) in SOPC membranes with CL. For CL concentrations of 5 and 9.2%, the lysis tension was lowered by 3.5 and 5.1 mN/m, respectively, which is the opposite of what we observe. A possible explanation for the discrepancy is that unsaturated chains (CL was more unsaturated than SOPC) lower the bending modulus and the lysis tension.[51,52] The breakdown voltage for black lipid membranes with the charged lipid phosphatidylserine or PC was about equal and independent of ionic strength.[62] In the same experiments, the chain volume was found to correlate negatively with membrane rupture. Keeping chain volume fixed, as in our model, and instead varying the headgroup volumes, the trend we observed suggests that the correlation carries over from absolute into relative chain volumes, i.e., lower headgroup volumes giving higher lysis tensions.

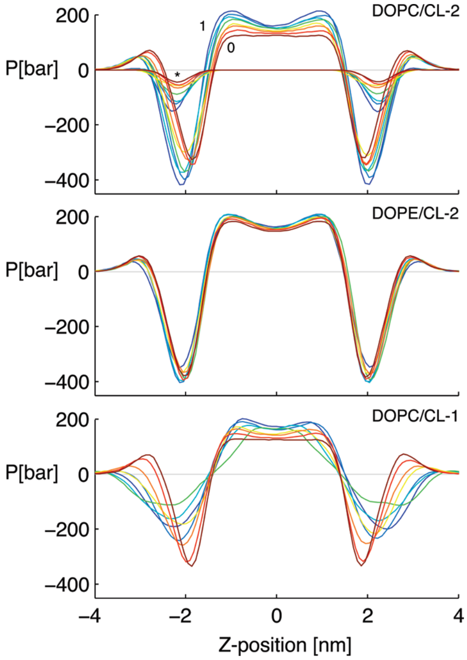


Figure 4. Pressure profiles as a function of distance from center of bilayers, from top to bottom: DOPC/CL-2 with the electrostatic component (*), DOPE/CL-2, and DOPC/CL-1. Colors denote X_{CL} from 0 (red) through 1 (blue).

Pores have two different curvatures: one negative in the bilayer plane and one positive, which has a radius determined by the monolayer thickness, tracing lines orthogonal to the first curvature. Recent simulations show that effects of the negative curvature can be neglected even for radii corresponding to pore closure.[63] Using the relationship between line tension, γ_L, and surface tension, Γ, we calculate an approximate radius at which the calculated line tension for the pore edge balances the surface tension created by opening the pore, i.e., $r = \gamma_L/\Gamma$. The calculated critical pore size by this method was 1.8 ± 0.1 nm across the range of concentrations at the surface tension corresponding to the critical tension and 2.3 ± 0.2 nm at the border between stable and metastable pores. These radii are larger than what was observed in the simulations, where the hydrocarbon pore was ~1.3 nm, and the water cylinder in the pore was ~1.1 nm but in qualitative agreement.

Local Pressure Profiles. We investigated the connection between bilayer behavior and interactions in the bilayer by calculating local pressure profiles along the bilayer normal. The pressure profiles for all compositions, shown in Figure 4, were calculated according to Lindahl et al.,[41] and the difference between the lateral and normal pressures was binned into 100 bins (approximately 0.1 nm per bin) along the z-axis. In the DOPC/CL-2 system, there were systematic changes to all three main regions of the pressure profile: the repulsive chain region, the attractive interface region, and the repulsive headgroup region. The pressure in the chain region increased with CL concentration. In the interface region, there was a shift of about 0.2 nm away from the center of the bilayer in going from $X_{\mathrm{CL}} = 0$ to $X_{\mathrm{CL}} = 1$. The interface pressure was also increasingly negative. Finally, the outer region had a similar outward shift as the interface peak, but the pressure decreased in magnitude. These changes can be explained by a decrease in the effective headgroup volume, which decreases the headgroup pressure and packs the lipids tighter. The effect can be correlated to the area per lipid, which decreased with the CL mole fraction (see Figure 3E). In the DOPE/CL-2 system, the local pressure profile was practically unchanged with concentration, but had high chain pressures and increased interface attraction as well as lower headgroup repulsion relative to the DOPC/CL-2 system. More drastic changes were observed in the DOPC/CL-1 pressure profiles. The most significant effects were: a reduced interface pressure, a stronger X_{CL} dependence, and a flattening of the headgroup repulsion profile. Reduced headgroup pressure, brought about by decreasing the phosphate charges, is consistent with the partial segregation seen in these systems. A general broadening of the interface was seen, which is explained in part by the increase in undulations. The slight asymmetry seen for the in pressure profile is also due to undulations, which make the interface less well-defined and the pressure statistics poorer.

From a decomposition of the lateral pressures into the respective interactions in the DOPC/CL-2 system, it was found that the electrostatic interactions had a net negative pressure, corresponding to a positive surface tension (see Figure 4). The dominant component of the electrostatic interactions was the CL−ion interaction which generated pressures on the order of −400 bar with $X_{\mathrm{CL}} = 0$.

Spontaneous Curvature. As a measure of the tendency to form inverted phases, we used the first moment of the lateral pressure profile, which is proportional to the spontaneous curvature of the monolayer, c_0, and to the bending modulus (here denoted κ): $\kappa c_0 = \int_0^\infty z\Sigma(z)dz$ and is independent of the position of the reference point along the bilayer normal, if the total surface tension is zero and if $\Sigma(z) = \langle P_z(z) - P_{\|}(z)\rangle$, where $P_{\|}(z)$ is the pressure in the bilayer plane. Negative spontaneous curvature was observed for all systems over the entire CL concentration range (see Figure 3E). The effects of a smaller zwitterionic headgroup (DOPE as compared to DOPC) and of a less charged headgroup (CL-1 as compared to CL-2) were all visible in the spontaneous curvature. Interestingly, the combination of DOPE and CL-2 had a more negative spontaneous curvature for low CL concentrations than that of the CL-1 system with reduced charge.

Order Parameters. The sequential particle−particle order parameters give a molecule centric view of the effects of local environment. Overall, order parameters were very similar for CL and DOPC/DOPE, with differences located mainly in the headgroups, see Figure 5. In the chain region the order was increased slightly with increasing CL-2 concentrations, and the change was the largest near the end of the tails. Headgroup order changes as a function of X_{CL} were minor for both components in the DOPC/CL-2 system. Chain order was slightly higher for DOPE/CL-2 than for DOPC/CL-2, but variations with composition were minute. This is consistent with the pressure profile differences between the systems, where higher chain pressures for DOPE were observed. In the reduced charge systems, the order parameter showed a tendency to decrease for all segments in the polar part of the lipid but tended to increase in the two last segments. The increased chain order can be explained by the lower area per lipid caused by a reduced repulsive interaction between lipids. Undulations and locally increased curvature, both of which were seen with CL-1, tend to weaken order in the polar part of the lipid.

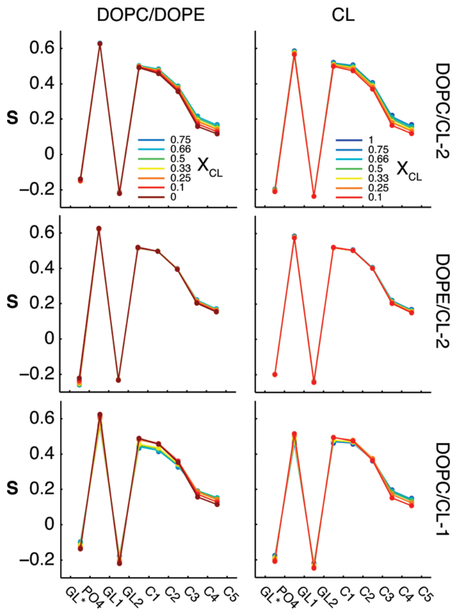


Figure 5. Sequential order parameters, *S*, as a function of position in DOPC/DOPE (left) and CL (right). $X_{\mathrm{CL}} = 0$ (red) through $X_{\mathrm{CL}} = 1$ (blue).

As mentioned above, the effect of increasing the water content perturbs the system only slightly. The order parameters thus were at most decreased by 4% in the 147 w/l system, relative to the 47 w/l system.

Counterion Profiles. A potential shortcoming of these CG models is the treatment of electrostatics. As a guide in judging the effect of neglecting the long-ranged interactions, we calculated the counterion profile in the bilayer normal direction (see Figure 6). Predictably, the ion profiles were essentially flat outside a cutoff distance from the charge headgroups, but the majority of the counterions were adsorbed to the interface, which is the behavior observed in atomistic simulations.[34–36] For mixed DOPC/CL-2 bilayers, the profile showed a minimum just outside the headgroup region, which we attribute to the effect of charge interactions with the P−N dipole and to the shifted balance in the interactions between water and ion particles with the choline particle type. In DOPE bilayers, the dipole tilt is significantly higher, essentially in the bilayer plane, and the ethanolamine particle type is more similar in its interactions with water and ions.

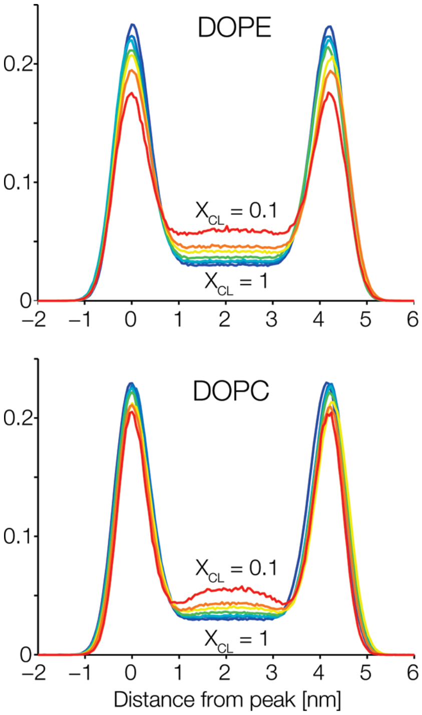


Figure 6. Counterion density distributions for DOPE/CL-2 and DOPC/CL-2 as a function of position along the bilayer normal, with the origin set at the maximum of the distribution. Normalized with the total number of ions in the system. Arbitrary units on the vertical axis. X_{CL} = 0.1 (red) through X_{CL} = 1 (blue).

Interestingly, the local minimum seen in the PC simulations has also been observed in atomistic simulations.[64,65]

Discussion

The main finding from this coarse grained (CG) model of cardiolipin (CL) membranes was that the rigidity of the bilayer was correlated to the effective headgroup volume, so that small headgroups were associated with lower bending moduli. In contrast to continuum models, where bending deformation of the membrane is directly proportional to the elasticity of the monolayers, the CL with full headgroup charge in our study exhibited an inverse relation between these two quantities. We note that the changes in area and bending moduli for the DOPC and DOPE systems were small and that these differences may not be large enough to be distinguishable experimentally over the range of CL concentrations. The spontaneous curvature aggregate κc_0, calculated from the pressure profiles, was increasingly negative with CL concentrations and lower for DOPE than for DOPC. This is consistent with the notion that small headgroup volumes tend to give negative curvatures, and here we have shown that this was true over the entire concentration range, with κc_0 monotonically decreasing as a function of X_{CL}. In the case of DOPC, the CL headgroup charge was clearly connected to the spontaneous curvature, and we observed microdomain formation when the charge was reduced on each CL. The area elasticity modulus in the CL-1 system was also low, which can be explained as an effect of the decreased bending modulus, which leads to a undulation dominated area dilation with logarithmic increase in area for a given surface tension.[66] The bending modulus in CL-1 was on the order of k_BT, which is consistent with the large undulations seen, and also indicates that the lamellar phase becomes destabilized upon charge neutralization. This effect was not sensitive to the choice of charge partition (−1/0 or −0.5/−0.5) in the headgroup. For DOPE, the X_{CL} dependence of κc_0 was lower, which is explained by headgroup sizes in DOPE and CL-2 being quite similar.

Net attractive interactions (negative lateral pressures) were found for the electrostatics in the headgroup region, increasing with the mole fraction of CL. The dominant component of the electrostatic interactions was between counterions and CL. Smaller in magnitude, and opposite in sign, was the CL−CL component of the pressure. With increasing CL concentration, the combination of increased chain pressure, decreased headgroup repulsion, and a net attractive electrostatic component gave a monotonously decreasing κc_0. It should be pointed out that κc_0 is independent of the position of the pivot plane for pressure profiles with zero surface tension but that the contribution from the components is not. For pivot plane positions close to the interface, the electrostatic component was still found to give a negative contribution to κc_0, i.e. negative curvatures. Thus, introducing headgroups of the same charge is not necessarily associated with reduced interface cohesion, and effective headgroup volume must be taken into account to accurately predict the change in mechanical properties induced by charged lipids.

We found that the line tension increased significantly with CL concentration. We understand this also as a consequence of the average effective headgroup volume, which will force inverse cone-shaped lipids away from pores, thus stabilizing the bilayer. Our two methods, the stability of a porated bilayer under tension and the bilayer ribbon can be seen as opposite extremes of pore radius (~1 nm and infinite, respectively). Because both methods gave the same trend in the line tension, we conclude that a small pore radius is not stabilized due to the negative curvature of the monolayers. This is in agreement with the results of Wohlert et al.[63] The increase in line tension with CL concentration, also at physiological levels of CL, stabilizes the membrane against pore expansion and rupture, at least on the time and length scales of the present simulations.

A notable feature of the CL−DOPC mixtures was the systematic change in chain pressure just beneath the interface that we observed as a function of the composition. With pressure differences on the order of 100 bar, CL can affect other lipids and membrane proteins even if direct binding of the charged headgroup is not possible. The observed trends can be understood in the context of effective headgroup volume: DOPE and CL with smaller headgroups having an overall larger chain pressure than DOPC. Similarly, small headgroups were associated with high area modulus (high chain pressures), as long as CL charge was high.

The lateral pressure profile is generally built up from large, and to a high degree canceling, components. The balance of forces across the membrane is sensitively dependent on the respective interactions, and the coarse nature of the model presented here does not allow us to predict the behavior of specific compositions or of spatial detail smaller than approximately 0.5 nm. In our view, we should instead use the systematic variations of easily controlled parameters to get a feeling for the minimum required parameters to model CL membranes. The CG model, which this work is based on, has been shown to reproduce many of the crucial properties of zwitterionic membranes, and our CL model only changes the headgroup interactions and the connectivity. A limiting factor in improving the present model is the lack of experimental results, such as area per lipid, bending modulus, and critical tension, controlled for CL chain composition. The headgroup properties of CL are still not well understood, and further experiments will be needed to determine reliably the CL area and the charge in mixtures with zwitterionic lipids.

The simulations presented in this study, with the exception of the pore simulations, were stable on the microsecond time scale. For DOPC and pure CL-2, we infer from previous modeling work that the equilibrium state is the bilayer. We have observed hexagonal phase formation for DOPE and CL-1 at 310 K, starting from isotropic lipid−water mixtures (DOPE) and stalked bilayer stacks (CL-1) (data not shown). A limitation to the present work is that the true equilibrium properties of the lipid mixtures are not fully known. We thus emphasize that the mechanical properties and the structural features found here only indicate the trends and that they should be interpreted as perturbations of the equilibrium states of the pure and stable bilayer membranes. As such, the local lipid segregation we found with CL-1 mixtures should not primarily be seen as evidence of domain formation in the sense reported recently[67–69] but rather as a mechanism of dissipating curvature frustration when the barrier to phase transition is high due to the high water content. Changing CL charge from −2 to −1 reduces electrostatic repulsion and increases “hydrogen bonding” (determined by the Lennard-Jones parameters), which together lowers the effective size of the headgroup. Because the total curvature in the domains remains fairly small and only small changes in bilayer thickness occurred, it is unlikely that domain formation is driven by a change in the electrostatic screening length, as seen in micelle fission by Sammalkorpi et al.[70]

The transition in bending modulus seen for CL-1 is compatible with the nonequilibrium vesicle system of Khalifat et al.,[11] showing that local pH manipulations (from global pH 8 to estimated local pH 4−5) can cause structural changes in the membrane. We note that the inclusion of CL in either its −1 or −2 state decreased the rigidity of the bilayer, which is compatible with highly curved mitochondrial membranes. Considering a second pK_a in the 7.5−9.5 region for CL, a possible mapping between the charge states in the simulations and the experiments is −2 at pH 8 and −1 at pH 4.

An important limitation of this work is the treatment of the electrostatics. We have quantified this in the counterion profiles, which agree qualitatively with atomistic simulations. However, in the aqueous phase the profiles deviate predictably—due to the cutoff—from the smoothly decaying behavior predicted from a Poisson−Boltzmann treatment. More rigorous coarse graining of electrostatics is being developed by others,[71,72] showing promising results and could be used to improve the model presented here. Another direction is to rationalize the cutoff as an effect of screening by a low amount of (virtual) salt. For Debye lengths comparable to a 1.2 nm cutoff, the monovalent salt solutions are in the 10−60 mM range (with limits taken for effective dielectric constants 15 and 80, respectively), which is low compared to the physiologically relevant salt concentration (approximately 200 mM). Ultimately, using a uniform effective dielectric constant might prove too coarse for charged species at the interface, but the work presented here gives some predictions that can be tested experimentally and with other molecular electrostatic models.

Acknowledgment. We thank Erik Lindahl for fruitful discussions. This work was supported by grants from the Swedish Research Council (VR) and the Carl Trygger Foundation.

Supporting Information Available: Details of the updated coarse grained cardiolipin model. This material is available free of charge via the Internet at http://pubs.acs.org.

Abbreviations. CL, cardiolipin; DOPC, 1,2-dioleoyl-glycero-3-phosphatidylcholine; DOPE, 1,2-dioleoyl-glycero-3-phosphatidylethanolamine; POPC, 1-palmitoyl-2-oleoyl-glycero-3-phosphatidylcholine; POPG, 1-palmitoyl-2-oleoyl-glycero-3-phosphatidylglycerol; POPA, 1-palmitoyl-2-oleoyl-glycero-3-phosphatidic acid; SOPC, 1-stearoyl-2-oleoyl-glycero-3-phosphatidylcholine; DMPC, 1,2-dimyristoyl-glycero-3-phosphatidylcholine; DPPC, 1,2-dipalmitoyl-glycero-3-phosphatidylcholine.

References

(1) Bloom, M.; Evans, E.; Mouritsen, O. G. Physical Properties of the Fluid Lipid-Bilayer Component of Cell Membranes: A Perspective. *Q. Rev. Biophys.* **1991**, *24*, 293–397.

(2) Zimmerberg, J.; Gawrisch, K. The Physical Chemistry of Biological Membranes. *Nat. Chem. Biol.* **2006**, *2*, 564–567.

(3) Marsh, D. Protein Modulation of Lipids, and Vice-Versa, in Membranes. *Biochim. Biophys. Acta* **2008**, *1778*, 1545–1575.

(4) Frey, T. G.; Mannella, C. A. The Internal Structure of Mitochondria. *Trends Biochem. Sci.* **2000**, *25*, 319–324.

(5) John, G. B.; Shang, Y.; Li, L.; Renken, C.; Mannella, C. A.; Selker, J. M.; Rangell, L.; Bennett, M. J.; Zha, J. The Mitochondrial Inner Membrane Protein Mitofilin Controls Cristae Morphology. *Mol. Biol. Cell* **2005**, *16*, 1543–1554.

(6) Mannella, C. A. The Relevance of Mitochondrial Membrane Topology to Mitochondrial Function. *Biochim. Biophys. Acta* **2006**, *1762*, 140–147.

(7) Mannella, C. A. Structure and Dynamics of the Mitochondrial Inner Membrane Cristae. *Biochim. Biophys. Acta* **2006**, *1763*, 542–548.

(8) Benard, G.; Rossignol, R. Ultrastructure of the Mitochondrion and its Bearing on Function and Bioenergetics. *Antioxid. Redox Signaling* **2008**, *10*, 1313–1342.

(9) Heath-Engel, H. M.; Shore, G. C. Mitochondrial Membrane Dynamics, Cristae Remodelling and Apoptosis. *Biochim. Biophys. Acta* **2006**, *1763*, 549–560.

(10) Phillips, R.; Ursell, T.; Wiggins, P.; Sens, P. Emerging Roles for Lipids in Shaping Membrane-Protein Function. *Nature* **2009**, *459*, 379–385.

(11) Khalifat, N.; Puff, N.; Bonneau, S.; Fournier, J. B.; Angelova, M. I. Membrane Deformation Under Local pH Gradient: Mimicking Mitochondrial Cristae Dynamics. *Biophys. J.* **2008**, *95*, 4924–4933.

(12) Hoch, F. L. Cardiolipins and Biomembrane Function. *Biochim. Biophys. Acta* **1992**, *1113*, 71–133.

(13) Daum, G. Lipids of Mitochondria. *Biochim. Biophys. Acta* **1985**, *822*, 1–42.

(14) Schlame, M.; Rua, D.; Greenberg, M. L. The Biosynthesis and Functional Role of Cardiolipin. *Prog. Lipid Res.* **2000**, *39*, 257–288.

(15) Kates, M.; Syz, J. Y.; Gosser, D.; Haines, T. H. PH-Dissociation Characteristics of Cardiolipin and its 2′-Deoxy Analogue. *Lipids* **1993**, *28*, 877–882.

(16) Haines, T. H.; Dencher, N. A. Cardiolipin: A Proton Trap for Oxidative Phosphorylation. *FEBS Lett.* **2002**, *528*, 35–39.

(17) Lewis, R. N.; McElhaney, R. N. Surface Charge Markedly Attenuates the Nonlamellar Phase-Forming Propensities of Lipid Bilayer Membranes: Calorimetric and (31)P-Nuclear Magnetic Resonance Studies of Mixtures of Cationic, Anionic, and Zwitterionic Lipids. *Biophys. J.* **2000**, *79*, 1455–1464.

(18) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P. de Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

(19) Nichols-Smith, S.; Teh, S. Y.; Kuhl, T. L. Thermodynamic and Mechanical Properties of Model Mitochondrial Membranes. *Biochim. Biophys. Acta* **2004**, *1663*, 82–88.

(20) Dahlberg, M. Polymorphic Phase Behavior of Cardiolipin Derivatives Studied by Coarse-Grained Molecular Dynamics. *J. Phys. Chem. B* **2007**, *111*, 7194–7200.

(21) Seddon, J. M.; Kaye, R. D.; Marsh, D. Induction of the Lamellar-Inverted Hexagonal Phase Transition in Cardiolipin by Protons and Monovalent Cations. *Biochim. Biophys. Acta* **1983**, *734*, 347–352.

(22) Ioannou, P. V.; Golding, B. T. Cardiolipins: Their Chemistry and Biochemistry. *Prog. Lipid Res.* **1979**, *17*, 279–318.

(23) Matsumoto, K.; Kusaka, J.; Nishibori, A.; Hara, H. Lipid Domains in Bacterial Membranes. *Mol. Microbiol.* **2006**, *61*, 1110–1117.

(24) Osman, C.; Haag, M.; Potting, C.; Rodenfels, J.; Dip, P. V.; Wieland, F. T.; Brugger, B.; Westermann, B.; Langer, T. The Genetic Interactome of Prohibitins: Coordinated Control of Cardiolipin and Phosphatidylethanolamine by Conserved Regulators in Mitochondria. *J. Cell Biol.* **2009**, *184*, 583–596.

(25) Gohil, V. M.; Thompson, M. N.; Greenberg, M. L. Synthetic Lethal Interaction of the Mitochondrial Phosphatidylethanolamine and Cardiolipin Biosynthetic Pathways in Saccharomyces Cerevisiae. *J. Biol. Chem.* **2005**, *280*, 35410–35416.

(26) Gohil, V. M.; Greenberg, M. L. Mitochondrial Membrane Biogenesis: Phospholipids and Proteins Go Hand in Hand. *J. Cell Biol.* **2009**, *184*, 469–472.

(27) Helfrich, W. Elastic Properties of Lipid Bilayers: Theory and Possible Experiments. *Z. Naturforsch., C: J. Biosci.* **1973**, *28*, 693–703.

(28) Szleifer, I.; Kramer, D.; Ben-Shaul, A.; Gelbart, W. M.; Safran, S. A. Molecular Theory of Curvature Elasticity in Surfactant Films. *J. Chem. Phys.* **1990**, *92*, 6800–6817.

(29) Winterhalter, M.; Helfrich, W. Effect of Surface Charge on the Curvature Elasticity of Membranes. *J. Phys. Chem.* **1988**, *92*, 6865–6867.

(30) Andelman, D. Electrostatic Properties of Membranes: The Poisson−Boltzmann Theory. In *Handbook of Biological Physics*; Lipowsky, R., Sackmann, E., Eds.; Elsevier Science: Amsterdam, The Netherlands, 1995; pp 603.

(31) Fogden, A.; Ninham, B. W. Electrostatics of Curved Fluid Membranes: The Interplay of Direct Interactions and Fluctuations in Charged Lamellar Phases. *Adv. Colloid Interface Sci.* **1999**, *83*, 85–110.

(32) Fuller, N.; Benatti, C. R.; Rand, R. P. Curvature and Bending Constants for Phosphatidylserine-Containing Membranes. *Biophys. J.* **2003**, *85*, 1667–1674.

(33) Rowat, A. C.; Hansen, P. L.; Ipsen, J. H. Experimental Evidence of the Electrostatic Contribution to Membrane Bending Rigidity. *Europhys. Lett.* **2004**, *67*, 144–149.

(34) Dahlberg, M.; Maliniak, A. Molecular Dynamics Simulations of Cardiolipin Bilayers. *J. Phys. Chem. B.* **2008**, *112*, 11655–11663.

(35) Rog, T.; Martinez-Seara, H.; Munck, N.; Oresic, M.; Karttunen, M.; Vattulainen, I. Role of Cardiolipins in the Inner Mitochondrial Membrane: Insight Gained through Atom-Scale Simulations. *J. Phys. Chem. B* **2009**, *113*, 3413–3422.

(36) Pöyry, S.; Róg, T.; Karttunen, M.; Vattulainen, I. Mitochondrial Membranes with Mono- and Divalent Salt: Changes Induced by Salt Ions on Structure and Dynamics. *J. Phys. Chem. B* **2009**, *113*, 15513–15521.

(37) Dickey, A.; Faller, R. Examining the Contributions of Lipid Shape and Headgroup Charge on Bilayer Behavior. *Biophys. J.* **2008**, *95*, 2636–2646.

(38) Bennun, S. V.; Hoopes, M. I.; Xing, C.; Faller, R. Coarse-Grained Modeling of Lipids. *Chem. Phys. Lipids* **2009**, *159*, 59–66.

(39) Dahlberg, M.; Marini, A.; Mennucci, B.; Maliniak, A. Quantum Chemical Modeling of the Cardiolipin Headgroup. *J. Phys. Chem. A* **2010**, *114*, 4375–4387.

(40) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comp.* **2008**, *4*, 435–447.

(41) Lindahl, E.; Edholm, O. Spatial and Energetic-Entropic Decomposition of Surface Tension in Lipid Bilayers from Molecular Dynamics Simulations. *J. Chem. Phys.* **2000**, *113*, 3882–3893.

(42) Jendrasiak, G. L.; Hasty, J. H. The Hydration of Phospholipids. *Biochim. Biophys. Acta* **1974**, *337*, 79–91.

(43) Tolpekina, T. V.; den Otter, W. K.; Briels, W. J. Simulations of Stable Pores in Membranes: System Size Dependence and Line Tension. *J. Chem. Phys.* **2004**, *121*, 8014–8020.

(44) Rand, R. P.; Parsegian, V. A. Hydration Forces between Phospholipid Bilayers. *Biochim. Biophys. Acta* **1989**, *988*, 351–376.

(45) Marrink, S. J.; deVries, A. H.; Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B* **2004**, *108*, 750–760.

(46) Kucerka, N.; Gallova, J.; Uhrikova, D.; Balgavy, P.; Bulacu, M.; Marrink, S. J.; Katsaras, J. Areas of Monounsaturated Diacylphosphatidylcholines. *Biophys. J.* **2009**, *97*, 1926–1932.

(47) Goormaghtigh, E.; Chatelain, P.; Caspers, J.; Ruysschaert, J. M. Evidence of a Complex between Adriamycin Derivatives and Cardiolipin: Possible Role in Cardiotoxicity. *Biochem. Pharmacol.* **1980**, *29*, 3003–3010.

(48) Goormaghtigh, E.; Huart, P.; Praet, M.; Brasseur, R.; Ruysschaert, J.-. Structure of the Adriamycin-Cardiolipin Complex: Role in Mitochondrial Toxicity. *Biophys. Chem.* **1990**, *35*, 247–257.

(49) Lewis, R. N.; Zweytick, D.; Pabst, G.; Lohner, K.; McElhaney, R. N. Calorimetric, X-Ray Diffraction and Spectroscopic Studies of the Thermotropic Phase Behavior and Organization of Tetramyristoyl Cardiolipin Membranes. *Biophys. J.* **2007**, *92*, 3166–77.

(50) Pinheiro, T. J. T.; Duralski, A. A.; Watts, A. Phospholipid Headgroup-Headgroup Electrostatic Interactions in Mixed Bilayers of Cardiolipin with Phosphatidylcholines Studied by H-2 NMR. *Biochemistry* **1994**, *33*, 4896–4902.

(51) Olbrich, K.; Rawicz, W.; Needham, D.; Evans, E. Water Permeability and Mechanical Strength of Polyunsaturated Lipid Bilayers. *Biophys. J.* **2000**, *79*, 321–327.

(52) Rawicz, W.; Olbrich, K. C.; McIntosh, T.; Needham, D.; Evans, E. Effect of Chain Length and Unsaturation on Elasticity of Lipid Bilayers. *Biophys. J.* **2000**, *79*, 328–339.

(53) Shoemaker, S. D.; Vanderlick, T. K. Intramembrane Electrostatic Interactions Destabilize Lipid Vesicles. *Biophys. J.* **2002**, *83*, 2007–2014.

(54) Baoukina, S.; Monticelli, L.; Amrein, M.; Tieleman, D. P. The Molecular Mechanism of Monolayer-Bilayer Transformations of Lung Surfactant from Molecular Dynamics Simulations. *Biophys. J.* **2007**, *93*, 3775–3782.

(55) Lindahl, E.; Edholm, O. Mesoscopic Undulations and Thickness Fluctuations in Lipid Bilayers from Molecular Dynamics Simulations. *Biophys. J.* **2000**, *79*, 426–433.

(56) Harmandaris, V. A.; Deserno, M. A Novel Method for Measuring the Bending Rigidity of Model Lipid Membranes by Simulating Tethers. *J. Chem. Phys.* **2006**, *125*, 204905.

(57) Chen, Z.; Rand, R. P. The Influence of Cholesterol on Phospholipid Membrane Curvature and Bending Elasticity. *Biophys. J.* **1997**, *73*, 267–276.

(58) Evans, E.; Heinrich, V. Dynamic Strength of Fluid Membranes. *C. R. Phys.* **2003**, *4*, 265–274.

(59) Wang, H.; de Joannis, J.; Jiang, Y.; Gaulding, J. C.; Albrecht, B.; Yin, F.; Khanna, K.; Kindt, J. T. Bilayer Edge and Curvature Effects on Partitioning of Lipids by Tail Length: Atomistic Simulations. *Biophys. J.* **2008**, *95*, 2647–2657.

(60) Karatekin, E.; Sandre, O.; Guitouni, H.; Borghi, N.; Puech, P. H.; Brochard-Wyart, F. Cascades of Transient Pores in Giant Vesicles: Line Tension and Transport. *Biophys. J.* **2003**, *84*, 1734–1749.

(61) Genco, I.; Gliozzi, A.; Relini, A.; Robello, M.; Scalas, E. Electroporation in Symmetric and Asymmetric Membranes. *Biochim. Biophys. Acta* **1993**, *1149*, 10–18.

(62) Diederich, A.; Bähr, G.; Winterhalter, M. Influence of Surface Charges on the Rupture of Black Lipid Membranes. *Phys. Rev. E* **1998**, *58*, 4883.

(63) Wohlert, J.; den Otter, W. K.; Edholm, O.; Briels, W. J. Free Energy of a Trans-Membrane Pore Calculated from Atomistic Molecular Dynamics Simulations. *J. Chem. Phys.* **2006**, *124*, 154905.

(64) Pandit, S. A.; Bostick, D.; Berkowitz, M. L. Mixed Bilayer Containing Dipalmitoylphosphatidylcholine and Dipalmitoylphosphatidylserine: Lipid Complexation, Ion Binding, and Electrostatics. *Biophys. J.* **2003**, *85*, 3120–3131.

(65) Vacha, R.; Siu, S. W.; Petrov, M.; Bockmann, R. A.; Barucha-Kraszewska, J.; Jurkiewicz, P.; Hof, M.; Berkowitz, M. L.; Jungwirth, P. Effects of Alkali Cations and Halide Anions on the DOPC Lipid Membrane. *J. Phys. Chem. A* **2009**, *113*, 7235–7243.

(66) Evans, E.; Rawicz, W. Entropy-Driven Tension and Bending Elasticity in Condensed-Fluid Membranes. *Phys. Rev. Lett.* **1990**, *64*, 2094–2097.

(67) Sennato, S.; Bordi, F.; Cametti, C.; Coluzza, C.; Desideri, A.; Rufini, S. Evidence of Domain Formation in Cardiolipin-Glycerophospholipid Mixed Monolayers. A Thermodynamic and AFM Study. *J. Phys. Chem. B* **2005**, *109*, 15950–15957.

(68) Domenech, O.; Redondo, L.; Picas, L.; Morros, A.; Montero, M. T.; Hernandez-Borrell, J. Atomic Force Microscopy Characterization of Supported Planar Bilayers that Mimic the Mitochondrial Inner Membrane. *J. Mol. Recognit.* **2007**, *20*, 546–553.

(69) Domenech, O.; Morros, A.; Cabanas, M. E.; Montero, M. T.; Hernandez-Borrell, J. Thermal Response of Domains in Cardiolipin Content Bilayers. *Ultramicroscopy* **2007**, *107*, 943–947.

(70) Sammalkorpi, M.; Karttunen, M.; Haataja, M. Micelle Fission through Surface Instability and Formation of an Interdigitating Stalk. *J. Am. Chem. Soc.* **2008**, *130*, 17977–17980.

(71) Izvekov, S.; Swanson, J. M. J.; Voth, G. A. Coarse-Graining in Interaction Space: A Systematic Approach for Replacing Long-Range Electrostatics with Short-Range Potentials. *J. Phys. Chem. B* **2008**, *112*, 4711–4724.

(72) Shi, Q.; Liu, P.; Voth, G. A. Coarse-Graining in Interaction Space: An Analytical Approximation for the Effective Short-Ranged Electrostatics. *J. Phys. Chem. B* **2008**, *112*, 16230–16237.

CT900654E

JCTC Journal of Chemical Theory and Computation

The Calculation of NMR Chemical Shifts in Periodic Systems Based on Gauge Including Atomic Orbitals and Density Functional Theory

Dmitry Skachkov, Mykhaylo Krykunov, Eugene Kadantsev, and Tom Ziegler*

Department of Chemistry, University of Calgary, Calgary, Alberta, Canada T2N 1N4

Received January 22, 2010

Abstract: We present here a method that can calculate NMR shielding tensors from first principles for systems with translational invariance. Our approach is based on Kohn−Sham density functional theory and gauge-including atomic orbitals. Our scheme determines the shielding tensor as the second derivative of the total electronic energy with respect to an external magnetic field and a nuclear magnetic moment. The induced current density due to a periodic perturbation from nuclear magnetic moments is obtained through numerical differentiation, whereas the influence of the responding perturbation in terms of the external magnetic field is evaluated analytically. The method is implemented into the periodic program BAND. It employs a Bloch basis set made up of Slater-type or numeric atomic orbitals and represents the Kohn−Sham potential fully without the use of effective core potentials. Results from calculations of NMR shielding constants based on the present approach are presented for isolated molecules as well as systems with one-, two- and three-dimensional periodicity. The reported values are compared to experiment and results from calculations on cluster models.

1. Introduction

NMR shielding tensors can convey very important information about the local electronic structure around a nucleus in a periodic solid. It is thus not surprising that solid-state NMR is an active field of experimental research. This area has in recent years been supplemented with a number of computational schemes that are able to evaluate NMR shielding tensors from first principle.[1−6] All these methods determine the shielding tensor as the second derivative of the total electronic energy with respect to an external magnetic field and a nuclear magnetic moment in one of two ways. In the first approach, the external magnetic field is considered as the initial perturbation inducing a current density and the nuclear magnetic dipole as the second perturbation responding to the induced current density. This is the order for the perturbations adopted in molecular NMR calculations as well as a recent periodic gauge-including projector augmented-wave (GIPAW) method developed by Mauri et al.[2,3] in which the external magnetic field is further considered as oscillating in order to adopt to the periodic symmetry of the solid. In the second converse approach, the order of the two perturbations is interchanged so that now the current density is induced by magnetic dipoles, whereas the external magnetic field is the responding perturbation. The converse approach has been pioneered by Thonhauser et al.[4,5] in conjunction with GIPAW corrections and supercell techniques. It has the merit for solids that the first perturbation, due to the magnetic dipoles, can be considered periodic. However, special care must still be exercised in connection with the constant and nonperiodic external magnetic field. Sebastiani and coauthors[6] applied an infinitesimal magnetic field and employed localized Wannier orbitals constructed from plane waves with continuous set of gauge transformations (CSGT) gauge corrections.[7] In order to obtain sufficiently localized Wannier functions, Sebastiani too employed a supercell technique. All three implementations mentioned above make use of pseudopotentials and plane waves.

The objective of this work is to develop a method for calculating the NMR chemical shift in periodic systems within the full potential program BAND[8−11] in which use is made of atom-centered basis functions. Some of the magnetic properties (EPR g- and A-tensors) have already

* Corresponding author e-mail: ziegler@ucalgary.ca.

10.1021/ct100046a © 2010 American Chemical Society
Published on Web 04/07/2010

been implemented in BAND.[12,13] In BAND the Bloch basis set is constructed from Slater-type orbitals (STOs) and/or numeric atomic orbitals (NAOs). The electronic density matrix near the nuclei is very important for NMR shielding and both STOs and NAOs afford a potentially accurate description of the Kohn−Sham (KS) orbitals in this region. Atomic centered basis functions allow for further use of gauge-included atomic orbitals (GIAOs) to ensure gauge invariant results.

We introduce in Section 2 a method for calculating the NMR chemical shift in periodic systems based on atom-centered basis functions with the computational details characteristic for BAND discussed in Section 3. We have tested our implementation on single molecules, diatomic chains as well as one-dimensional (1D) polymers, two-dimensional (2D) sheets, and a three-dimensional (3D) crystal of diamond. Our results are discussed in Section 4, where we make comparisons to experiment and other computational methods.

2. NMR Shielding Tensor in Periodic Systems

The NMR shielding tensor $\hat{\sigma}^N$ for nucleus N is defined as the second derivative of the total electronic energy with respect to an external magnetic field $\mathbf{B}$ and a nuclear magnetic moment $\boldsymbol{\mu}_N$. For a periodic system, the NMR shielding tensor is the second derivative of the total electronic energy per unit cell:

$$\sigma^N_{\alpha\beta} = \frac{\partial^2 E(\mathbf{B}, \boldsymbol{\mu}_N)}{\partial \mu_{N\alpha} \partial B_\beta}\bigg|_{\substack{\mathbf{B}=0\\ \boldsymbol{\mu}_N=0}} \qquad (1)$$

where $\mu_{N\alpha}$ and B_β are Cartesian components of the magnetic moment $\boldsymbol{\mu}_N$ and the magnetic field $\mathbf{B}$, respectively.

We shall use Kohn−Sham density functional theory[14] (DFT) in this work. The total energy is given in DFT as a functional of the electronic density. The density, in turn, is represented as a sum of the auxiliary Kohn−Sham orbitals, Ψ_i:

$$\rho(\mathbf{k}, \mathbf{r}) = \sum_i n_i \Psi_i^*(\mathbf{k}, \mathbf{r}) \Psi_i(\mathbf{k}, \mathbf{r}) \qquad (2)$$

here n_i is the occupation numbers for the KS orbitals, and the summation is over occupied orbitals. These orbitals are obtained as the self-consistent solution to the set of equations:

$$\mathbf{H}\Psi_i = \varepsilon_i \Psi_i \qquad (3)$$

where the Hamiltonian has the form:

$$\mathbf{H} = \frac{1}{2}\mathbf{p}^2 + V^{\mathrm{KS}}, \quad V^{\mathrm{KS}} = V_{\mathrm{XC}} + V_{\mathrm{NUC}} + V_{\mathrm{C}} \qquad (4)$$

and $\mathbf{p}$ is the momentum operator ($\mathbf{p} = -i\nabla$), V^{KS} is the effective Kohn−Sham potential, which is made up of an exchange−correlation potential accounting for many-body effects V_{XC}, an attractive potential due to nuclei V_{NUC}, and a classical electron repulsion potential V_{C}. Finally, Ψ_i and ε_i are an one electron Kohn−Sham orbital and an eigenvalue to (3), respectively.

The magnetic field is introduced into the Hamiltonian using the "minimum-coupling ansatz,"[15] where a magnetic vector potential is added to the momentum operator:

$$\mathbf{p} \rightarrow \mathbf{p} + \frac{\mathbf{A}}{c} \qquad (5)$$

By making use of eqs 4 and 5, the full Hamiltonian for the system can be written in the form:

$$\mathbf{H} = \mathbf{H}^0 + \frac{1}{c}\mathbf{A}\mathbf{p} + \frac{1}{2c^2}\mathbf{A}^2, \quad \mathbf{H}^0 = \frac{1}{2}\mathbf{p}^2 + V^{\mathrm{KS}} \qquad (6)$$

where $\mathbf{A}$ is a vector potential. In NMR spectroscopy, the vector potential $\mathbf{A}$ is made up of contributions from an external magnetic field and from nuclear magnetic moments, respectively:

$$\mathbf{A} = \mathbf{A}^{(\boldsymbol{\mu}_N)} + \mathbf{A}^{(\mathbf{B})} \qquad (7)$$

where $\mathbf{A}^{(\mathbf{B})}$ is a magnetic vector potential due to a constant external magnetic field $\mathbf{B}$ that takes the form

$$\mathbf{A}^{(\mathbf{B})} = \frac{1}{2}[\mathbf{B} \times \mathbf{r}] \qquad (8)$$

whereas $\mathbf{A}^{(\boldsymbol{\mu}_N)}$ is a magnetic vector potential due to the magnetic dipoles and given by

$$\mathbf{A}^{(\boldsymbol{\mu}_N)}(\mathbf{r}) = \sum_{\mathbf{T}} \frac{[\boldsymbol{\mu}_N \times \mathbf{r}_{NT}]}{|\mathbf{r}_{NT}|^3} \qquad (9)$$

where $\mathbf{r}_{NT} = \mathbf{r} - \mathbf{R}_N - \mathbf{T}$, $\mathbf{R}_N$ is the position of a probe atom N, and $\mathbf{T}$ is the crystal vector. The infinite sum of dipole contributions is conditionally convergent in the 3D case and can be properly defined via analytic continuation techniques.[16]

Keeping in eq 6 terms containing the vector potentials to first order in $\mathbf{A}^{(\boldsymbol{\mu}_N)}$ and $\mathbf{A}^{(\mathbf{B})}$ affords:

$$\mathbf{H} \approx \mathbf{H}^0 + \frac{1}{c}\mathbf{A}^{(\boldsymbol{\mu}_N)}\mathbf{p} + \frac{1}{c}\mathbf{A}^{(\mathbf{B})}\mathbf{p} + \frac{1}{c^2}\mathbf{A}^{(\boldsymbol{\mu}_N)}\mathbf{A}^{(\mathbf{B})} \qquad (10)$$

In order to obtain solutions to eq 3 for periodic systems, the Kohn−Sham orbitals are expanded in terms of a complex Bloch basis set $\varphi_{\mu k}$:[16]

$$\Psi_i(\mathbf{k}, \mathbf{r}) = \sum_\mu c_{\mu ik} \varphi_{\mu k}(\mathbf{r}) \qquad (11)$$

where $\varphi_{\mu k}$ is calculated as a Bloch sum of equivalent atomic orbitals $\varphi_\mu(\mathbf{r} - \mathbf{R}_\mu - \mathbf{T})$ separated by the crystal vector and centered on $\mathbf{R}_\mu + \mathbf{T}$:

$$\varphi_{\mu k}(\mathbf{r}) \equiv \varphi_{\mu k}(\mathbf{r} - \mathbf{R}_\mu) = \sum_{\mathbf{T}} \varphi_\mu(\mathbf{r} - \mathbf{R}_\mu - \mathbf{T}) e^{i\mathbf{k}\mathbf{T}} \qquad (12)$$

To avoid gauge problems due to the use of a finite atomic basis set, we employ gauge-including atomic orbitals.[17]

In this case the Bloch basis set $\varphi_{\mu k}$ depends on the magnetic field $\mathbf{B}$:

$$\varphi_{\mu k}(\mathbf{r})^{\mathrm{GIAO}} = \varphi_{\mu k}(\mathbf{r}) e^{-i/2c([\mathbf{B} \times \mathbf{R}_\mu]\mathbf{r})} \qquad (13)$$

In order to calculate the second derivative of the total electronic energy, we follow the procedure due to Gauss[18] and write the expression for the energy Lagrangian with the usual orthonormality constraint:

$$\tilde{E} = \int d\mathbf{k} \sum_i n_i \langle \Psi_i(\mathbf{k},\mathbf{r})|\mathbf{h}|\Psi_i(\mathbf{k},\mathbf{r})\rangle + \frac{1}{2}\int d\mathbf{k} \sum_i n_i \langle \Psi_i(\mathbf{k},\mathbf{r})|V_\mathrm{C}|\Psi_i(\mathbf{k},\mathbf{r})\rangle + E_\mathrm{XC}[\rho] - \int d\mathbf{k} \sum_i \varepsilon_{ij} n_i (\langle \Psi_i(\mathbf{k},\mathbf{r})|\Psi_i(\mathbf{k},\mathbf{r})\rangle - \delta_{ij}) \quad (14)$$

Here the integration over **k**-space is introduced, E_XC is the exchange–correlation energy,[19] and **h** contains the sum of the operators for the electronic kinetic energy and the electron nuclear attraction plus the magnetic vector potentials:

$$\mathbf{h} = -\frac{\nabla^2}{2} + V_\mathrm{NUC} + \frac{1}{c}\mathbf{A}^{(\boldsymbol{\mu}_N)}\mathbf{p} + \frac{1}{c}\mathbf{A}^{(\mathbf{B})}\mathbf{p} + \frac{1}{c^2}\mathbf{A}^{(\boldsymbol{\mu}_N)}\mathbf{A}^{(\mathbf{B})}$$

For the second derivative we have the following expression:[18,20]

$$\frac{d^2\tilde{E}}{d\mathbf{B}d\boldsymbol{\mu}_N} = \int d\mathbf{k}\left\{\sum_{\mu\nu} P_{\mu\nu k}\frac{\partial^2 h_{\mu\nu k}}{\partial\mathbf{B}\partial\boldsymbol{\mu}_N} + \frac{\partial P_{\mu\nu k}}{\partial\boldsymbol{\mu}_N}\frac{\partial h_{\mu\nu k}}{\partial\mathbf{B}} + \frac{\partial P_{\mu\nu k}}{\partial\boldsymbol{\mu}_N} \times \left[\left\langle\frac{\partial\varphi_{\mu k}}{\partial\mathbf{B}}\middle|V_\mathrm{C} + V_\mathrm{XC}\middle|\varphi_{\nu k}\right\rangle + \left\langle\varphi_{\mu k}\middle|V_\mathrm{C} + V_\mathrm{XC}\middle|\frac{\partial\varphi_{\nu k}}{\partial\mathbf{B}}\right\rangle\right] - \frac{\partial W_{\mu\nu k}}{\partial\boldsymbol{\mu}_N}\frac{\partial S_{\mu\nu k}}{\partial\mathbf{B}}\right\} \quad (15)$$

where we use the following definitions:

$V_\mathrm{XC} = \dfrac{\partial E_\mathrm{XC}[\rho]}{\partial\rho}$ is the exchange–correlation potential

$P_{\mu\nu k} = \sum_i n_i c^*_{\mu ik} c_{\nu ik}$ is the density P matrix

$W_{\mu\nu k} = \sum_i n_i c^*_{\mu ik}\varepsilon_{ik} c_{\nu ik}$ is the energy-weighted P matrix

$S_{\mu\nu k} = \langle\varphi_{\mu k}|\varphi_{\nu k}\rangle$is the overlap matrix

$$h_{\mu\nu k} = \langle\varphi_{\mu k}|\mathbf{h}|\varphi_{\nu k}\rangle$$

The first term in eq 15 is the diamagnetic part of the shielding tensor; all other terms consist the paramagnetic shielding tensor. For the diamagnetic tensor, we have

$$\hat{\sigma}^{N,d} = \int d\mathbf{k}\sum_{\mu\nu} P_{\mu\nu k}\frac{\partial^2 h_{\mu\nu k}}{\partial\mathbf{B}\partial\boldsymbol{\mu}_N}\bigg|_{\substack{\mathbf{B}=0\\ \boldsymbol{\mu}_N=0}} \quad (16)$$

In order to evaluate $\hat{\sigma}^{N,d}$ of eq 16, use is made of the GIAO basis functions (eq 13) and the following expression for the derivatives of the GIAOs:

$$\frac{\partial\varphi^\mathrm{GIAO}_{\mu k}}{\partial\mathbf{B}} = i\left[\frac{\mathbf{r}}{2c}\times\mathbf{R}_\mu\right]\varphi^\mathrm{GIAO}_{\mu k} \quad (17)$$

By taking into account (eq 17) and employing the expression (eq 8) for $\mathbf{A}^{(\mathbf{B})}$, the diamagnetic tensor takes the final form:

$$\sigma^{N,d}_{\alpha\beta} = \frac{1}{2c}\int d\mathbf{k}\sum_i n_i \sum_{\mu\nu} c^{(0)*}_{\mu ik} c^{(0)}_{\nu ik}\left\{\left\langle\varphi_{\mu k}\middle|\left[(\mathbf{r}-\mathbf{R}_\nu)\times\frac{1}{c}\frac{\partial\mathbf{A}^{(\boldsymbol{\mu}_N)}}{\partial\mu_{N\alpha}}\right]_\beta\middle|\varphi_{\nu k}\right\rangle + \left\langle\varphi_{\mu k}\middle|[(\mathbf{r}-\mathbf{R}_\nu)\times(\mathbf{R}_\nu-\mathbf{R}_\mu)]_\beta\frac{1}{c}\frac{\partial\mathbf{A}^{(\boldsymbol{\mu}_N)}}{\partial\mu_{N\alpha}}\nabla\middle|\varphi_{\nu k}\right\rangle\right\} \quad (18)$$

where the sum (eq 9) of $\mathbf{A}^{(\boldsymbol{\mu}_N)}$ has only one atom from each cell and contains thus our probe atom N and its periodic images in other cells separated by the lattice vector **T**. From expression (eq 18), we subtract a term

$$\frac{1}{2c}\int d\mathbf{k}\sum_i n_i\sum_{\mu\nu} c^{(0)*}_{\mu ik} c^{(0)}_{\nu ik}\left\langle\varphi_{\mu k}\middle|[\mathbf{R}_\nu\times(\mathbf{R}_\nu-\mathbf{R}_\mu)]_\beta\frac{1}{c}\frac{\partial\mathbf{A}^{(\boldsymbol{\mu}_N)}}{\partial\mu_{N\alpha}}\nabla\middle|\varphi_{\nu k}\right\rangle \quad (19)$$

to make the diamagnetic part invariant with respect to a displacement of the coordinate origin.

The paramagnetic shielding tensor can be written as

$$\hat{\sigma}^{N,p} = \int d\mathbf{k}\left\{\sum_{\mu\nu}\frac{\partial P_{\mu\nu k}}{\partial\boldsymbol{\mu}_N}\frac{\partial h_{\mu\nu k}}{\partial\mathbf{B}} + \frac{\partial P_{\mu\nu k}}{\partial\boldsymbol{\mu}_N}\left[\left\langle\frac{\partial\varphi_{\mu k}}{\partial\mathbf{B}}\middle|V_\mathrm{C}+V_\mathrm{XC}\middle|\varphi_{\nu k}\right\rangle + \left\langle\varphi_{\mu k}\middle|V_\mathrm{C}+V_\mathrm{XC}\middle|\frac{\partial\varphi_{\nu k}}{\partial\mathbf{B}}\right\rangle\right] - \frac{\partial W_{\mu\nu k}}{\partial\boldsymbol{\mu}_N}\frac{\partial S_{\mu\nu k}}{\partial\mathbf{B}}\right\}\bigg|_{\substack{\mathbf{B}=0\\ \boldsymbol{\mu}_N=0}} \quad (20)$$

In order to calculate (eq 20), we use the expression:

$$\frac{\partial S_{\mu\nu k}}{\partial\mathbf{B}} = i\left\langle\phi^\mathrm{GIAO}_{\mu k}\middle|\left[\frac{\mathbf{r}}{2c}\times(\mathbf{R}_\nu-\mathbf{R}_\mu)\right]\middle|\varphi^\mathrm{GIAO}_{\nu k}\right\rangle \quad (21)$$

For the paramagnetic tensor, we apply analytic differentiation with respect to the external magnetic field components and the numerical differentiation with respect to the magnetic dipole moment components. We make use of numerical differentiation for the magnetic dipole moment to avoid[20] potential problems related to near degeneracies between occupied and virtual orbitals (for example, for 2D graphite).[21] Such problems do not occur for the external magnetic field as it is the second and responding perturbation. Thus, use can be made of analytic differentiation in this case. The final formula for the paramagnetic tensor takes the form:

$$\sigma^{N,p}_{\alpha\beta} = \frac{1}{2c}\int d\mathbf{k}\sum_i n_i\sum_{\mu\nu}(-i)\frac{\partial}{\partial\mu_{N\alpha}}(c^{(\boldsymbol{\mu}_N)*}_{\mu ik} c^{(\boldsymbol{\mu}_N)}_{\nu ik})\Big\{\langle\varphi_{\mu k}|[(\mathbf{r}-\mathbf{R}_\nu)\times\nabla]_\beta|\varphi_{\nu k}\rangle + \langle\varphi_{\mu k}|[\mathbf{r}\times(\mathbf{R}_\nu-\mathbf{R}_\mu)]_\beta(\mathbf{H}^0-\varepsilon^{(0)}_{ik})|\varphi_{\nu k}\rangle\Big\} \quad (22)$$

where $c^{(\boldsymbol{\mu}_N)}_{\mu ik}$ is the numerical solution for (eq 3) with only the perturbation by $\boldsymbol{\mu}_N$ included.

According to the recipe of Fukui,[22] we add the term (eq 19) to the paramagnetic contribution in order to make both paramagnetic and diamagnetic terms individually origin invariant.

In order to evaluate the numerical derivative

$$\frac{\partial}{\partial\mu_{N\alpha}}(c^{(\boldsymbol{\mu}_N)*}_{\mu ik} c^{(\boldsymbol{\mu}_N)}_{\nu ik}) \quad (23)$$

with respect to the magnetic dipole moment component, we make use of the fact that there are identical magnetic dipole moments $\boldsymbol{\mu}_N$ from *one* equivalent atom N (see Figure 1) in each cell.

The Hamiltonian perturbed by a periodic distribution of nuclear dipole moments $\boldsymbol{\mu}_N$ has the form

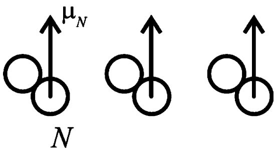


Figure 1. Magnetic dipole moments on equivalent atoms *N*.

$$\mathbf{H} = \mathbf{H}^0 + \frac{1}{c}\mathbf{A}^{(\boldsymbol{\mu}_N)}\mathbf{p} \tag{24}$$

where the vector magnetic potential from the periodic distribution of dipoles $\boldsymbol{\mu}_N$ has a form given in eq 9.

We stress again that $\mathbf{A}^{(\boldsymbol{\mu}_N)}$ is a sum made up of a single contribution from one magnetic dipole moment in each cell. The KS equation with the perturbed Hamiltonian (eq 24) are solved separately for μ_{Nx}, μ_{Ny}, and μ_{Nz}, respectively. We have found a single point numerical differentiation with displacements $\Delta\mu_{N\alpha} = 0.01$ au to be numerically stable.

Since the Bloch basis set is complex, the first-order change in density with respect to a magnetic field

$$\left.\frac{\partial\rho(\mathbf{k},\mathbf{r})}{\partial B_\beta}\right|_{B_\beta=0} = \frac{1}{2c}\sum_i n_i \sum_{\mu\nu} c_{\mu ik}^{(0)*} c_{\nu ik}^{(0)} i\varphi_{\mu k}^*[\mathbf{r}\times(\mathbf{R}_\nu - \mathbf{R}_\mu)]_\beta \varphi_{\nu k} \tag{25}$$

is not equal to zero in each **k** point. However, the first-order change in density in **k** will be canceled by the corresponding change in $-\mathbf{k}$. It can be shown based on the fact that for Bloch functions in a $-\mathbf{k}$ point we have

$$\varphi_{\mu,-k} = \varphi_{\mu k}^* \tag{26}$$

according to (eq 12). Further, $c_{\mu i,-k}^{(0)} = c_{\mu ik}^{(0)*}$ in order to keep properties of the Bloch functions (eq 11) (time-reversal symmetry).[23] Thus we have

$$\left.\frac{\partial\rho(\mathbf{k},\mathbf{r})}{\partial B_\beta}\right|_{B_\beta=0} = -\left.\frac{\partial\rho(-\mathbf{k},\mathbf{r})}{\partial B_\beta}\right|_{B_\beta=0}$$

Therefore, the calculation of NMR shielding tensors of periodic systems is based on uncoupled perturbation theory[24] as in the case of single molecules, since the total first-order changes in the density in both types of systems are zero.

In evaluating $\hat{\sigma}^{N,d}$ of eq 18 and $\hat{\sigma}^{N,p}$ of eq 22, the integration is over a single unit cell. However, the unit cell must be large enough so that the current density induced by the magnetic moment of the single probe atom *N* in that cell is practically falling off to zero at the borders of that cell. If this condition is satisfied, then it does not matter that the operator due to the magnetic field is not periodic. This is so since our assumption about the induced current density in conjunction with the use of GIAOs will ensure that integration over any unit cell to obtain $\hat{\sigma}^{N,d}$ and $\hat{\sigma}^{N,p}$ will give the same results. Thus, in order to calculate $\hat{\sigma}^{N,d}$ and $\hat{\sigma}^{N,p}$ in some cell $\mathbf{T}_0$, we need to change $\mathbf{r}\rightarrow\mathbf{r} + \mathbf{T}_0$ and $\mathbf{R}_\mu\rightarrow\mathbf{R}_\mu + \mathbf{T}_0$, and this does not change the diamagnetic tensor value since $\hat{\sigma}^{N,d}$ has only $\mathbf{r} - \mathbf{R}_\mu$ and $\mathbf{R}_\nu - \mathbf{R}_\mu$ terms. Moreover, integration in the cell $\mathbf{T}_0$ is equivalent to origin shift by $\mathbf{T}_0$, and we have already discussed the origin invariance of paramagnetic tensor.

The problem of the operator representing the interaction between the electrons and the external magnetic field is not periodic has been treated in different ways by various authors.[25-27] Following the original suggestion by Thornhauser et al.,[26] we consider the magnetic dipole as the first perturbation in what the authors have termed a converse approach, since the external magnetic field traditionally has been considered as the first perturbation. In the converse approach, one can use periodicity of the perturbing potential. However, such an approach does not completely circumvent the problem of the nonperiodic operator due to the external magnetic field.

To incorporate this aspect it is important to note that the shielding constant for nuclei *N*, as defined in eq 1, is related to the interaction energy $\Delta E_{\alpha\beta}$ between the current density ΔJ_α induced by the nuclear magnetic moment component $\mu_{N,\alpha}$ on *N* and the external magnetic field component B_β by $\Delta E_{\alpha\beta} = \sigma_{\alpha\beta}^N \mu_{N,\alpha} B_\beta$. The induced current density ΔJ_α from nuclei *N* is not periodic. However, we can assume that it vanishes outside the border of some region (supercell). Thus $\Delta E_{\alpha\beta}$ and $\sigma_{\alpha\beta}^N$ can be evaluated by integration within this supercell. It is implicit in the definition of $\sigma_{\alpha\beta}^N$ that only contributions from the current density of nucleus *N* (and not its periodic images) shall be considered in evaluating $\sigma_{\alpha\beta}^N$. Note, we can still operate with a periodic magnetic vector potential $\mathbf{A}^{(\boldsymbol{\mu}_N)}$ as long as the magnetic moment due to $\mu_{N,\alpha}$ vanishes outside the border of the supercell to which *N* belongs.

3. Computational Details

The Bloch states are expanded in a mixed basis of Slater-type and numerical atomic orbitals with the radial part of each NAO stored on a grid. Such a basis is well suited for an accurate representation of the electron density near the nuclei. Use was made of a triple-ζ basis consisting of two STOs and one NAO for each *nl* subshell (1s, 2s, 2p, etc.). This basis was augmented with two STO polarization functions. This basis is referred to as TZ2P in the BAND's basis set database. In some cases, a STO component from one or more *nl* subshells had to be removed in order to avoid linear dependencies. The numerical accuracy parameter used by BAND has been set to five. Most of the calculations are carried out with BAND's parameter *kspace* equal to five and three. Here the *kspace* parameter of the BAND program describes the number of integration points in each **k** direction in reciprocal space. For odd *kspace* values, BAND uses quadratic integration schemes for 1D, 2D,[9] and 3D[10] Brillouin zones.

In order to calculate the crystal orbitals perturbed by the nuclear dipole moments $\boldsymbol{\mu}_N$ (coefficients $c_{\mu ik}^{(\boldsymbol{\mu}_N)}$), we calculate the matrix elements

$$\left\langle \varphi_{\mu k}\left|\frac{1}{c}\mathbf{A}^{(\boldsymbol{\mu}_N)}\mathbf{p}\right|\varphi_{\nu k}\right\rangle \tag{27}$$

involving the perturbing Hamiltonian (eq 24) and add them to the corresponding matrix elements containing the unperturbed Hamiltonian. The resulting matrix is subsequently diagonalized. Only one SCF cycle is required since the magnetic perturbation is purely imaginary. Thus, no first-

Table 1. Calculated Shielding Constants (in ppm) for Molecular Water in a Cubic Super-Cell Compared to Experiment and Molecular ADF Calculations

	BAND[a]			
atom	STOs basis set	mixed NAOs/STOs basis set	ADF	experiment[b]
O	331.21	329.36	331.71	344.0
H	31.79	31.82	31.79	30.1

[a] Cubic supercell dimension $a = 30$ Å. [b] Experimental data from ref 35.

order change is induced in the density and the corresponding Coulomb and exchange−correlation potentials.

All calculations are based on spin restricted SCF calculations employing the generalized gradient approximation (GGA) for the exchange−correlation energy. The parametrization of the exchange−correlation energy follows that of Becke[28] for the exchange and Perdew[29,30] for the correlation.

Since Bloch eigenstates are only defined within a phase factor, the solutions $c_{\mu ik}^{(\boldsymbol{\mu}_N)}$ and $c_{\mu ik}^{(0)}$ may differ by a phase. In order to avoid this problem, use was made of the following expression for calculating the derivative (eq 23):

$$\frac{\partial}{\partial \boldsymbol{\mu}_N}(c_{\mu ik}^{(\boldsymbol{\mu}_N)*}c_{\nu ik}^{(\boldsymbol{\mu}_N)}) \approx \frac{c_{\mu ik}^{(\boldsymbol{\mu}_N)*}c_{\nu ik}^{(\boldsymbol{\mu}_N)} - c_{\mu ik}^{(0)*}c_{\nu ik}^{(0)}}{|\boldsymbol{\mu}_N|} \qquad (28)$$

instead of a direct numerical differentiation of the solutions $c_{\mu ik}^{(\boldsymbol{\mu}_N)}$.

Only time-reversal symmetry is used in integration over the Brillouin zone, since symmetry cannot be employed as the perturbation (eq 24) does not commute with the symmetry operations.

4. Results

We have tested our BAND implementation for the calculation of NMR shielding tensors by comparing our results with experiment, with calculations reported in the literature, and with calculations using the molecular ADF code.[31−34]

Single Molecule of Water. The simplest test system for the calculation of NMR shielding by a periodic code is a single molecule in a big box. Calculations have been carried out on one water molecule in a large cubic supercell with $a = 30$ Å, employing two different basis sets and an experimental geometry. In the first case, we have employed a pure STO TZ2P basis from the ADF database without NAOs. In the second case, use was made of a mixed NAO/STO TZ2P basis from the database of BAND, as described in Section 3, Computational Details. The results are listed in the Table 1.

It is clear from Table 1 that BAND and ADF afford quite similar results for the same TZ2P basis consisting of STOs only. Employing a mixed TZ2P STO/NAO basis introduces only a minor change in the results obtained by BAND. Thus employing a pure STO or a mixed STO/NAO TZ2P basis in BAND is likely to afford results of similar accuracy compared to experimental results.

One-Dimensional Periodic Systems. In order to test our implementation on periodic systems, we have carried out a set of calculations for 1D systems. The systems consisted of diatomic chains and polymers.

Table 2. Shielding Constants for Chains of Diatomic Molecules

			isotropic shielding constant for molecular chains, ppm		
chain	period, Å	atom probed by NMR	BAND[a]	ADF[b]	isotropic shielding constant for single molecule, ppm
H_2	2.38	H	19.59	19.53	22.17
F_2	3.00	F	−181	−195	−258.7
HCl	2.70	H	20.5	20.1	31.9
		Cl	762	775	954

[a] BAND calculations with *kspace* = 5 and 5 molecules in one supercell; TZ2P NAO/STO basis. [b] ADF calculations based on a cluster of 40 molecules; TZ2P STO basis.

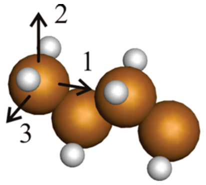


Figure 2. Polyethylene cells and principal axis of the shielding tensor.

Diatomic Chains. Calculated isotropic shielding constants for chains of H_2, F_2, and HCl molecules are listed in Table 2. All constants obtained by BAND were compared to results from ADF calculations on a cluster model consisting of 40 molecules.

We find in general for chains of diatomic molecules that a total number of five molecules are required in a unit cell to satisfy our boundary condition of diminishing induced current density at the edges. Thus, making use of only three molecules per cell changed the calculated constants by a few ppm. On the other hand, increasing the number of molecules to seven had only a marginal influence on the calculated constants (∼0.1%). The calculated shielding constants had converged with the use of five **k** points (BAND's parameter *kspace* = 5).

Polyethylene (PE) has two carbon and four hydrogen atoms in each primitive cell with a period of $T = 2.553$ Å (see Figure 2). The bond angles and lengths used in the calculation were taken from experimental X-ray data.[36,37] For the C−H bond length, we adopted a value of $r_{CH} = 1.09$ Å.[37] The results for the principal and isotropic values of the chemical shift are listed in the Table 3. Also shown are experimental findings[38] along with ADF results from a calculation on a molecular cluster, $CH_3-(CH_2)_{16}-CH_3$. The induced paramagnetic current

$$\mathbf{J}^p(\mathbf{r}) = \int d\mathbf{k} \sum_i n_i \frac{1}{2c} \mathrm{Im}(\Psi_i^{(\boldsymbol{\mu}_N)*}(\mathbf{k},\mathbf{r}) \nabla \Psi_i^{(\boldsymbol{\mu}_N)}(\mathbf{k},\mathbf{r}))$$

where $\Psi_i^{(\boldsymbol{\mu}_N)}(\mathbf{k},\mathbf{r})$ are the KS orbitals perturbed by the nuclear moment $\boldsymbol{\mu}_N$ of our probing carbon atom, and its periodic image in the other supercells is shown in Figure 3 for polyethylene. Figure 3 depicts the paramagnetic current in one supercell consisting of three primitive cells of PE. The big circles show carbon atoms, the small circles show

Table 3. Carbon Chemical Shifts (in ppm) for Polyethylene with Respect to TMS[a]

	BAND[b]	ADF cluster[c]	experiment[d]
δ_{11}	16.2	18.3	15.5
δ_{22}	36.6	48.8	33.9
δ_{33}	55.0	45.2	51.1
δ_{iso}	35.9	37.4	33.5

[a] Tetramethylsilane. [b] BAND calculations with *kspace* = 3 and 3 primitive cells as a supercell; TZ2P NAO/STO basis. [c] ADF calculations based on a $CH_3-(CH_2)_{16}-CH_3$ cluster, where the calculated shift corresponds to one of the two central carbons; TZ2P STO basis. [d] From ref 38.

hydrogen atoms. The intensity of color is proportional to the absolute value of the current. The paramagnetic current is located mostly on carbon atoms. It is almost zero on the hydrogens. It is clear from Figure 3 that using one primitive cell to calculate the paramagnetic shielding tensor is not enough and that we need to take into consideration the nearest cells. To reach convergence for the shielding tensor, we need to take three primitive cells as one big supercell and make integration in **k**-space with three **k**-points (*kspace* = 3).

The chemical shift is calculated with respect to tetramethylsilane (TMS) as $\delta^{^{13}C} = \sigma_{TMS}^{^{13}C} - \sigma_{PE}^{^{13}C}$, where the TMS isotropic shielding tensor is calculated by the ADF program. For the ADF cluster calculation on $CH_3-(CH_2)_8-C^*H_2-(CH_2)_7-CH_3$, we obtained a value of 37.4 ppm compared to the BAND result of 35.9 ppm and the experimental value of 33.5 ppm. Thus there seems to be good agreement between experiment and theory. It should, however, be pointed out that the experimental value corresponds to PE folded in 3D. Nevertheless the comparison is still valid since 3D PE "locally" can be considered linear.

***Trans*-polyacethylene** (PA) has two atoms of carbon and two atoms of hydrogen in a primitive cell with a period of $T = 2.457$ Å (Figure 4). The bond angles and lengths used in the calculation were taken from experimental X-ray data.[39] The result for the isotropic value of the chemical shift is listed in the Table 4. Also shown are experimental findings along with ADF results from a calculation on a molecular cluster, $H-(CH)_{80}-H$. BAND result is 142.3 ppm, convergence is reached with a *kspace* parameter equal to 5 (total number of **k**-points equal to 5) and 5 primitive cells as a supercell. For the ADF cluster, we have obtained a value for the shielding constant of 134.6 ppm. The experimental estimate is 137.3 ppm.

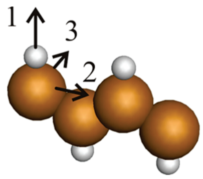


Figure 4. *Trans*-polyacethylene cells and principal axis of the shielding tensor.

Table 4. Carbon Chemical Shift (in ppm) for *Trans*-PA with Respect to TMS

	BAND[a]	ADF cluster[b]	experiment[c]
δ_{11}	221.0	218.4	219
δ_{22}	155.1	140.1	144
δ_{33}	50.7	45.5	47
δ_{iso}	142.3	134.6	137.3

[a] BAND calculations with *kspace* = 5 and 5 primitive cells as a supercell; TZ2P NAO/STO basis. [b] ADF calculations based on a $H-(CH)_{80}-H$ cluster with the central carbon as the NMR probe; TZ2P STO basis. [c] From ref 40.

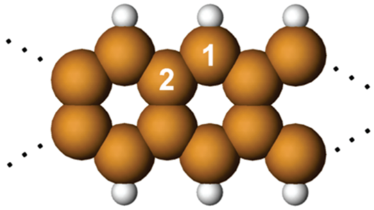


Figure 5. Carbon nanoribbon.

Carbon nanoribbon is a 1D polymer with four atoms of carbon and two atoms of hydrogen in a primitive cell (Figure 5). The C−C bond length used in the calculation is $r_{CC} = 1.418$ Å. The result for the isotropic value of the chemical shifts is listed in the Table 5. Also shown are calculated results by other authors along with ADF results from a calculation on a molecular cluster consisting of 44 benzene rings. BAND result is 138.4 and 147.1 ppm for two types of carbon atoms (see Figure 5). Convergence is reached with *kspace* = 5 (total number of 5 **k**-points) and 5 primitive cells as a supercell. For the ADF cluster, we have obtained values

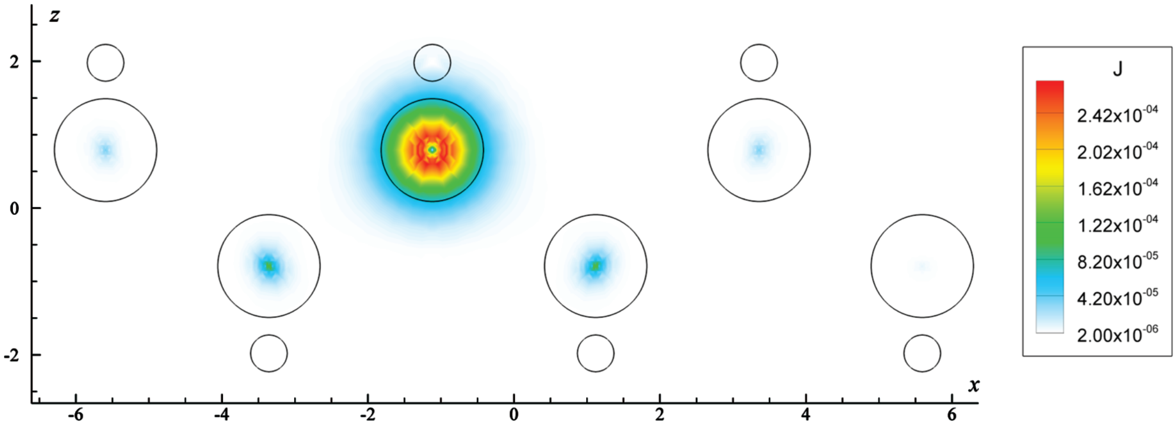


Figure 3. The paramagnetic current for polyethylene; *x* and *z* coordinates and absolute value of the current are in au.

Table 5. Carbon Chemical Shift (in ppm) for Carbon Nanoribbon with Respect to TMS

	atom probed by NMR	BAND[a]	ADF cluster[b]	calculated by other authors
carbon	C_1	138.4	131.9	128.0[c]
nanoribbon	C_2	147.1	137.1	132.2[c]

[a] BAND calculations with *kspace* = 5 and 5 primitive cells as a supercell; TZ2P NAO/STO basis. [b] ADF calculations based on a $H_2{-}(C_4H_2)_{43}{-}C_2H_2$ cluster with the central carbon as the NMR probe; TZ2P STO basis. [c] From ref 5.

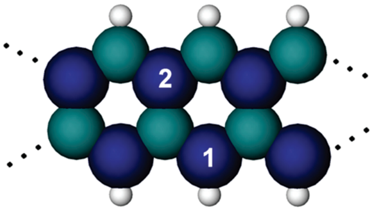


Figure 6. Boron nitride nanoribbon.

Table 6. Nitrogen Chemical Shift (in ppm) for a Boron Nitride Nanoribbon with Respect to $CH_3{-}NO_2$

	atom probed by NMR	BAND[a]	ADF cluster[b]
BN nanoribbon	N_1	−261.1	−270.9
	N_2	−220.6	−228.8

[a] BAND calculations with *kspace* = 3 and 5 primitive cells as a supercell; TZ2P NAO/STO basis. [b] ADF calculations based on a $H_2{-}(B_2N_2H_2)_{44}{-}BN{-}H_2$ cluster with the central carbon as the NMR probe; TZ2P STO basis.

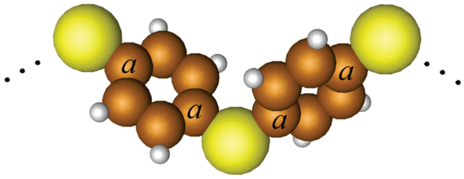


Figure 7. Poly(*p*-phenylene sulfide) cell.

Table 7. ^{13}C Chemical Shift (in ppm) for PPS with Respect to TMS

	atom probed by NMR	BAND[a]	experiment[b]
PPS	C_a	134.8	135.1
	C_b	131.1	131.8

[a] BAND calculations with *kspace* = 3 and one primitive cell; TZ2P NAO/STO basis. [b] From refs 43 and 44.

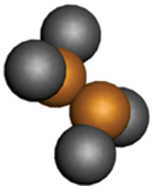


Figure 8. Teflon cell.

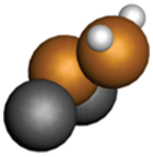


Figure 9. PVDF cell.

Table 8. ^{19}F Chemical Shifts (in ppm) for PVDF[a] and Teflon[b]

polymer	atom probed by NMR	BAND	ADF cluster	experiment
PVDF	F	−135.0[c]	−133.9[d]	91.6, 94.8, 113.6, 115.6[f]
teflon	F	−585.8[c]	−588.1[e]	−549[g]

[a] With respect to $CFCl_3$. [b] With respect to F_2. [c] BAND calculations with *kspace* = 3 and 3 primitive cells as a supercell; TZ2P NAO/STO basis. [d] ADF calculations based on a $CH_3{-}(CF_2{-}CH_2)_{19}{-}H$ cluster; TZ2P STO basis. [e] ADF calculations based on a $F{-}(CF_2)_{28}{-}F$ cluster; TZ2P STO basis. [f] From ref 47. [g] From ref 48.

for the shielding constants of 131.9 and 137.1 ppm, respectively. The result by Thonhauser et al. is 128.0 and 132.2 ppm.

Boron nitride nanoribbon is a 1D polymer with two atoms each of boron, nitrogen, and hydrogen in a primitive cell (Figure 6). The B−N bond length used in the calculation is r_{BN} = 1.446 Å. The results for the nitrogen shielding shifts with respect to nitromethane are listed in the Table 6. Also shown are ADF results from a calculation on a molecular cluster consisting of 44 BN rings. BAND result is −261.1 ppm for the first atom and −220.6 ppm for the second. Convergence is reached with *kspace* = 3 (3 **k**-points) and 5 primitive cells as a supercell. For the ADF cluster, we have obtained values for the shielding constants of −270.9 and −228.8 ppm, respectively.

Poly(*p*-phenylene sulfide) (PPS) cell consist of two atoms of sulfur, twelve atoms of carbon, and eight atoms of hydrogen (Figure 7) with period of 10.26 Å. The bond angles and lengths used in the calculation were taken from experimental data.[41,42] There are two types of carbon atoms due to symmetry: type *a* (marked in Figure 7) and *b* (all other carbon atoms). The result for the isotropic value of the chemical shifts for two types of atoms is listed in the Table 7. Calculations have shown that to reach convergence for such a big system one primitive cell is enough. The results calculated in BAND are very closed to the experimental values.

Polytetrafluoroethylene (PTFE). The PTFE (Teflon) polymer cell holds two carbons and four atoms of fluorine (Figure 8). All geometrical parameters used in the calculations were taken from experimental X-ray data.[45] The calculated isotropic values for the ^{19}F chemical shifts with respect to F_2 are listed in Table 8. Also displayed are ADF results from a calculation on a molecular cluster, $F{-}(CF_2)_{28}{-}F$ as well as experimental findings. The BAND calculations afford a ^{19}F isotropic shift of −585.8 ppm. Convergence was reached with three **k**-points and three primitive cells in a supercell. For the ADF cluster calculation, we have obtained a ^{19}F shielding constant of −588.1 ppm, and the experimental value is −549 ppm. It seems that DFT in this case falls somewhat short of experiment.

Poly(vinylidene fluoride). The PVDF polymer has two atoms each of carbon, hydrogen, and fluorine in one unit cell (Figure 9). All the structural data used in the calculations were based on experimental X-ray data.[46] We list in Table 8 the value of the ^{19}F isotropic chemical shifts with respect

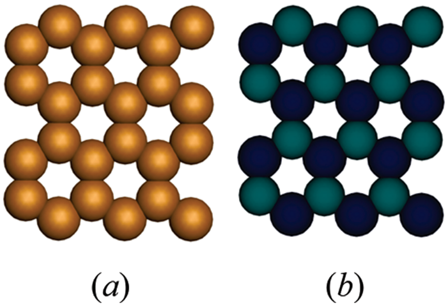


Figure 10. Twenty-four atom square supercell for graphite (*a*) and boron nitride (*b*) sheets.

Table 9. ^{13}C Chemical Shift (in ppm) for 2D Graphite with Respect to TMS

	atom probed by NMR	BAND	ADF cluster	calculated by other authors	experiment
2D graphite	C	127.1[a]	119.2[b]	118.0[c]	155,179[d]

[a] BAND calculations with *kspace* = 5 and the supercell with 24 atoms; TZ2P NAO/STO basis. [b] ADF calculations based on a $C_{188}H_{38}$ cluster; TZ2P STO basis. [c] From ref 5. [d] From ref 49.

to $CFCl_3$ from BAND calculations. Comparisons are further given with experimental findings and results from ADF calculations on the molecular cluster $H-(CH_2-CF_2)_9-CH_3$. For BAND, we reach a converged value of −135.0 ppm with 3 **k**-points and 3 primitive units in a supercell. For the ADF cluster calculation, we have obtained a similar value of −133.9 ppm. The experimental estimates contain four different peaks due to structural defects and ranges from 91.6 to 115.6 ppm.[47]

Two- and Three-Dimensional Periodic Systems. Planar 2D graphite has a hexagonal lattice with two atoms in a primitive cell (Figure 10). The C−C bond length is equal to 1.418 Å. We compile the isotropic ^{13}C chemical shift values with respect to TMS from BAND in Table 9. In the same table are given experimental findings as well as ADF results from a calculation on a molecular cluster consisting of 75 benzene rings. The converged BAND result is 127.1 ppm. It was obtained from a square supercell with 24 atoms and *kspace* = 5 (total number 45 of **k**-points). For the ADF cluster calculation, we have determined a shielding constant of 119.2 ppm. The experimental values range from 155 to 179 ppm.[49] It is possible that both theoretical models fall short of the experimental value because DFT at the GGA level used here is unable to describe dispersion. Also we do not consider the Knight shift or the semimetallic behavior exhibited by graphite at low temperatures. Nevertheless, we include our graphite results in order to compare with other implementations where use has been made of the same approximations as here.

Planar 2D boron nitride has a hexagonal lattice with two atoms in a primitive cell with B−N bond length is equal to 1.446 Å. We compile the isotropic ^{15}N chemical shift values with respect to nitromethane from BAND in Table 10. In the same table are given experimental findings as well as ADF results from a calculation on a molecular cluster consisting of 75 BN rings. The converged BAND result is −272.5 ppm. It was obtained from a square supercell with 24 atoms and *kspace* = 3 (total number 15 of **k**-points). For the ADF cluster calculation, we have determined a shielding constant of −264.4 ppm. The experimental value for hexagonal boron nitride powder is −285 ppm.

Table 10. ^{15}N Chemical Shift (in ppm) for 2D Boron Nitride with Respect to CH_3-NO_2

	atom probed by NMR	BAND	ADF cluster	calculated by other authors	experiment
2D BN	N	−272.5[a]	−264.4[b]	−287.0[c]	−285[d]

[a] BAND calculations with *kspace* = 3 and the supercell with 24 atoms; TZ2P NAO/STO basis. [b] ADF calculations based on a $B_{94}N_{94}H_{38}$ cluster; TZ2P STO basis. [c] From ref 50 (cluster consisting of 22 atoms). [d] From ref 50.

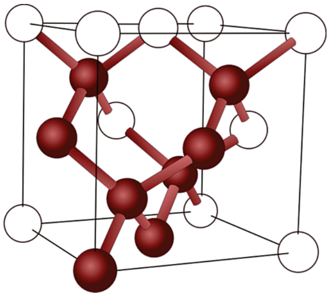


Figure 11. Cubic supercell of diamond with eight atoms.

Table 11. Chemical Shift (in ppm) for Diamond with Respect to TMS

	atom probed by NMR	BAND	calculated by other authors	experiment
diamond	C	35.8[a]	49.6[b] 36.17[d]	34.54[c] 35.7−38.3[e]

[a] BAND calculations with *kspace* = 5 and the cubic supercell with 8 atoms; TZ2P NAO/STO basis. [b] From ref 4. [c] From ref 51. [d] From ref 52. [e] From ref 53.

3D crystal of diamond has lattice parameter 3.567 Å. Convergence of the shielding tensor is reached with a 3D **k**-space mesh consisting of 123 **k**-points (*kspace* parameter of BAND is equal to 5) and a cubic supercell of 8 atoms (Figure 11). The value calculated by BAND for the ^{13}C isotropic chemical shifts with respect to TMS is listed in Table 11. Also shown are results by other authors and the experimental values. There is in general a good agreement between experiment and theory.

5. Conclusion

We have developed a Kohn−Sham density functional theory (DFT)-based approach for the calculation of NMR shielding tensors in periodic systems. This implementation is gauge-origin invariant. Our implementation differs from others in employing Slater-type or/and numerical atomic orbitals with use of the complete Kohn−Sham potential without recourse to effective potentials. We can thus describe even core orbitals variationally. Integration in the reciprocal space is carried out in one-half of the Brillouin zone. The calculation of NMR chemical shifts for single molecules as well as one-, two- and three-dimensional periodic systems has been used to validate our implementation. Our calculated results agree in most cases with experiment and with results from cluster models and other methods. In our converse procedure, the calculation of the shielding tensor $\sigma_{\alpha\beta}^{N}$ for nuclei N requires, first, the evaluation of the current density ΔJ_α induced by the three components ($\alpha = 1, 3$) of the nuclear magnetic moment $\boldsymbol{\mu}_N$, followed by the response of ΔJ_β to the three components ($\beta = 1, 3$) of the external magnetic field **B**. When use is made of functionals without current density dependence, the work needed to evaluate $\sigma_{\alpha\beta}^{N}$ by the converse method is exactly the same as in traditional approaches, where the order of the perturbations has been reversed. This is so since no change in density is induced by either of the magnetic perturbations **B** or $\boldsymbol{\mu}_N$. Thus no iterative set of coupled equations is required to be solved, and a direct expression for $\sigma_{\alpha\beta}^{N}$ can be given that does not depend on the order in which the perturbations are applied. The evaluation of ΔJ_α by a noniterative finite difference procedure rather than analytical differentiation might add some cost. We do not yet have sufficient data to assess the relative merits of periodic NMR calculations compared to those of cluster approaches. However, the use of similar basis sets in both techniques will ultimately allow us to make a valid comparison. For systems with large band gaps, one can employ small supercells, while for systems with a small or vanishing band gap, it is necessary to make use of much larger supercells. The full variational approach taken here lends itself readily to calculation on heavier nuclei, and this will be the subject of a forthcoming investigation.

Acknowledgment. T.Z. would like to thank the Canadian government for a Canada Research Chair in Theoretical Inorganic Chemistry and the NSERC for financial support. We would like to thank Dr. Pier Philipsen for technical help and Dr. Georg Schreckenbach for discussions.

References

(1) Zurek, E.; Autschbach, J. *Int. J. Quantum Chem.* **2009**, *109*, 3343–3367.

(2) Mauri, F.; Pfrommer, B. G.; Louie, S. G. *Phys. Rev. Let.* **1996**, *77*, 5300–5303.

(3) Pickard, C. J.; Mauri, F. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2001**, *63*, 245101–245113.

(4) Thonhauser, T.; Ceresoli, D.; Mostofi, A. A.; Marzari, N.; Resta, R.; Vanderbilt, D. *J. Chem. Phys.* **2009**, *131*, 101101; arxiv.org:0709.4429v2.

(5) Thonhauser, T.; Ceresoli, D.; Marzari, N. *Int. J. Quantum Chem.* **2009**, *109*, 3336–3342.

(6) Sebastiani, D.; Parinello, M. *J. Phys. Chem. A* **2001**, *105*, 1951–1958.

(7) Keith, T. A.; Bader, R. F. W. *Chem. Phys. Let.* **1993**, *210* (1−3), 223–231.

(8) te Velde, G.; Baerends, E. J. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1991**, *44*, 7888–7903.

(9) Wiesenekker, G.; te Velde, G.; Baerends, E. J. *J. Phys. C: Solid State Phys* **1988**, *21*, 4263–4283.

(10) Wiesenekker, G.; Baerends, E. J. *J. Phys.: Condens. Matter* **1991**, *3*, 6721–6742.

(11) te Velde, G.; Baerends, E. J.; Philipsen, P. H. T.; Wiesenekker, G.; Groeneveld, J. A.; Berger, J. A.; de Boeij, P. L.; Klooster, R.; Kootstra, F.; Romaniello, P.; Snijders, J. G.; Kadantsev, E. S.; Ziegler, T. *BAND*, 2009.01, *SCM: Theoretical Chemistry*, Vrije Universiteit: Amsterdam, The Netherlands; http://www.scm.com/.

(12) Kadantsev, E. S.; Ziegler, T. *J. Phys. Chem. A* **2008**, *112*, 4521–4526.

(13) Kadantsev, E. S.; Ziegler, T. *J. Phys. Chem. A* **2009**, *113*, 1327–1334.

(14) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133–A1138.

(15) McWeeny, R. *Methods of molecular quantum mechanics*; Academic Press: San Diego, 1992.

(16) te Velde, G. PhD Thesis. Vrije Universiteit: Amsterdam, The Netherlands, 1990; (available on-line: http://www.scm.com/Doc/BAND_thesis/BAND_Thesis.pdf).

(17) Ditchfield, R. *Mol. Phys.* **1974**, *27* (4), 789–807.

(18) Gauss, J. Molecular properties. In *Modern methods and algorithms of quantum chemistry, Proceedings*; Grotendorst, J., Ed.; John von Neumann Institute for Computing, Jülich, NIC series, 2000, *3*, 541−592.

(19) Jones, R. O. Introduction to density functional theory and exchange-correlation energy functionals. In *Computational Nanoscience: Do It Yourself!*; Grotendorst, J., Blügel, S., Marx, D., Eds.; John von Neumann Institute for Computing, Jülich, Germany, 2006, *31*, 45−70.

(20) Pople, J. A.; Krishnan, R.; Schlegel, H. B.; Binkley, J. S. *Int. J. Quantum Chem., Quantum Chem. Symp.* **1971**, *13*, 225–241.

(21) Kaxiras, E. *Atomic and electronic structure of solids*. Cambridge University Press: New York, NY, 2003.

(22) Fukui, H. *Magn. Reson. Rev.* **1987**, *11*, 205–274.

(23) Martin, R. M. *Electronic structure. Basic theory and practical methods*; Cambridge University Press: Cambridge, England, 2004.

(24) Amos, A. T.; Musher, J. I. *Mol. Phys.* **1967**, *13*, 509–515.

(25) Resta, R.; Ceresoli, D.; Thonhauser, T.; Vanderbilt, D. *ChemPhysChem* **2005**, *6*, 1815–1819. Thonhauser, T.; Ceresoli, D.; Vanderbilt, D.; Resta, R. *Phys. Rev. Lett.* **2005**, *95*, 137205–137214. Ceresoli, D.; Thonhauser, T.; Vanderbilt, D.; Resta, R. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2006**, *74*, 024408–024413. Xiao, D.; Shi, J.; Niu, Q. *Phys. Rev. Lett.* **2005**, *95*, 137204–137214. Shi, J.; Vignale, G.; Xiao, D.; Niu, Q. *Phys. Rev. Lett.* **2007**, *99*, 197202−197206.

(26) Thonhauser, T.; Mostofi, A. A.; Marzari, N.; Resta, R.; Vanderbilt, D. arxiv.org:0709.4429v1.

(27) Mauri, F.; Louie, S. G. *Phys. Rev. Lett.* **1996**, *76*, 4246–4249.

(28) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098–3100.

(29) Perdew, J. P. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *33*, 8822–8824.

(30) Perdew, J. P. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *34*, 7406.

(31) Baerends, E. J.; Autschbach, J.; Bashford, D.; Bérces, A.; Bickelhaupt, F. M.; Bo, C.; Boerrigter, P. M.; Cavallo, L.; Chong, D. P.; Deng, L.; Dickson, R. M.; Ellis, D. E.; van Faassen, M.; Fan, L.; Fischer, T. H.; Fonseca Guerra, C.; Ghysels, A.; Giammona, A.; van Gisbergen, S. J. A.; Götz, A. W.; Groeneveld, J. A.; Gritsenko, O. V.; Grüning, M.; Harris, F. E.; van den Hoek, P.; Jacob, C. R.; Jacobsen, H.; Jensen, L.; van Kessel, G.; Kootstra, F.; Krykunov, M. V.; van Lenthe, E.; McCormack, D. A.; Michalak, A.; Mitoraj, M.; Neugebauer, J.; Nicu, V. P.; Noodleman, L.; Osinga, V. P.; Patchkovskii, S.; Philipsen, P. H. T.; Post, D.; Pye, C. C.; Ravenek, W.; Rodríguez, J. I.; Ros, P.; Schipper, P. R. T.; Schreckenbach, G.; Seth, M.; Snijders, J. G.; Solà, M.; Swart, M.; Swerhone, D.; te Velde, G.; Vernooijs, P.; Versluis, L.; Visscher, L.; Visser, O.; Wang, F.; Wesolowski, T. A.; van Wezenbeek, E. M.; Wiesenekker, G.; Wolff, S. K.; Woo, T. K.; Yakovlev, A. L.; Ziegler, T. *ADF* , 2009.01; SCM: Theoretical Chemistry; Vrije Universiteit: Amsterdam, The Netherlands, 2009.

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

(33) Schreckenbach, G.; Ziegler, T. *J. Phys. Chem.* **1995**, *99*, 606–611.

(34) Schreckenbach, G.; Ziegler, T. *Int. J. Quantum Chem.* **1996**, *60*, 753–766.

(35) Computational chemistry comparison and benchmark database; National Institute of Standards and Technology: Gaithersburg, MD; http://cccbdb.nist.gov/. Accessed 2005.

(36) Caminiti, R.; Pandolfi, L.; Ballirano, P. *J. Macromol. Sci.* **2000**, *B39* (4), 481–492.

(37) Yamanobe, T.; Sorita, T.; Comoto, T.; Ando, I. *J. Mol. Struct.* **1985**, *131*, 267–275.

(38) Yamanobe, T. Structure and dynamics of crystalline and noncrystalline phases in polymers. In *Solid State NMR of Polymers*; Ando, I., Ed.; Elsevier & Technology Books: Amsterdam, The Netherlands, 1998, pp 267−1306.

(39) Perego, G.; Luglia, G.; Pedretti, U.; Cesari, M. *Makromol. Chem.* **1988**, *189*, 2657–2669.

(40) Terao, T.; Maeda, S.; Yamabe, T.; Akagi, K.; Shirakawa, H. *Chem. Phys. Let.* **1984**, *103* (5), 347–351.

(41) Tabor, B. J.; Magre, E. P.; Boon, J. *Eur. Polym. J.* **1971**, *7*, 1127–1133.

(42) Napolitano, R.; Pirozzi, B.; Salvione, A. *Macromolecules* **1999**, *32*, 7682–7687.

(43) Wade, B.; Abhiraman, A. S.; Wharry, S.; Sutherlin, D. *J. Polym. Sci.* **1990**, *B28*, 1233–1249.

(44) Lowman, D. W.; Fagerburg, D. R. *Bull. Magn. Reson.* **1992**, *14* (1−4), 148–152.

(45) Iwasaki, M. *J. Polym. Sci., Part A: Polym. Chem.* **2003**, *1* (4), 1099–1104.

(46) Hasegawa, R.; Takanashi, Y.; Chatani, Y.; Tadokoro, H. *Polym. J. (Tokyo)* **1972**, *3* (5), 600–610.

(47) Tonelli, A. E.; Schilling, F. C.; Cais, R. E. *Macromolecules* **1982**, *15*, 849–853.

(48) Gabuda, S. P.; Kozlova, S. G.; Paasonen, V. M.; Nazarov, A. S. *J. Struct. Chem.* **2000**, *41* (1), 67–71.

(49) Sagunama, M.; Mizutami, U.; Kondow, T. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1980**, *22*, 5079–5084.

(50) Marian, C. M.; Gastreich, M. *Solid State Nucl. Magn. Reson.* **2001**, *19*, 29–44.

(51) Merwin, L. H.; Johnson, C. E.; Weimer, W. A. *J. Mater. Res.* **1994**, *9* (3), 631–635.

(52) Mauri, F.; Pfrommer, B. G.; Louie, S. G. *Phys. Rev. Lett.* **1997**, *79* (12), 2340–2343.

(53) Alam, T. M. *Mater. Chem. Phys.* **2004**, *85*, 310–315.

CT100046A

JCTC Journal of Chemical Theory and Computation

On the Calculation of Vibrational Frequencies for Molecules in Solution Beyond the Harmonic Approximation

Chiara Cappelli,*,† Susanna Monti,‡ Giovanni Scalmani,¶ and Vincenzo Barone§

Dipartimento di Chimica e Chimica Industriale, Università di Pisa, Via Risorgimento 35, I-56126 Pisa, Italy, Istituto per i Processi Chimico-Fisici del Consiglio Nazionale delle Ricerche, via Moruzzi, 1 I-56124 Pisa, Italy, Gaussian, Inc., 340 Quinnipiac Street, Building 40, Wallingford, Connecticut 06492, and Scuola Normale Superiore, Piazza dei Cavalieri, 7 I-56126 Pisa, Italy

Received January 25, 2010

Abstract: We report some results on the calculation of vibrational spectra of molecules in condensed phase with accounting simultaneously for anharmonicity and solute−solvent interactions, the latter being described by means of the polarizable continuum model (PCM). Density functional theory force fields are employed as well as a new implementation of the PCM cavity and its derivatives. The results obtained for formaldehyde and simple peptide prototypes show that our approach is able to yield a quantitative agreement with experiments for vacuo-to-solvent harmonic and anharmonic frequency shifts.

Introduction

Infrared (IR) and Raman spectroscopies are among the most powerful techniques for characterizing medium-size molecules, but proper assignment of spectra is often not straightforward especially for unstable species or nonstandard bonding situations. In the last years, development of nonlinear techniques (e.g., two-dimensional IR, 2D-IR)[1] has allowed the direct examination not only of vibrational frequencies but also of the specific anharmonicities (both diagonal and off-diagonal) of vibrational modes.[2] This additional information demands a more quantitative interpretation of the different contributions determining the overall vibrational spectrum.

Thanks to the progresses in hardware and software, the a priori prediction of accurate low-lying vibrational levels of semirigid polyatomic molecules by means of quantum-mechanical (QM) methodologies is becoming viable. It is now widely recognized that the computation of semidiagonal quartic force fields at the coupled clusters with single, double, and perturbative inclusion of triple excitations[3] [CCSD(T)] level in conjunction with sufficiently large basis sets (at least of triple-ζ quality for second-row atoms) followed by an effective second-order perturbative treatment (PT2) usually provides results with an accuracy of the order of 10−15 cm^{-1} for fundamental transitions.[4–6] Although the perturbative vibrational treatment remains highly cost-effective for quite large systems, the unfavorable scaling of the CCSD(T) model with the number of active electrons limits the determination of quartic force fields to molecules containing at most five to six atoms. Additionally, a simple reduction of computational cost by combining correlated QM methods with a small basis set should not be recommended, due to the quite unpredictable accuracy of the results. Thus, extension of computational studies to larger systems requires cheaper, yet reliable, electronic structure approaches.

Recently, several authors have reported anharmonic force fields for small- and medium-sized semirigid molecules computed by methods rooted in the density functional theory (DFT).[7–10] Among the functionals tested, the so-called hybrid functionals provide satisfactory results when used with a basis set of at least double-ζ plus polarization quality supplemented by diffuse sp functions. An even more effective

* Corresponding author. Telephone: +39 050 2219248. Fax: +39 0502219260. E-mail: chiara@dcci.unipi.it.

† Università di Pisa.

‡ Istituto per i Processi Chimico-Fisici del Consiglio Nazionale delle Ricerche.

¶ Gaussian, Inc.

§ Scuola Normale Superiore.

10.1021/ct100048g © 2010 American Chemical Society
Published on Web 04/01/2010

approach in terms of good accuracy, obtained at a computationally reduced cost, is based on the additivity of DFT anharmonic corrections to CCSD(T) harmonic force fields. This is well-known to further improve the agreement with experimental data.[11,12]

The next step involves the consideration of environmental effects because most of the experimental determinations (and of the biologically and technologically interesting processes) are performed in the condensed phase. In this framework, continuum solvation methods[13,14] are particularly attractive due to their reliability coupled to computational costs fully comparable with those of the corresponding computations in the gas phase. We will use in the following the polarizable continuum model (PCM),[15] in view of its physical soundness coupled with effective implementations for several QM models.

Going back to the evaluation of anharmonic force fields, the effectiveness of the PT2 vibrational approach in the gas phase is related to the availability of reliable and relatively fast procedures for computing analytical second derivatives of the energy, with respect to the atomic coordinates and the extracting normal modes. By using a finite difference approach, it is relatively easy to differentiate once more along the direction of the normal modes to obtain all of the third and a subset of the fourth derivatives (the ones with no more than three distinct indexes), which are required in the PT2 model. The key issue for an effective extension to the condensed phase is related to the handling of the cavity containing the solute molecule, which is closely related to the so-called molecular surface. When the derivatives of the energy in solution, with respect to the atomic positions, need to be computed, the molecular surface must be a continuous and smooth function of the same atomic positions. The importance of this issue has been recognized in recent years.[16–18] Recently, a robust and reliable method fulfilling these characteristics, originally proposed in ref 19, has been extended to second derivatives, and the corresponding fully analytical expression for the second derivatives of the PCM contribution to the energy has been derived and is now available in the Gaussian09 (G09) suite of programs.[20]

On these grounds, we report in the present work some results related to one of the most challenging issues in the calculation of reliable vibrational spectra in condensed phase, namely, the simultaneous inclusion of anharmonicity and solute−solvent interactions. In view of their importance, most of the results refer to amides and peptides, but the computational approach and the general trends are not specific to given molecular structures and/or solvents, thus, rather providing a first exploration of a much wider topic.

Finally, it is worth remarking that to date QM calculations of anharmonicities that explicitly account for solvent effects have received only little attention in the literature. To the best of our knowledge, the preceding contributions to this matter are very limited,[21–27] all resorting to fitting of the potential energy surface (PES) and subsequent calculation of anharmonic vibrational levels.

Theoretical Background and Computational Details. In order to calculate anharmonic vibrational frequencies including solvent effects accurately, the PCM,[14] and in particular, the integral equation formalism (IEF) version of PCM,[28–30] has been used in the calculations reported in this paper. It is well-known that standard techniques can be used to compute analytical second derivatives of the energy with respect to the atomic coordinates[31,32] and to extract the normal modes when dealing with systems in the gas phase. Indeed, by means of a finite difference approach, it is relatively easy to differentiate once more along the direction of the normal modes to obtain all of the third and a subset of the fourth derivatives (the ones with no more than three distinct indexes),[10,33] which are enough to provide an accurate treatment of the anharmonicity.

Indeed, in order to apply the approach described in the previous paragraph to calculations in solution, some issues must be addressed. First, the PCM requires a molecular cavity to be defined in the dielectric continuum to host the solute, typically using a set of interlocking spheres centered at the positions of the atoms (vide infra). The surface of this cavity needs to be discretized into finite elements (historically called *tesserae*) so that the surface integrals required by the solvation model can be effectively calculated. Once the PCM equations are solved, each surface element is assigned a portion of the apparent surface charge (ASC) that represents the solvent polarization due to the presence of the solute. This ASC is typically expressed in terms of a collection of point charges located at representative points of the surface elements. When the derivatives of the energy in solution, with respect to the atomic positions, need to be computed, the issue to be addressed is whether the definition and discretization of the molecular surface is a continuous and smooth function of the same atomic positions.

In recent years, the importance of this issue has been recognized[16,17] mainly because smooth energy derivatives are needed in the study of solvent effects on the equilibrium structure of molecules. The simplest approach to address the problem of continuity and discretization of the cavity is to focus primarily on the geometrical details of how the surface elements are generated and how the regions of intersection of the spheres are handled, while the fact that the ASC is apportioned in point charges is usually considered a problem of minor importance. In the late 90s, a discretization scheme able to provide a smooth partition of both the molecular surface and the ASC was proposed by York and Karplus (YK).[19] According to the YK approach, the generation of the surface area elements is smooth because elements from one sphere can penetrate somewhat into nearby spheres, while their surface area is reduced using a smooth switching function. Clearly, such a method would be impossible to apply as long as the ASC is partitioned in point charges as there is no guarantee that two surface elements from two different spheres will not be superimposed in the intersection region of the two spheres. The natural solution to this problem is to drop the use of point charges in favor of a continuous description of the ASC using a set of charges each described by a small three-dimensional Gaussian function. In fact, two (or more) charges represented by Gaussian functions can be exactly superimposed, and their interaction energy does not diverge, as it would do if point charges were used. The obvious drawback of the YK scheme

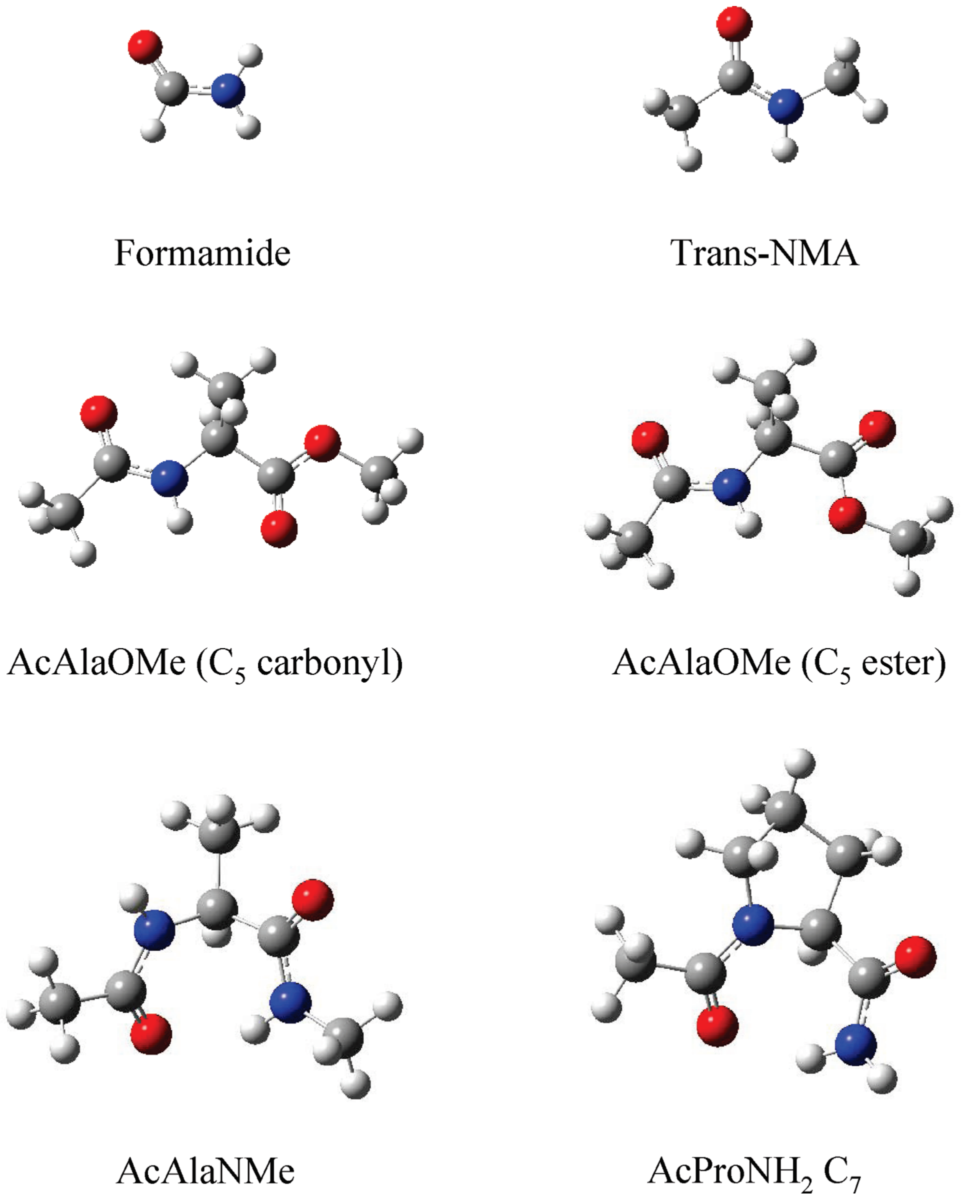


Figure 1. Structures of mono- and dipeptides studied in the present paper.

is that every interaction among the ASC charges and between the ASC charges and the solute becomes an integral over all space.

The second issue to address in order to be able to compute anharmonic vibrational frequencies in solution is that the YK scheme, described in the previous paragraph, must be generalized to second derivatives and applied to the calculation of derivatives with respect to atomic position of the PCM contribution to the energy. The formalism of the derivatives of the PCM contribution to the energy and the terms arising from the second derivatives of the surface discretization has been already reported.[34–36] However, they have never been implemented free from any approximation.

In recent years, the YK methodology has been extended to second derivatives, and the corresponding fully analytical expression for the second derivatives of the PCM contribution to the energy has been derived and implemented in the code and is now available in the G09 suite of programs.[20] The complete continuous surface charge formalism of PCM within the YK discretization scheme and all the implementation details are beyond the scope of this paper.[37] Here we just want to underline the fact that this state-of-the-art implementation fulfills all the requirements needed to carry out further numerical differentiation of the energy in solution along the normal modes and, thus, to compute solvent effects on anharmonic normal modes reliably.

Formaldehyde and some simple mono- and dipeptide prototypes (see Figure 1) were chosen for comparison with a previous study by Wang and Hochstrasser[38] on the calculation of amide modes anharmonicity in vacuo.

All structures were optimized at the DFT level by using the B3LYP hybrid functional[39] and the 6-311++G(d,p) basis set both in vacuo and solution. Solvent effects were described through a continuum approach by means of the IEF[28–30] version of PCM,[15] as implemented in the G09.[20] The molecular cavity surrounding the molecular solute was built by interlocking spheres, according to G09 default settings. The size of the cavity was also varied by applying different choices of the cavity size scaling factor α.

Free energy gaps and Boltzmann populations in vacuo were obtained by including zero-point and thermal contributions. The same quantities in solvent were obtained in a similar way by further including nonelectrostatic (repulsion, dispersion, and cavitation) energy contributions,[14] calculated over the same cavity used in the evaluation of the electrostatic term.

Anharmonic terms were obtained by the perturbative approach as reported in refs 10 and 33. Fermi resonances were handled in all calculations by suitable settings in the G09 anharmonic calculations. Although G09 can selectively calculate anharmonicities for a selection of normal modes, due to the limited size of the systems under study the calculations were done on all normal modes.

Results and Discussion

In this section, the calculation of anharmonic frequencies for formaldehyde and some simple mono- and dipeptide prototypes are reported. The discussion of the results is organized as follows. First, the PCM approach is tested against experiments in the case of formaldehyde in acetonitrile, in order to establish the quality of the PCM results in reproducing solvent to vacuo shifts as a function of the choice of the size of the PCM molecular cavity surrounding the solute.

Then, the model is applied to the description and prediction of anharmonic shifts of the amide vibrations of simple mono- and dipeptide prototypes in aqueous solution.

Formaldehyde in Acetonitrile: Benchmark of the Molecular Cavity with Respect to Experiment. Before discussing in detail the result of the calculation, it is mandatory to spend a few words on the construction and use of the molecule-shaped cavity. This is one of the most important features which distinguishes the PCM from other continuum models making use of much simpler cavities, such as spheres or ellipsoids, which in many cases are not well suited to reproduce the whole solvent effect.[40] As a matter of fact, the shape and size of the molecular cavity are the only adjustable parameters in cavity-based models (for a given solvent, i.e., once the dielectric constant has been set), and thus they are responsible for the uncertainty and arbitrariness of the results of the calculations.[14,25,40]

The choice of the molecular cavity in PCM is not univocal, such as in the self-consistent isodensity polarized continuum model (SCIPCM) approach,[41] but depends on the number, position, and radii of the spheres which are used to build the cavity itself. In the current implementation of PCM, the spheres are placed on the molecule nuclei, but the number of spheres (i.e., the number of nuclei in the same sphere) and their radii are adjustable parameters, whose choice is far from being trivial and often is left to the code default settings or to user's sensibility. However, the importance of a good definition of the radii is well-known; many studies can be quoted[25,42,43] but, until now, no definitive rules have been found.

In principle, the size and shape of the molecular cavity cannot be defined once and for all, it has only a limited physical meaning (the interface between the solute and the solvent) but a crucial numerical role, being the boundary in the definition of the PCM operators.

The definition of the molecular cavity in PCM calculations has been done so far basically in two different ways. The simplest (and the original one) consists of using one sphere for each atom, with the radius equal to the atomic van der Waals radius (Bondi[44] and Pauling[45] sets of radii are often used), whereas the other one is the use of an united-atom type cavity,[42] with radii obtained by fitting the solvation free energy at a given QM level with the corresponding experimental values. Of course the fitting with respect to solvation free energy is only one of many possible criteria, any molecular property being, in principle, exploitable for obtaining a reliable fitting and the best parameters for the cavity.

However, in some way this is an ill-posed question as the difficulty of representing a complex phenomenon, such as solvent effects, cannot be limited to a check on a single property. On the other hand, it is also quite impossible to find a universal definition of the best parameters valid for any kind of phenomenon, process, or property. The best strategy is probably to adopt a given set of cavity parameters chosen among sufficiently 'safe' values (for example those derived from some experimental data or from well established theoretical models) and then to check the stability of the results obtained under this assumption by varying the same parameters in a range, without giving up on the reliability of the model. If the choice adopted is sufficiently sound, then the set of computed results will not be dramatically affected by reasonable variations of the parameters involved in the cavity definition.[25]

As far as the reproduction of anharmonic solvent to vacuo frequency shifts for formaldehyde in acetonitrile is concerned, we would like to underline the fact that the reason for the choice of such a system is two-fold. First, considering that we are basically interested in testing the performance of the method as a function of the cavity parameters, formaldehyde is a sufficiently small carbonyl compound for which a great number of calculations can be run in a reasonable time. Second, being that the IR spectrum of formaldehyde is relatively simple, experimental results reported in the literature should be reasonably accurate. Notwithstanding this, the analysis reported, involving a specific QM level of calculation (B3LYP/6-311++G**) and a given solvation method (PCM-IEF), is not exhaustive, implying that other choices would have possibly given different results.

Calculated harmonic and anharmonic frequencies of formaldehyde in the gas phase together with their experimental counterparts taken from refs 26 and 27 are reported in Table 1. The comparison between calculated and experimental data is satisfactory, thus showing the adequacy of the chosen combination of DFT functional and basis set. As expected, the introduction of anharmonic effects substantially increases the agreement between calculated and experimental absolute frequency values, being ν_3 the best case (-2 cm^{-1}) and $2\nu_5$ and $2\nu_2$ the worst cases (88 and -81 cm $^{-1}$, respectively).

Table 1. Calculated B3LYP/6-311++G** Harmonic and Anharmonic Frequencies (in cm^{-1}) of Formaldehyde in the Gas Phase

	harmonic	anharmonic	exptl[a]
ν_4	1202	1181	1167
ν_6	1260	1240	1249
ν_3	1531	1499	1501−1500
ν_2	1815	1789	1745−1746
$2\nu_2$		3558	3470−3472
ν_1	2885	2722	2782
ν_5	2942	2762[b]	2843

[a] From refs 26, 27. [b] The calculated value is 2722 cm^{-1} if Fermi resonance is considered.

Calculated harmonic and anharmonic frequency values of formaldehyde in acetonitrile are reported in Tables 2 and 3, respectively, as a function of the size of the molecular cavity around the solute. In all cases, the cavity is built by one sphere on each atom, with the radius equal to G09 default settings, i.e., R(C) = 1.926, R(O) = 1.75, and R(H) = 1.443 Å. As it appears from the examination of the data, a definite trend of absolute values is not evident. Although, at least in principle, values obtained increasing the cavity size should go toward values in vacuo (where the size of the hypothetic cavity is virtually infinite); this is observed only in the case of harmonic frequencies, whereas, as far as anharmonic terms are concerned, larger fluctuations can be noticed. The tendency to reach in vacuo values is less evident and seems to depend on the normal mode under consideration. From the results obtained it could be speculated that higher frequency modes tend to the limit more slowly. In all tables, SMD refers to the use of the optimized cavity reported by Truhlar and co-workers[46] available in G09.

Moving to the comparison between calculated and experimental absolute values, as expected, the inclusion of anharmonic corrections substantially improves the prediction of experimental values. Of course, the discrepancy between absolute values not only depends on the representation of solvent effects but also is strongly influenced by the QM level of description.

In order to get better insight into solvent induced effects and to remove the intrinsic uncertainty due to the particular choice of the QM level, the frequency shift obtained moving from vacuo to solvent should be evaluated. In the latter case, in fact, only the quality of the solvation model is put into evidence.

Solvent to vacuo shifts are reported in Table 4, as a function of the size of the cavity, and the comparison with experiment is shown in a pictorial way in Figure 2. The quality of the agreement with experiment strongly depends on the cavity size, being generally better for larger cavities. The best agreement for formaldehyde in acetonitrile seems to be obtained with α = 1.3, i.e., a cavity a bit larger than the default setting in G09 (α = 1.1). Figure 2 also shows values previously calculated by Begue et al.,[26,27] using a fitting of the PES calculated by exploiting the SCIPCM continuum model.[41] Both the quality of our PCM results with α = 1.3, which is very close to the one obtained by Begue et al., and the agreement with the experimental data are very satisfactory thus confirming the reliability of the PCM approach to model solvent effects on vibrational properties and spectroscopies[25,47−51] also in the case of tricky effects, such as anharmonic ones.

To end this section, it is worth noticing that the SMD cavity, which has been parametrized using a large training set of neutral and ionic solvation free energies for various solutes in water and organic solvents, i.e., is specific for the quantitative description of solvation energies, seems not to be adequate to describe vibrational anharmonic frequencies. Such an inaccuracy is reasonably due to the fact that the SMD cavity is the smallest of the series of exploited cavities, so that the use of larger cavities is to be advised for the evaluation of solvent effects on vibrational frequencies by means of the PCM.

Amide Modes in Peptide prototypes. As already mentioned in the Introduction, IR spectroscopy can be very useful to identify and characterize peptide structure, conformational preference, and reactivity in solution. Thus, it is important to give an accurate and a detailed theoretical description of peptide vibrational spectra in order to put a clear interpretation on the experimental evidence.

In this section, a few mono- and dipeptide prototypes in water solution (Figure 1) are considered, and the focus is on amide-A, -I, and -II, which are the most exploited for peptide structure determination.

Calculated harmonic and anharmonic frequencies of selected modes of formamide and *trans-N*-methyl acetamide (NMA) in the gas phase are listed in Table 5, where the experimental values[52,53] are also shown for comparison. The agreement between calculated and experimental values is very good, and the results are even better than those obtained for formaldehyde, thus confirming the appropriateness of the chosen combination of DFT functional and basis set for the description of amide modes (the calcd vs exptl discrepancies are of the order of few cm^{-1}).

Considering that we are interested in evaluating effects due to an aqueous solvent, we have chosen to report the PCM results obtained by using α = 1.25, which is consistent with the characteristics of the medium.

Selected harmonic and anharmonic frequencies of the amide modes for the systems chosen are reported in Table 6, and calculated anharmonicities are reported and compared with experimental findings taken from the literature in Table 7. The values in the table are obtained as follows:

$$\Delta_{ii} = 2\nu_i - \nu_{2i} \tag{1}$$

$$\Delta_{ij} = \nu_i + \nu_j - \nu_{ij} \tag{2}$$

Diagonal anharmonicities are larger than those of off-diagonal ones, and amide-A anharmonicities are larger than those of amide-I and -II, which are almost comparable for all molecules.

The comparison between calculated and experimental data shows a very good correlation between the two sets (see also Figure 3). Our PCM values are always within three error bars and generally correlate with experiments better than the data previously reported by Wang and Hochstrasser,[38] obtained from B3LYP/6-31+G** calculations in vacuo. Also worth noting, the ca. 30 cm^{-1} difference in reproducing

Table 2. Selected Harmonic Frequencies (in cm^{-1}) of Formaldehyde in Acetonitrile[a]

	α = 1.1	α = 1.15	α = 1.2	α = 1.25	α = 1.3	α = 1.4	SMD	exptl[b]
ω_4	1214	1213	1212	1211	1210	1209	1219	
ω_6	1257	1257	12587	1258	1258	1258	1255	1247
ω_3	1526	1527	1528	1528	1529	1529	1521	1503
ω_2	1783	1786	1790	1792	1795	1798	1774	1723−1726
ω_1	2921	2917	2914	2911	2908	2904	2918	2797−2808
ω_5	2992	2986	2982	2978	2975[c]	2969	2987	2876

[a] Calculated data at the B3LYP/6-311++G** level with various choices of the molecular cavity. The cavity is made by four interlocking spheres centered at atoms, with the following radii values: *R*(C) = 1.926, *R*(O) = 1.750, and *R*(H) = 1.443 Å, each multiplied by the α factor in the table. [b] From refs 26 and 27. [c] It is 2822 with Fermi resonance.

Table 3. Selected Anharmonic Frequencies (in cm^{-1}) of Formaldehyde in Acetonitrile[a]

	α = 1.1	α = 1.15	α = 1.2	α = 1.25	α = 1.3	α = 1.4	SMD[b]	exptl[c]
ν_4	1190	1193	1175	1185	1193	1190	1219	
ν_6	1244	1253	1235	1239	1237	1247	1232	1247
ν_3	1500	1508	1493	1499	1496	1505	1480	1503
ν_2	1757	1762	1761	1766	1767	1774	1744	1723−1726
$2\nu_2$	3496	3506	3504	3514	3515	3529	3468	3434
ν_1	2756	2753	2746	2748	2747	2743	2760	2797−2808
ν_5	2763	2768	2767	2773	2775	2774	2760	2876

[a] Calculated data at the B3LYP/6-311++G** level with various choices of the molecular cavity. The cavity is made by four interlocking spheres centered at atoms, with the following radii values: *R*(C) = 1.926, *R*(O) = 1.750, and *R*(H) = 1.443 Å, each multiplied by the α factor in the table. [b] The SMD cavity for formaldehyde in acetonitrile is *R*(C) = 1.85, *R*(O) = 2.186, and *R*(H) = 1.2 Å and α = 1.00. [c] From refs 26 and 27.

Table 4. Anharmonic Solvent to Vacuo Sol−Vac Shifts (in cm^{-1}) of Formaldehyde[a]

	α = 1.1	α = 1.15	α = 1.2	α = 1.25	α = 1.3	α = 1.4	SMD[b]	exptl[c]
$\Delta\nu_4$	9	12	−6	4	12	9	38	
$\Delta\nu_6$	4	13	−5	−1	−3	7	−8	−2
$\Delta\nu_3$	1	9	−6	0	−3	6	−19	3/2
$\Delta\nu_2$	−32	−26	−27	−22	−22	−15	−45	−23/−19
$\Delta 2\nu_2$	−62	−52	−54	−44	−43	−29	−90	−36
$\Delta\nu_1$	34	31	24	26	25	21	38	15/26
$\Delta\nu_5$	1	6	5	11	13	12	−2	33

[a] Calculated data at the B3LYP/6-311++G** level with various choices of the molecular cavity. The cavity is made by four interlocking spheres centered at atoms, with the following radii values: *R*(C) = 1.926, *R*(O) = 1.750, and *R*(H) = 1.443 Å, each multiplied by the α factor in the table. [b] The SMD cavity for formaldehyde in acetonitrile is *R*(C) = 1.85, *R*(O) = 2.186, and *R*(H) = 1.2 Å and α = 1.00. [c] From refs 26 and 27.

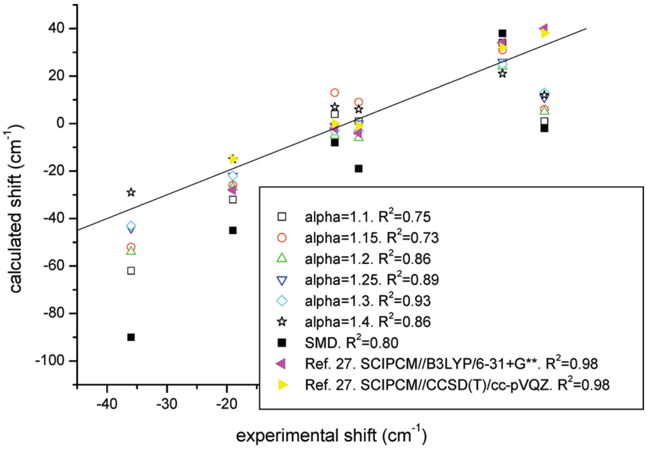


Figure 2. Formaldehyde in acetonitrile. Correlation between experimental and calculated sol−vac frequency shifts; values in cm^{-1}. Experimental data taken from ref 27.

Table 5. B3LYP/6-311++G** Harmonic and Anharmonic Frequencies (in cm^{-1}) of the Amide-A, -I, and -II Modes of Formamide and *trans*-NMA in the gas phase

	formamide		*trans*-NMA	
	calcd	exptl[a]	calcd	exptl[b]
ω_A	3716		3643	
ω_I	1791		1744	
ω_{II}	1618		1559	
ν_A	3513		3493	3498
ν_{2A}	6943		6843	
ν_I	1760	1755	1713	1708
ν_{2I}	3504		3408	
ν_{II}	1577	1580	1501	1511
ν_{2II}	3150		2989	
ν_{AI}	5271		5203	
ν_{III}	3338		3213	

[a] Ref 52. [b] Ref 53.

amide-I mode of *trans*-NMA is reasonably due to the discarding of the directional component of the hydrogen-bond effects, so that resorting to a supermolecule approach (*trans*-NMA + water clusters) would probably go toward the right direction. However, part of the effect is also due to the use of DFT wave functions.

***trans* (s)-*N*-Methyl Acetylproline Amide.** As confirmed by previous studies,[48] the PCM is able to give a reliable description of conformational effects and spectroscopic properties, such as IR/VCD, Raman/VROA, UV/CD, ORD, and NMR of *trans* (s)-*N*-acetylproline amide (AcProNH$_2$)

Table 6. B3LYP/6-311++G** Harmonic and Anharmonic Frequencies (in cm^{-1}) of the Amide-A, -I, and -II Modes of Various Mono- and Dipeptides in Water[a]

	formamide	*trans*-NMA	AcAlaOMe (ester)	AcAlaOMe (carbonyl)	$AcProNH_2$ (C_7)
ω_A	3704	3645	3628	3603	3455
ω_I	1735	1690	1698	1690	1714
ω_{II}	1621	1549	1532	1537	1604
ν_A	3510	3508	3464	3424	3283
ν_{2A}	6938	6876	6783	6702	6412
ν_I	1704	1653	1672	1657	1676
ν_{2I}	3393	3290	3326	3297	3334
ν_{II}	1565	1558	1514	1492	1547
ν_{2II}	3112	3102	3011	2970	3078
ν_{AI}	5212	5158	5132	5076	4958
ν_{III}	3265	3210	3183	3149	3218

[a] The PCM cavity is defined in terms spheres centered on each atom with the following radii: R(C) = 1.926, R(O) = 1.750, and R(H) = 1.443 Å, further multiplied by α = 1.25.

Table 7. B3LYP/6-311++G** Calculated Amide Anharmonicities (cm^{-1}) of Various Mono- and Dipeptides in Water[a]

		$\Delta_{A\,A}$	$\Delta_{I\,I}$	$\Delta_{II\,II}$	$\Delta_{A\,I}$	$\Delta_{I\,II}$
formamide	exptl	129				
	calcd	81.3	15.8	18.6	2.4	5.0
trans-NMA	exptl		16			3.5 ± 0.5
	calcd	140.4	17.0	12.8	4.0	1.0
AcAlaOMe (C5 ester)	exptl	144 ± 7			1.4 ± 0.4	
	calcd	145.1	17.5	16.6	3.6	2.1
AcAlaOMe (C5 carbonyl)	exptl	144 ± 7			2.6 ± 0.8	
	calcd	144.5	17.5	14.7	4.6	1.1
$AcProNH_2$ (C7)	exptl	165 ± 15[b]	13 ± 2	13 ± 2	3.5[b]	4.1 ± 0.6
	calcd	154.4	18.4	15.7	1.3	5.2

[a] Experimental findings taken from ref 38, and references therein are also reported for comparison. The PCM cavity is defined in terms spheres centered on each atom with the following radii: R(C) = 1.926, R(O) = 1.750, and R(H) = 1.443 Å, further multiplied by α = 1.25. [b] Experimental value for AcProNHMe.[58]

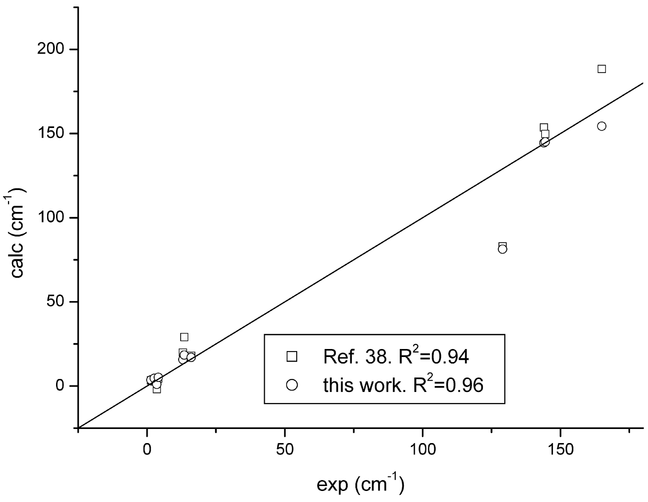


Figure 3. Selected mono- and dipeptides (see text). Correlation between experimental and calculated anharmonicities; values in cm^{-1}. Experimental values taken from ref 38 and references therein. Calculated values at the B3LYP/6-31+G** level in vacuo, taken from the same reference, are also reported.

in aqueous solution and, thus, can be confidently applied to evaluate anharmonic effects for this system in water and CH_2Cl_2.

It has been shown[48] previously, that only 3_{10} helix I and C_7 are stable minima in the gas phase, whereas three structures, i.e., 3_{10} helix I, P_{II}, and C_7, with different conformational weights, coexist in water. In particular, $AcProNH_2$ assumes almost exclusively the C_7 conformation (with only 1% of the 3_{10} helix I) at room temperature in the gas phase. The situation changes in water solution, where P_{II}, 3_{10}, and C_7 structures are present, and their percentage populations are 68, 28, and 4%, respectively. The combination of such a conformational ranking with PCM calculated response and spectroscopic properties led to a very good description of experimental spectra, thus showing, once again, the reliability of the PCM approach to describe the solvation of $AcProNH_2$.

Moreover, discrepancies between calculated (harmonic) and experimental vibrational frequencies were in the range of 30−50 cm^{-1}, which coincides with that between calculated harmonic and anharmonic amide-I frequencies in the gas phase for the methylated analogue (see Wang and Hochstrasser, ref 38). Thus, $AcProNH_2$ seems an ideal candidate for further testing the current implementation of PCM anharmonic frequency evaluation. In the following tables, the attention will be focused on amide-I modes only, for which experimental results are available in the literature.[54–56] Data obtained for other amide modes are reported in the Supporting Information.

Calculated anharmonic amide-I frequencies of the three conformers together with amide-I−amide-I off-diagonal anharmonicities are reported in Table 8 for three different cavities, i.e., the G09 default value, the α = 1.25 value, and the cavity exploited in ref 48. The inspection of the table reveals that vibrational anharmonic frequencies are sensitive to the peptide conformation, in close analogy with harmonic ones. In all cases, going beyond the harmonic approximation decreases the absolute values of a few tenths of cm^{-1}, i.e.,

Table 8. B3LYP/6-311++G** Calculated Amide I Anharmonic Frequencies and Anharmonicities (cm^{-1}) of the Various Conformations of $AcProNH_2$ in Water, with Different Choices of the Molecular Cavity

	C_7	3_{10}	P_{II}	average	exptl[a]
			$\alpha = 1.25$		
ω_i	1714	1711	1696		
ω_j	1652	1677	1647		
ν_i	1676	1681	1665	1676	1650
ν_j	1614	1639	1615	1617	1608
Δ_{ij}	1.24	0.08	0.80	1.07	
			Ref 48		
ω_i	1677	1667	1682		
ω_j	1632	1650	1642		
ν_i	1647	1650	1654	1653	1650
ν_j	1617	1613	1603	1606	1608
Δ_{ij}	0.84	0.41	1.55	1.20	
			$\alpha = 1.1$[b]		
ω_i	1695	1689	1704		
ω_j	1639	1657	1647		
ν_i	1660	1657	1672	1665	1650
ν_j	1601	1637	1610	1620	1608
Δ_{ij}	2.05	0.12	0.69	0.59	

[a] Ref 55. [b] Default value in G09.

Table 9. B3LYP/6-311++G** Calculated Boltzmann Populations of $AcProNH_2$ in Water with Different Choices of the PCM Molecular Cavity[a]

	C_7	3_{10}	P_{II}
$\alpha = 1.1$	0.34	0.36	0.30
$\alpha = 1.25$	0.68	0.19	0.13
ref 48	0.04	0.28	0.68
SMD	0.00	0.24	0.76
in vacuo[b]	0.99	0.01	0.00

[a] All data obtained by including ZPE, thermal ($T = 298$ K), and nonelectrostatic contributions. In vacuo results are reported for comparison. [b] Ref 48.

Table 10. B3LYP/6-311++G** Calculated Amide-I Harmonic and Anharmonic Frequencies and Anharmonicities (cm^{-1}) of the the Various Conformations of $AcProNH_2$ in CH_2Cl_2 with $\alpha = 1.3$

	C_7	3_{10}	P_{II}
ω_i	1725	1725	1739
ω_j	1660	1689	1677
ν_i	1688	1685	1690
ν_j	1626	1651	1651
Δ_{ij}	1.8	−0.1	0.1
exptl[a]		1.5 ± 0.4	

[a] Ref 57.

a similar range as the discrepancy between calculated harmonic and anharmonic amide-I frequencies in the gas phase reported by Wang and Hochstrasser for the methylated analogue.[38]

Calculated Boltzmann populations for $AcProNH_2$ in aqueous solution at room temperature, as a function of the size of the molecular cavity, are reported in Table 9. Data refer to 298 K and were obtained by considering free energies corrected for zero point energies (ZPE) and thermal contributions, and with the inclusion of nonelectrostatic solvent effects (cavitation, dispersion, and repulsion; see ref 40 for details). The weight of the C_7 conformer increases with the cavity size, whereas an opposite trend is observed for P_{II}. Such a behavior is expected because by increasing the cavity size, the limit of the in vacuo calculation should be reached. In particular, the picture does not change substantially by passing from $\alpha = 1.1$ (default cavity in G09) to the cavity reported in the previous study (see ref 48) or to the one used in the SMD, even if in the latter case the opposite limit with respect to vacuo is reached (i.e., the weight of the C_7 conformation is negligible). The use of $\alpha = 1.25$ substantially reaches the in vacuo limit where the P_{II} conformer is not present. Indeed, the stabilization of P_{II} is due to its interaction with water.

Calculated and experimental[55] frequencies are reported in Table 8 (average values).

In agreement with the results reported in ref 48, an appropriate evaluation of solvent effects on populations is crucial for the correct description of $AcProNH_2$ spectroscopic properties. In fact, use of the default G09, ref 48, or SMD cavities leads to a prevalence of the P_{II} conformations over the others, with only a very small (or even negligible) amount of the C_7. On the contrary, use of $\alpha = 1.25$ makes the results go toward in vacuo data, i.e., a prevalence of the C_7 conformation is predicted (see Table 9). The latter ranking, however, makes the average results go farther from experiments. In particular, the use of the cavity previously reported in ref 48 almost matches calculations with experimental values.

To the best of our knowledge, no experimental data for $AcProNH_2$ anharmonicities in aqueous solution have been reported so far in the literature. However, experimental values in dichloromethane have instead been measured.[57,58] Thus, our study has been extended to $AcProNH_2$ in dichloromethane. Data obtained by exploiting $\alpha = 1.3$ are listed in Table 10. Notice that, due to the lower polarity of dichloromethane with respect to water, a slightly larger α value has been chosen in this case. As already pointed out for water, the inclusion of anharmonic effects makes the calculated frequencies decrease, and once again the frequencies of amide-I band are sensitive to the surrounding environment. Even more sensitive are amide-I anharmonici-

Table 11. B3LYP/6-311++G** Calculated Boltzmann Populations of $AcProNH_2$ in CH_2Cl_2 with Different Choices of the PCM Molecular Cavity[a]

	C_7	3_{10}	P_{II}
$\alpha = 1.1$	0.63	0.14	0.23
$\alpha = 1.25$	0.73	0.15	0.12
$\alpha = 1.3$	0.88	0.05	0.07
SMD	0.07	0.32	0.62
same as ref 48	0.12	0.30	0.58

[a] All data obtained by including nonelectrostatic, ZPE, and thermal (298 K) contributions.

ties, which increase even by a factor of 100 moving from one conformation to another.

Comparison of anharmonicities of the single conformers with experimental findings seems to evidence a prevalence of the C_7 conformer in the dichloromethane solution or even a mixture with a relevant amount of C_7. This is in agreement with previous experimental findings reported in the literature for $AcProNH_2$ in dichloromethane and deuterated chloroform.[59,60]

As already pointed out in the previous paragraphs, the comparison between experimental and "average" calculated values is not straightforward and requires accurate and reliable evaluation of molecular properties and also a consistent prediction of the conformation hierarchy. Calculated PCM conformational weights are reported in Table 11.

The picture changes with respect to water (compare Table 9). In fact, the G09 default cavity and the largest $\alpha = 1.25$ and 1.3 cavity lead to a prevalence of the C_7 conformer, whereas other choices, including SMD, show a prevalence of the P_{II} conformer, which is, however, consistent neither with the anharmonicity values of Table 10 nor with previous experimental studies.[59,60]

In summary, by resorting to the results reported in the present study, it seems reasonable to state once again that the choice of a universal definition for the molecular cavity is far from trivial and that such a choice can hugely influence the outcome of the calculation. In particular, as far as the present examples are concerned, it seems reasonable to suggest the use of a larger PCM molecular cavity for the calculations of both molecular properties and conformational effects in medium-polarity solvents, whereas a small cavity is recommended for polar solvents, at least as far as conformational properties are concerned. However, due to a lack of extended studies on this matter in the last years, efforts to get a better parametrization of the molecular cavity seem to be absolutely necessary.

Summary, Conclusions, and Future Developments

We have reported some results related to the description of vibrational spectra of molecules in condensed phase with the simultaneous account of anharmonicity and solute−solvent interactions. The computational approach and the general trends, even if here applied to few small-to-medium sized systems, should not be considered as specific of given molecular structures and/or solvent, so that the present study represents a first exploration of a much wider topic.

The current implementation of the PCM in the Gaussian suite of programs[20] yields a continuous surface, which is smooth enough to be further differentiated to obtain the quantities needed for the evaluation of anharmonic vibrational frequencies.

The comparison of the calculated results with their experimental counterparts suggests the use of a molecular cavity of a different size for the evaluation of energetics and vibrational spectroscopic properties, depending on the nature of the solvent. However, this discrepancy could be overtaken by finer parametrization of the nonelectrostatic contributions, which enter the evaluation of the solvation free energy.[14,61–63] Work in this direction is currently in progress in our group.

Nevertheless, the quality of the results which have been obtained, in connection with the low cost and versatility of the PCM, shows that it is a valuable method for a quantitative description of vacuo-to-solvent harmonic and anharmonic frequency shifts. Also, due to features of the current implementation, which enables the user to discriminate between normal modes and then to choose to perform the calculation only on selected ones, the PCM evaluation of anharmonic frequencies in solution is nowadays applicable to large systems, also by combining the PCM description to QM/MM approaches.

The only real limitation still remaining is the difficulty of the continuum approach to evaluate specific solute−solvent interactions, for which a proper treatment requires, in almost all cases, resorting to techniques rooted in the molecular dynamics.

Acknowledgment. One of the authors (C.C.) thanks Prof. Jacopo Tomasi for useful discussion. Most of the calculations were performed on the computational resources made available to the authors at CNR-Village (Virtual Integrated Laboratory for Large-scale Applications in a Geographically distributed Environment; http://village.ipcf.cnr.it/).

Supporting Information Available: Calculated harmonic and anharmonic amide-II, amide-III, and amide-A frequencies of $AcProNH_2$ in aqueous and dichloromethane solutions. This material is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Cho, M. *Chem. Rev.* **2008**, *108*, 1331.

(2) Kim, Y. S.; Hochstrasser, R. M. *J. Phys. Chem. B* **2009**, *113*, 8231.

(3) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.

(4) Martin, J. M. L.; Lee, T. J.; Taylor, P. R.; Francois, J.-P. *J. Chem. Phys.* **1995**, *103*, 2589.

(5) Stanton, J. F.; Gauss, J. *J. Chem. Phys.* **1998**, *108*, 9218.

(6) Tew, D. P.; Klopper, W.; Heckert, M.; Gauss, J. *J. Phys. Chem. A* **2007**, *111*, 11242.

(7) Burcl, R.; Handy, N. C.; Carter, S. *Spectrochim. Acta, Part A* **2003**, *59*, 1881.

(8) Boese, A. D.; Martin, J. *J. Phys. Chem. A* **2004**, *108*, 3085.

(9) Barone, V. *J. Phys. Chem. A* **2004**, *108*, 4146.

(10) Barone, V. *J. Chem. Phys.* **2005**, *122*, 014108.

(11) Carbonniere, P.; Lucca, T.; Pouchan, C.; Rega, N.; Barone, V. *J. Comput. Chem.* **2005**, *26*, 384.

(12) Puzzarini, C.; Barone, V. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6991.

(13) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161.

(14) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3093.

(15) Miertus, S.; Scrocco, E.; Tomasi, *J. Chem. Phys.* **1981**, *55*, 117.

(16) Li, H.; Jensen, J. H. *J. Comput. Chem.* **2004**, *25*, 1449.

(17) Su, P.; Li, H. *J. Chem. Phys.* **2009**, *130*, 074109.

(18) Lange, A. W.; Herbert, J. M. *J. Phys. Chem. Lett.* **2010**, *1*, 556.

(19) York, D. M.; Karplus, M. *J. Phys. Chem. A* **1999**, *103*, 11060.

(20) Frisch, M. J. ; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision 02; Gaussian, Inc.: Wallingford, CT, 2009.

(21) Olivares del Valle, F. J.; Aguilar, M.; Tolosa, S.; Contador, J. C.; Tomasi, *J. Chem. Phys.* **1990**, *143*, 371.

(22) Aguilar, M. A.; Olivares del Valle, F. J.; Tomasi, *J. Chem. Phys.* **1991**, *150*, 151.

(23) Louis, J.; Daniel, R.; Dillet, V. *Mol. Phys.* **1996**, *89*, 1521.

(24) Rivail, J.-L.; Rinaldi, D.; Dillet, V. *J. Chim. Phys.* **1998**, *95*, 1818.

(25) Cappelli, C.; Mennucci, B.; da Silva, C. O.; Tomasi, J. *J. Chem. Phys.* **2000**, *112*, 5382–5392.

(26) Begue, D.; Carbonniere, P.; Barone, V.; Pouchan, C. *Chem. Phys. Lett.* **2005**, *416*, 206.

(27) Begue, D.; Elissalde, S.; Iratcabal, P.; Pouchan, C. *J. Phys. Chem. A* **2006**, *110*, 7793.

(28) Cancès, E.; Mennucci, B. *J. Math. Chem.* **1998**, *23*, 309–326.

(29) Cancès, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032.

(30) Mennucci, B.; Cancès, E.; Tomasi, J. *J. Phys. Chem. B* **1997**, *101*, 10506.

(31) Johnson, B. G.; Frisch, M. J. *J. Chem. Phys.* **1994**, *100*, 7429.

(32) Stratman, R. E.; Burant, J. C.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1997**, *106*, 10175.

(33) Barone, V. *J. Chem. Phys.* **2004**, *120*, 3059.

(34) Mennucci, B.; Cammi, R.; Tomasi, J. *J. Chem. Phys.* **1999**, *110*, 6858.

(35) Cossi, M.; Scalmani, G.; Rega, N.; Barone, V. *J. Chem. Phys.* **2002**, *117*, 43.

(36) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Comput. Chem.* **2003**, *24*, 669.

(37) Scalmani, G.; Frisch, M. J. *J. Chem. Phys.* **2010**, *132*, 114110.

(38) Wang, J.; Hochstrasser, R. M. *J. Phys. Chem. B* **2006**, *110*, 3798.

(39) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(40) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.

(41) Foresman, J. B.; Keith, T. A.; Wiberg, K. B.; Snoonian, J.; Frisch, M. J. *J. Phys. Chem.* **1996**, *100*, 16098.

(42) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210.

(43) Caricato, M.; Mennucci, B.; Tomasi, J. *Mol. Phys.* **2006**, *104*, 87.

(44) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441.

(45) *Handbook of Chemistry and Physics*; West, R. C., Ed.; Chemical Rubber Company: Cleveland, OH, 1981.

(46) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378.

(47) Cappelli, C. Continuum Solvation Approaches to Vibrational Properties. In *Continuum Solvation Models in Chemical Physics: Theory and Applications*; Mennucci, B., Cammi, R., Eds.; Wiley: Chichester, U.K., 2007, p 167.

(48) Cappelli, C.; Mennucci, B. *J. Phys. Chem. B* **2008**, *112*, 3441.

(49) Cappelli, C.; Corni, S.; Tomasi, J. *J. Phys. Chem. A* **2001**, *105*, 10807.

(50) Cappelli, C.; Corni, S.; Mennucci, B.; Cammi, R.; Tomasi J. *J. Phys. Chem. A* **2002**, *106*, 12331.

(51) Cappelli, C.; Mennucci, B.; Monti, S. *J. Phys. Chem. A* **2005**, *109*, 1933–1943.

(52) King, S. T. *J. Phys. Chem.* **1971**, *75*, 405.

(53) Ataka, S.; Takeuchi, H.; Tasumi, M. *J. Mol. Struct.* **1984**, *113*, 147.

(54) Hahn, S.; Lee, H.; Cho, M. *J. Chem. Phys.* **2004**, *121*, 1849–1865.

(55) Oh, K.-I.; Han, J.; Lee, K.-K.; Hahn, S.; Han, H.; Cho, M. *J. Phys. Chem. B* **2006**, *110*, 13335–13365.

(56) Lee, K.-K.; Hahn, S.; Oh, K.-I.; Choi, J. S.; Joo, C.; Lee, H.; Han, H.; Cho, M. *J. Phys. Chem. B* **2006**, *110*, 18834.

(57) Rubtsov, I. V.; Hochstrasser, R. M. *J. Phys. Chem. B* **2002**, *106*, 9165.

(58) Rubtsov, I. V.; Wang, J.; Hochstrasser, R. M. *J. Phys. Chem. B* **2003**, *107*, 3384.

(59) Karaiskaj, D.; Sul, S.; Jiang, Y.; Ge, N.-H. In *Ultrafast Phenomena XIV*; Kobayashi, T., Okada, T., Kobayashi, T., Nelson, K. A., De Silvestri, S., Eds.; Springer: Berlin, 2005, p 545.

(60) Sul, S.; Karaiskaj, D.; Jiang, Y.; Ge, N.-H. *J. Phys. Chem. B* **2006**, *110*, 19891.

(61) Orozco, M.; Marchan, I.; Soteras, I. Continuum analysis of conformational sampling in solution. In *Continuum Solvation Models in Chemical Physics: Theory and Applications*; Mennucci, B., Cammi, R., Eds.; Wiley: Chichester, U.K., 2007, p 499.

(62) Curutchet, C.; Orozco, M.; Luque, F. J.; Mennucci, B.; Tomasi, J. *J. Comput. Chem.* **2006**, *27*, 1769.

(63) Bidon-Chanal, A.; Huertas, O.; Orozco, M.; Luque, F. J. *Theor. Chem. Acc.* **2008**, *123*, 11.

CT100048G

JCTC Journal of Chemical Theory and Computation

Theoretical Study on the Redox Cycle of Bovine Glutathione Peroxidase GPx1: pK_a Calculations, Docking, and Molecular Dynamics Simulations

Syed Tahir Ali, Sajid Jahangir, Sajjad Karamat, Walter M. F. Fabian,
Krzysztof Nawara, and Juraj Kóňa*

Institute of Chemistry, Karl Franzens University, Graz, Heinrichstrasse 28, A-8010 Graz, Austria, Department of Chemistry, Federal Urdu University of Arts, Science and Technology, Gulshan-e-Iqbal Science Campus, Karachi, Sindh, Pakistan, and Institute of Chemistry, Center for Glycomics, Slovak Academy of Sciences, Dúbravska cesta 9, 845 38 Bratislava, Slovak Republic

Received July 2, 2009

Abstract: Three approaches of computational chemistry [quantum mechanics (QM) calculations, docking, and molecular dynamics (MD) simulations] were used to investigate the redox cycle of bovine erythrocyte glutathione peroxidase from class 1 (GPx1, EC 1.11.1.9). The pK_a calculations for two redox states of the active-site selenocysteine of GPx1 (selenol, Sec45−SeH, and selenenic acid, Sec45−SeOH) were estimated using a bulk solvent model (B3LYP-IEFPCM and B3LYP-CPCM-COSMO-RS). The calculated pK_a values of Sec45−SeH and Sec45−SeOH were corrected via a simple linear fit to a training set of organoselenium compounds, which consisted of aliphatic selenols and aromatic selenenic acids with available experimental pK_a values. Based on docking calculations, binding sites for both molecules of the cofactor glutathione (GSH) are described. MD simulations on the dimer of GPx1 have been performed for all chemical states of the redox cycle: without GSH and with one or two molecules of GSH bound at the active site. Conformational analyses of MD trajectories indicate high mobility of the Arg177 and His79 residues. These residues can approach the vicinity of Sec45 and take part in the catalytic mechanism. On the basis of the calculated data, new atomistic details for a generally accepted mechanism of GPx1 are proposed.

1. Introduction

Glutathione peroxidase (GPx, EC 1.11.1.9) was the first selenoprotein identified in mammals.[1] It protects cells from oxidative damage by catalyzing the reduction of H_2O_2, lipidhydroperoxides, and other organic peroxides, using glutathione (γ-glutamylcysteinylglycin, GSH) as the reducing substrate.[2] The X-ray structure of the tetrameric *Bos taurus* erythrocyte glutathione peroxidase from class 1 (GPx1) showed two asymmetric units containing two dimers, each with one selenocysteine residue (Sec) per active site at the monomer unit.[3] The active sites of GPx1 are found in flat depressions on the molecular surface and are located at a contact region of the dimer units. Biochemical,[2−7] kinetic[8] and crystallographic[3,7] studies have suggested that the Sec residue directly participates in the catalytic process of the reduction of hydroperoxide by GPx. It has been experimentally suggested that the catalytically active form of the enzyme is the selenolate anion (E−Se$^-$).[3,4] The proposed mechanism of the overall catalytic cycle is shown in Figure 1a. In the first redox step, E−Se$^-$ is oxidized to the selenenic acid (E−SeOH) with the accompanying reduction of a hydroperoxide substrate to a corresponding alcohol. In the second step, the E−SeOH reacts with GSH to produce a selanyl sulfide adduct (E−SeSG). In the third step, a second molecule of GSH attacks E−SeSG to regenerate the active form of the enzyme, and the oxidized form of GSH (GSSG) is formed as a byproduct. This step has been suggested to

* Corresponding author. Telephone: +421-2-59410322. Fax: +421-2-59410222. E-mail: chemkona@savba.sk.

10.1021/ct9003355 © 2010 American Chemical Society
Published on Web 04/22/2010

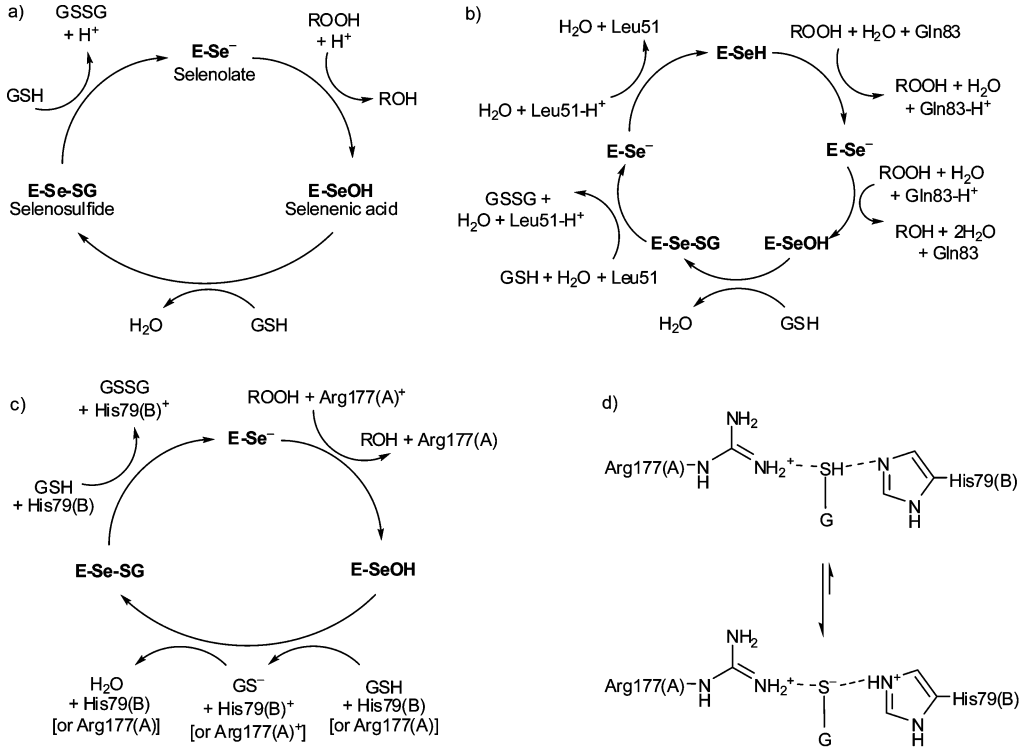


Figure 1. (a) Catalytic cycle of GPx1 after Epp et al.[3] The resting state of Sec is selenolate. A proton donor and acceptor in the first and third redox steps are not known. (b) Catalytic cycle of GPx3 proposed by Prabhakar et al.[10] and tested by density functional theory (DFT) calculations. The resting state of Sec is selenol. The concomitant proton transfers occur between Sec and amide nitrogen (side chain) of Gln83 and between GSH and amide nitrogen (backbone) of Leu51, facilitated by solvent water molecules. (c) A proposed catalytic mechanism of GPx1 in this work. The His79(B) and Arg177(A) residues play a role of catalytic acid/base based on docking, MD simulations, and pK_a calculations at the QM level. (d) In the presence of Arg177(A), a pK_a value of the thiol group of GSH can be depressed below 7, and the equilibrium shifted on the side of the thiolate state of GSH in the presence of His79(B).

be the rate-determining step of the entire mechanism.[9] As can be seen in the reaction scheme (Figure 1), a proton must be supplied for the first redox step and be abstracted in the third reaction step to maintain overall stoichiometry of the catalytic process. The mode of action of these proton transfers is not clear because no ionizable amino acid residues were found in the proximity of the Sec reaction center. Recently, a redox mechanism based on the proton transfer via solvent water molecules was proposed for human plasma glutathione peroxidase 3 (GPx3)[10−12] and for the redox reactions of organoselenium compounds[13−16] using quantum mechanics (QM) calculations. In the proposed mechanism for GPx3[10−12] (Figure 1b), water molecules participate in proton exchange between the selenol (and thiol groups) of selenocysteine (and glutathione) and the protein backbone. This is based on the assumption that the resting form of Sec is selenol, rather than the experimentally proposed selenolate state,[3,4] e.g., pK_a value of Sec at GPx3 should be higher than 7 according to this mechanism. The mechanism had reasonable kinetics parameters, similar to those found for the redox reactions of selenocysteine,[17] organic selenols,[13,14,18−20] and thiols.[21−23]

The selenolate state of the Sec residue, rather than selenol, is further supported by its lower pK_a value of 5.3[24] compared with cysteine (8.3).[2,25] Moreover, the selenolate state of the active-site Sec in selenosubtilisin was validated by NMR experiments (pK_a < 4).[26] In contrast to the active site of the selenosubtilisin,[27] which consists of the Sec−His−Asp triad, no ionizing residues in direct interaction with the active-site Sec were found in the available crystal structures of GPx.[3,7,28,29] The ionizing Arg and His residues do place in the active site with a radius of 6−10 Å from Sec45 in GPx1. The His79 is located in the contact dimer region.[3] The sequence His79−Lys84, from the monomer unit B, participates at the active site of the monomer unit A and vice versa, the sequence from monomer A is involved at the active site of the monomer B. When GPx1 was treated with ammonium peroxodisulphate, loss of enzymatic activity with presumably His79 complexed with peroxodisulphate, was found as a major observation.[3] Difference Fourier analysis of the corresponding derivative revealed a positive difference density maximum, stretching symmetrically across the local axis at the positions of residues interpreted as His79(A) and His79(B). Presumably these histidine residues can influence the catalytic process, but this phenomenon is not understood in molecular terms at present.[3] Several other studies have demonstrated a catalytic role of the histidine in cysteine and selenocysteine redox proteins,[27,30−33] as thioredoxin reductases[35−37] or an OxyR transcription factor.[38,39]

Because the crystal structures of glutathione peroxidases present a nonreactive oxidized seleninic acid state (E−SeOO$^-$) or a Gly (or Ala) mutant of the Sec residue, we speculate that the position of side chains of the active-site Arg and His residues could be different in vivo compared with the crystal structures, mainly in the presence of one or

Table 1. Enzyme and Cofactor Configurations Used in the Docking and MD Simulations[a]

	Sec45(A), Sec45(B)	His79(A), His79(B)	GSH (in A)	Sec45(A), Sec45(B)	His79(A), His79(B)	Asn75−Gln81(A), Asn75−Gln81(B)
Dock01_GSH	R−SeH	His−NεH^0	R−SH			
Dock02_GSH_GSH	R−SeH	His−NεH^0	2 R−SH			
Dock03_GSH_GS	R−SeH	His−NεH^0	RSH + R−S$^-$			
Dock04_GSSH	R−SeH	His−NεH^0	R−S−S−R			
MD1_Se	R−Se$^-$	His−NεH^0	No	−	+	+
				−	−	−
MD1_S	R−S$^-$	His−NεH^0	No	+	−	−
				−	−	−
MD2_Se_GSH_GSH	R−Se$^-$	His−NεH^0	2 R−SH	−	−	−
				−	+	+
MD3_SeOH	R−SeOH	His−NεH^0	No	−	+	−
				−	−	−
MD4_SeO	R−SeO$^-$	His−NεH^0	No	−	+	−
				−	−	−
MD5_SeOH_GSH	R−SeOH	His−NεH^0	R−SH	+	+	+
				−	+	+
MD6_SeO_GSH	R−SeO$^-$	His−NεH^0	R−SH	−	+	+
				+	+	+
MD7_SeO_GSH_GSH	R−SeO$^-$	His−NεH^0	2 R−SH	−	−	−
				−	+	+
MD8_SeO_GSH_GS_HIP	R−SeO$^-$	His−NεH^0	R−SH + R−S$^-$	−	−	−
		His$^+$(B)		−	−	−
MD9_SeOH_GSH_GS_HIP	R−SeOH	His−NεH^0	R−SH + R−S$^-$	−	−	+
		His$^+$(B)		−	+	+
MD10_SeSG	R−Se−S−R	His−NεH^0	No	−	−	+
	R−Se$^-$(B)			+	−	−
MD11_SeSG_GSH	R−Se−S−R	His−NεH^0	R−SH	−	−	−
	R−Se$^-$(B)			+	+	+
MD12_Se_GSSG	R−Se$^-$	His−NεH^0	R−S−S−R	−	+	+
				−	−	+

[a] Plus brief results from the MD analyses concerning conformations of Sec45 and His79 and of the Asn75−Gln81 loop; (+) means structural changes compared with the X-ray structure of GPx1 (PDB ID: 1GP1), i.e., a flip of Sec45 or conformational change of the side chain of His79 or the Asn75−Gln81 loop.

two molecules of GSH bound at the active site (for example, two different conformations of the side chain of Arg180 in human GPx1 were observed, PDB ID: 2F8A).[40]

Here we calculated: (i) pK_a values of Sec45 in two redox states to predict their forms at the physiological pH; and (ii) binding positions of both molecules of GSH, which revealed the importance of the Trp158, Arg177, and His79 for proper binding and reactivity of GSH. We also performed MD simulations on a nanosecond time scale to analyze the mobility of side chains of ionizable residues in the active site of GPx1. We will show that Arg177 and His79 can fill positions in the vicinity of selenol and thiol groups of Sec and GSH, respectively.

2. Computational Details

2.1. Structural Models and Docking. Based on the crystal structure of bovine GPx1 (PDB ID: 1GP1),[3] the 13 following configurations were simulated (Table 1) GPx1: with Sec in selenolate state without (MD1_Se) and with oxidized glutathione (MD12_Se_GSSG), with two molecules of GSH (MD2_Se_GSH_GSH), with Sec mutated to cysteine (MD1_S), in selenenic state (MD3_SeOH) and its ionized state (MD4_SeO), with one molecule of the reduced GSH (MD5_SeOH_GSH and MD6_SeO_GSH), with two molecules of GSH (combination of different ionization states of Sec45−SeOH, GSH and His79; MD7_SeO_GSH_GSH, MD8_SeO_GSH_GS_HIP, MD9_SeOH_GSH_GS_HIP), and in selanyl sulfide state without (MD10_SeSG) and with the reduced GSH (MD11_SeSG_GSH). Missing structural information on glutathione bound at the active site as well as on selanyl sulfide intermediate (E−SeSG) of GPx1 was added by means of docking calculations, employing the GLIDE program of the Schrödinger package.[41] One molecule of GSH was docked into the active site (Dock01_GSH, the selected configurations of the enzyme are compiled in Table 1), and then the second molecule of GSH (neutral GSH as well as ionized GS$^-$) was docked into the active site with the first molecule of glutathione already bound (Dock02_GSH_GSH and Dock03_GSH_GS). The oxidized form of glutathione was also docked into the active site (Dock04_GSSG). The selanyl sulfide intermediate of GPx1 for the MD10_SeSG and MD11_SeSG_GSH simulations was built from a docked pose (no. 16, Dock03_GSH_GS, see Figure 3e and a three-dimensional (3-D) structure in the Supporting Information) of the complex consisting of GPx1 and two molecules of the reduced form of GSH bound at the active site. In this complex a molecule of GSH (with a proximal position of its thiol group to the selenol group of Sec45 of GPx1) was covalently bound to Sec45 and subsequently minimized with the rest of the enzyme frozen at the crystallographic geometry using the MacroModel of the Schrödinger package.[42] For the constrained minimization, the OPLS2001 force field[43,44] was used for both enzyme and glutathione.

The GLIDE program[41,45] uses a hierarchical series of filters to search for possible locations of the ligand in the active-site region of the receptor. The shape and properties of the receptor are represented on a grid by several sets of fields that provide progressively more accurate scoring of the ligand poses. Conformational flexibility is handled in GLIDE by an extensive conformational search, augmented by a heuristic screen. The scoring is carried out using the Schrödinger's discrete version of the ChemScore empirical scoring function. Much as for ChemScore itself, this algorithm recognizes favorable hydrophobic, hydrogen bonding, metal−ligand interactions, desolvation, and entropic effects and penalizes steric clashes.[46] For the docking calculations, default values of parameters were used. The receptor box for the docking conformational search was centered at Sec45 with a size of 30 × 30 × 30 Å, using partial atomic charges for the GPx1 receptor from the OPLS2001 force field.[43,44] The grid maps were created with no van der Waals radius and charge scaling for the atoms of the receptor. Flexible docking in standard,[45] which penalizes nonplanar conformation of amide bonds, was used for the glutathione ligands. The partial charges of the ligands were calculated at the ab initio level (for more details see the Section 2.2, MD simulations). The potential for nonpolar parts of the ligands was softened by scaling the van der Waals radii by a factor of 0.8 for atoms of the ligands with partial atomic charges less than specified cutoff of 0.15. The 5000 poses were kept per ligand for the initial docking stage with a scoring window of 100 kcal mol^{-1} for keeping initial poses; and the best 400 poses were kept per ligand for energy minimization. The ligand poses with root-mean-square (rms) deviations less than 0.5 Å and maximum atomic displacement less than 1.3 Å were discarded as duplicates. One hundred ligand poses with the best docking score were saved for subsequent analyses using the MAESTRO viewer of the Schrödinger package.[47]

2.2. MD Simulations. Simulations with the explicit solvent model (TIP3P)[48] were performed with the standard AMBER 99 force field by means of the GROMACS 3.3.1[49] program package. For nonstandard amino acids and their ionized states (Sec−SeH, Sec−Se$^-$, Sec−SeOH, Sec−SeO$^-$, and Sec−SeSG) as well as for all states of the glutathione (GSH, GS$^-$, and GSSG), parameters were derived from either AMBER and GAFF force fields or built from data obtained by ab initio Hartree−Fock calculations[50] [HF/6-31G(d)] using the Gaussian 03 package.[51] The electrostatic potential fitting algorithm of Merz−Singh−Kollman scheme[52,53] was used [the keywords POP = MK IOP(6/33 = 2, 6/42 = 6)] to estimate atomic charges from the HF/6-31G(d) calculations (atom types, charges and added force field parameters are provided in the Supporting Information). The ionization states of the ionizing residues of the enzyme were predicted by the PropKa program,[54,55] considering an in vivo pH of 7 (calculated pK_a values, and an output structure of GPx1 with predicted ionization configurations of the amino acid residues are available in the Supporting Information). Terminal and side-chain ionizing groups (amino and carboxyl) of glutathione molecules were treated in their ionized configurations (as $-NH_3^+$ or $-COO^-$) in all docking calculations and MD simulations.

$$\begin{tikzcd}
\mathrm{AH(g)} \arrow[r, "\Delta G(\mathrm{g})"] \arrow[d, "\Delta G(\mathrm{AH})_{\mathrm{gs}}"'] & \mathrm{A^{-}(g)} \arrow[d, "\Delta G(\mathrm{A^{-}})_{\mathrm{gs}}"] & + & \mathrm{H^{+}(g)} \arrow[d, "\Delta G(\mathrm{H^{+}})_{\mathrm{gs}}"] \\
\mathrm{AH(s)} \arrow[r, "\Delta G(\mathrm{s})"] & \mathrm{A^{-}(s)} & + & \mathrm{H^{+}(s)}
\end{tikzcd}$$

Figure 2. Thermodynamic cycles of ionizing groups in aqueous solution.

The Berendsen algorithm[56] for temperature and the Parrinello−Rahman algorithm for pressure coupling, with coupling constants of τ_t =1.0 ps and τ_p =1.0 ps, were used at a constant temperature (300 K) and pressure (101.325 kPa). Periodic boundary conditions were used together with the particle-mesh Ewald (PME) method[57] for treating long-range electrostatics. A time step of 1.0 fs, with the LINCS algorithm[58] to constrain bonds involving hydrogens, was used along simulations with a 10 Å nonbonded cutoff, and the nonbonded pairlist was updated every 20 time steps. The simulations, preceded by initial minimizations (1000 steps), were carried out over 10 ns, and coordinates were saved for analysis every 1 ps. The protein was solvated by more than 14 000 TIP3P[48] water molecules in a box (66 × 108 × 71 Å) using the LEAP[59] program of the AMBER 8 package[60] (total number of atoms in the simulated system is ca. 50 000).

2.3. pK_a Calculations. Based on the thermodynamic cycle[61−64] (Figure 2) pK_a values were calculated using the following formulas:

$$\mathrm{p}K_a = \Delta G/RT \ln 10$$

$$\mathrm{p}K_a(\mathrm{s}) = [\Delta G(\mathrm{g}) + \Delta G(\mathrm{A}^-)_{\mathrm{gs}} - \Delta G(\mathrm{AH})_{\mathrm{gs}} + \Delta G(\mathrm{H}^+)_{\mathrm{gs}}]/2.303RT$$

Here pK_a(s) is for a functional group in an aqueous solution. $G(\mathrm{H}^+) = -6.28$ kcal mol^{-1} and $\Delta G(\mathrm{H}^+)_{\mathrm{gs}} = -264.0$ kcal mol^{-1} were derived experimentally,[65] where the standard state is 1 M [$\Delta G(\mathrm{H}^+)_{\mathrm{gs}}$ was calculated from the Tissandier et al.[63,66] value of −265.9 kcal mol^{-1} and from the free energy change (1.89 kcal mol^{-1} at 298 K) associated with moving from a standard state that uses the concentrations of 101.325 kPa in the gas phase and 1 M in the aqueous phase, to a standard state that uses a concentration of 1 M in both the gas and aqueous phases.]

A thermodynamic cycle in Figure 2 was used for the calculations of pK_a values for a series of selenols (R−SeH, n = 6) and aryl selenenic acids (ArSeOH, n = 5) in aqueous solution.[61,63,70−73] Gibbs free energies were obtained by density functional theory (DFT) computations [B3LYP/6-31G(d,p)][74−76] using the Gaussian 03 package.[51] The solvent effect (aqueous solution) was estimated by the isoelectric focusing-polarizable continuum model (IEF-PCM).[69] Because of the lack of selenium parameters for the calculation of cavitation/dispersion/repulsion contributions, only the electrostatic component of the solvation energy was taken into account. The accuracy of the DFT methodology used was compared with DFT calculations with a larger basis set [6-31+G(d,p)[77,78] or aug-cc-pVDZ],[79] with an ab initio method [the second-order Møller−Plesset theory (MP2)][80] as well as with the another solvation model, CPCM variant using COSMO-RS radii (CPCM-COSMO-RS)[81] (keyword SCRF = COSMORS with the default setting of the Gaussian 03 package[51]). Experimental pK_a values are taken from the literature (selenols[2,82−84] and selenenic acids[85]).

The pK_a values of Sec45 in selenol and selenenic acid forms in GPx1 were calculated based on a simple structural model, in which Sec45, in a conformation identical to that adopted in the enzyme, was reduced to isolated selenocysteine (or its selenenic acid form), and the effects of the aqueous solution were simulated by the above-mentioned PCM models. For these purposes, the structure of GPx1 was optimized using an enzyme model (built from a snapshot selected from the MD trajectory of MD1_Se). The four enzyme redox states, E−SeH, E−Se$^-$, E−SeOH, and E−SeO$^-$, were optimized at the hybrid quantum mechanics/molecular mechanics (QM/MM) level [the ONIOM with mechanic embedding (ME) scheme,[67,68] B3LYP/6-31G(d,p):Amber][11,12] using the Gaussian 03 package[51] (Gaussian ONIOM outputs of the optimized geometries are available in the Supporting Information). We used the ME scheme since optimization of GPx3 with the ONIOM scheme with the electrostatic embedding (EE) did not improve the results and gave a slightly larger rms deviation when the ONIOM geometries were compared with the X-ray structure of the enzyme.[12] ONIOM-ME does not include the exact electrostatic interaction between the electron density and the point charges in the protein environment, however, it includes electrostatic effects due to geometrical changes of amino acid residues and molecules of aqueous solvent involved in the QM part of the system (in our case: Sec45, Gly46, Thr47, Gln80, Trp158, Arg177, and 4 molecules of H_2O). The entire enzyme structure contains 246 molecules of water located in the active site (taken from the MD simulations), from which 4 molecules (located around the Sec45 residue) were treated at the QM level and the rest at the MM level.

The optimized geometries of the above-mentioned four redox states of Sec45 in the enzyme model were taken for the subsequent pK_a calculations in the aqueous solution using the IEF-PCM[69] and CPCM-COSMO-RS[81] methods. The calculated pK_a values were corrected via a simple linear fit to a training set of the organoselenium compounds according to the formula: pK_a = k*[pK_a(raw)] + d (Table 4).

We also tried to evaluate effects of the protein environment on the calculated pK_a values of Sec45 at the ONIOM-ME and ONIOM-EE as well as at the ONIOM-EE//ONIOM-ME levels. [However, unfitted pK_a values of Sec45 based on the ONIOM schemes did not converge and had in some cases unreasonably high positive (>14, the calculations with basis set without diffuse functions) or low negative values (>−3, the calculations with the basis set augmented by the diffuse functions). Therefore, they will not be discussed in this work.]

3. Results and Discussion

3.1. Docking. The docking calculations were performed to provide missing structural information on both binding sites of GSH at the active site of GPx1. Because it is not clear at which redox state of GPx1 the two molecules of GSH bind into the active site, we used for our docking study GPx1 in its reduced selenol state. In the first experiment, we docked one molecule of the reduced GSH into a crystal geometry of the active site of GPx1 (Dock01_GSH). In the most favorable docking pose [with the best docking score of −5.4 kcal mol^{-1}, Figure 3b and Table S1 in the Supporting Information], GSH was positioned in a groove region, 7.5−11.0 Å apart from Sec45(A), where the thiol group of GSH was fitted in a hydrophobic area at the end of the groove [7.5 Å from the selenol group of Sec45(A)]. The hydrophobic pocket consists of Phe145(A), Trp158(A), and partially Arg177(A) [numbering according to the crystal structure of GPx1 (PDB ID: 1GP1)]. In the binding pose, a N-terminal amino group of the γ-glutamyl moiety of GSH interacts by hydrogen bonding with a side-chain carboxylate of Asp126 and the backbone carbonyl group of Ala129 (2.1 and 1.9 Å, see a 3-D structure in the Supporting Information), while a C-terminal carboxylate group of the glycin moiety of GSH interacts with Arg177 (1.9 Å). The tryptophan residue has been discussed by other authors[3,7,10,37,86] as an essential amino acid of the catalytic triad Sec−Gln−Trp in GPx, which modulates the selenium reaction center toward the feasible redox reactions. Based on our docking results, we believe that a catalytic Trp158 is mainly responsible for the proper binding of GSH. It provides the proper orientation of the thiol group of GSH close to the Sec45 reaction center. Therefore, it is evident from the docking pose that hydrophobic interactions between the tryptophan aromatic ring and the sulfur of GSH play a significant role in the redox mechanism of GPx1.

In the next experiment we docked another molecule of the reduced GSH into the active site of GPx1, with the first molecule of GSH already bound to the enzyme according to the previous docking calculations. GSH was docked into the enzyme either with a neutral (Dock02_GSH_GSH) or an ionized form (Dock03_GSH_GS) of its thiol group. For both ionization configurations similar docking poses for GSH were found. The second GSH was fitted in an enzyme region in the proximity of Sec45 (Figure 3a). The thiol group of GSH was placed into a small hydrophobic pocket located at the dimer contact surface near His79(B) (3.6 Å). (It should be noted that His79 from monomer A is not involved in the active site of this monomer, and instead is part of the active site of monomer B. And vice versa, His79 from monomer B is in the active site of monomer A.) Our results are in agreement with crystallographic experiments and their interpretation of the binding sites of GSH. When the reduced GPx1 was treated with an excess of GSH at pH 7, two binding sites of GSH per monomer of GPx1 were observed.[3] These binding sites were described by a positive, ill-defined, and noncontinuous difference electron density map (between Gpx1 and the complex GPx1:2GSH), with a maximum near to Sec45 and with a further density maximum stretching symmetrically from the dimer contact surface at the local axis [residues interpreted as His79(A) and His79(B)] into the inner part of each monomer.

An interatomic distance between thiol groups of the bound GSHs is 6.3 Å, while the distance between the thiol group of the GSH (interacted with Sec45) and the selenium atom

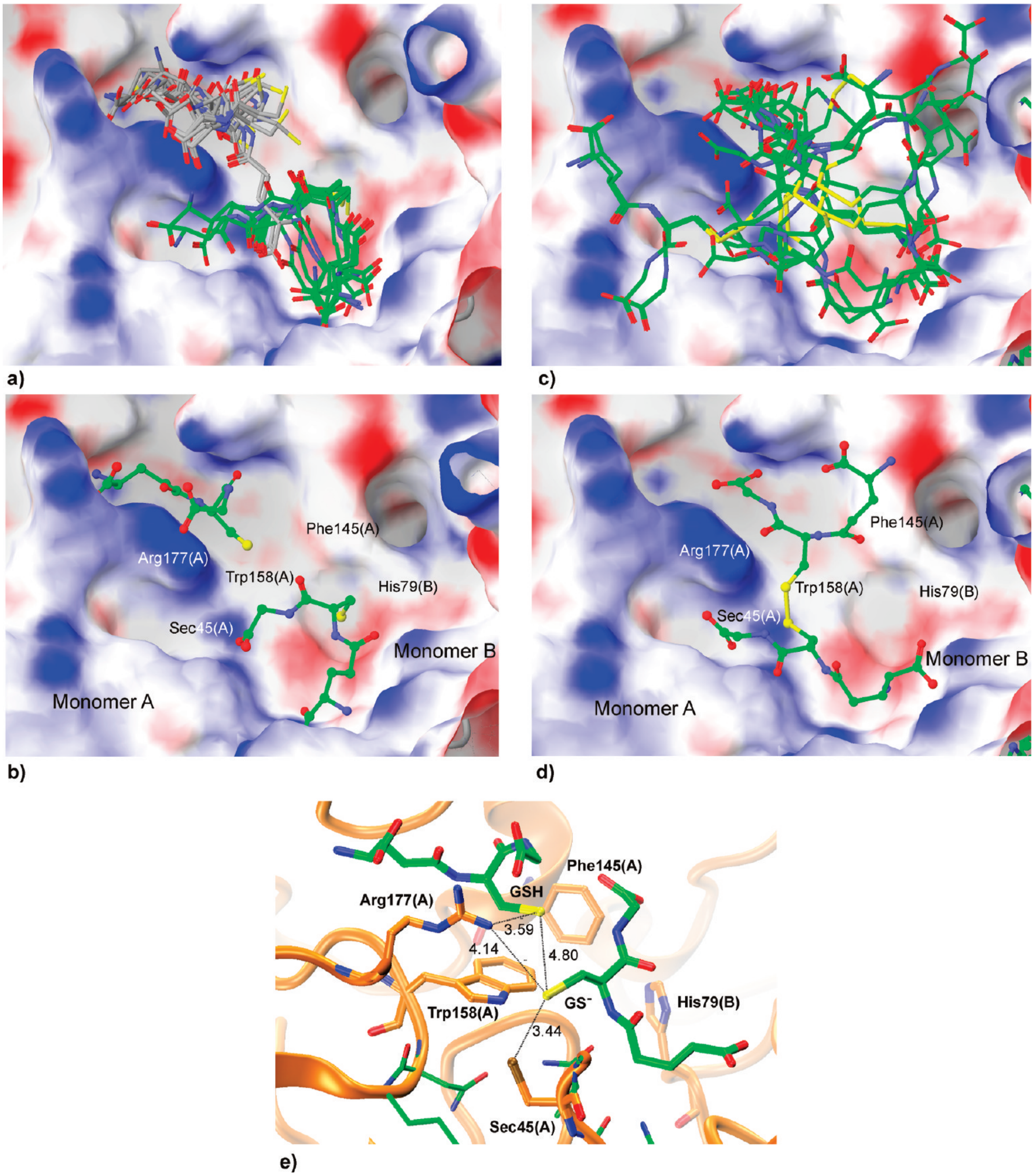


Figure 3. (a) An active site of GPx1 (visualized by surfaces) with two docked molecules of GSH (the best 10 poses visualized for every molecule of GSH). The GSH molecule (gray) placed close to Arg177(A) is from the Dock01_GSH, and GSH (green) positioned next to His79(B) is from the Dock02_GSH_GSH calculations. (b) Poses of both GSHs with the best docking score. (c) The best 10 poses of the docked GSSG. (d) A pose of GSSG with the best docking score. (e) The binding pose no. 16 of GS^- (Dock03_GSH_GS) with selected interatomic distances between sulfur (yellow), selenium (brown), and nitrogen (blue) atoms of GSH, GS^-, Sec45, and Arg177 (3-D structure in the Supporting Information). This pose could represent a starting reactive conformation for the second redox step of the GPx1 cycle.

of Sec45 is 8.0 Å. As can be seen from Figure 3b, after binding two molecules of GSH the active-site Sec45 residue is still not blocked (the analysis of the poses of GSHs with the best docking score) and could be oxidized to selenenic acid. Consequently, binding GSH prior to the formation of selenenic acid could protect Sec against the formation of higher oxidation states in an unwanted reaction with other molecules of the hydroperoxide substrate, i.e., the bound GSH can preferentially react with the selenenic acid intermediate and, thus, effectively compete with a hydroperoxide substrate. This assumption is indirectly supported by experiments with a radioactive GSH.[3] These demonstrated that GSH had been bound noncovalently to the reduced GPx1 and that its binding prior to the formation of any putative selanyl sulfide intermediate occurs via a reversible equilibrium as long as the system, consisting of the enzyme and GSH, was in the reduced state.[3]

Among the most preferable binding poses in the calculations with the ionized thiol group of GSH (pose no.16 with a docking score of -3.1 kcal mol^{-1} and no. 20 (-3.0 kcal mol^{-1}); for comparison the pose no. 1 possesses the best docking score of −4.2 kcal mol^{-1}], the binding poses of GSH with the thiolate oriented toward Sec45(A) (3.5 Å) still in vicinity of the thiol group of the first bound GSH (4.8 Å) and Arg177(A) (3.5 Å) [or His79(B) (4.8 Å)], were found (Figure 3e, see also 3-D structure in the Supporting Information). These poses could represent starting reactive conformations for the second and third redox step of the GPx1 cycle. The direct interaction between the side chain of His79(B) [or Arg177(A)] and the thiol group of GSH allowed us to consider them as catalytic residues involved in the redox mechanism of GPx1. They could assist in deprotonation of the thiol group of both GSHs to a more reactive thiolate (their role is also supported by our MD simulations and will be further discussed in the next section).

The docking study with oxidized glutathione (Dock04_GSSH) indicated a binding pose with the disulfide linkage of GSSG positioned close to Sec45 (4.5 Å) and placed in the vicinity of the Phe145-Trp158-Arg177 hydrophobic area (Figure 3d). GSSG was placed in the same regions as it was for the reduced GSHs in previous docking calculations but with a different position of its backbone (see 3-D structure in Supporting Information). For the oxidized GSSG, we observed a lower docking score compared with the reduced GSH (−6.5 kcal mol^{-1} for GSSG versus −10.1 kcal mol^{-1} for 2GSH). This is in agreement with the assumption that, after the last redox step, GSSG should be released from the active site to allow for binding with other molecules of hydroperoxide substrate and reduced GSH.

3.2. MD Simulations. The main goal of the simulations was to investigate the mobility of side chains of the active-site ionizing residues, mainly His79 and Arg177. These residues could facilitate deprotonation of the selenol as well as the thiol groups of Sec45 and GSH and increase the efficiency of the redox catalysis of GPx1. The MD simulations were performed on the enzyme dimer, and thus, we could simultaneously analyze conformational changes at the two active-sites of GPx1 and investigate a possible cooperation across the dimer interface. (It should be noted that both active sites in the dimer unit are located close to each other and separated by the contact region between the monomer units.)

In the crystal structure of GPx1, the active-site Sec45 residue (in the oxidized seleninic state, RSeOO$^-$) is located at the N-terminal end of an α-helix, which forms a $\beta\alpha\beta$ substructure together with the two adjacent parallel β strands. In most MD trajectories (in 21 of overall 26 active sites, Table 1), both Sec45(A) and Sec45(B) (or Cys45 in the case of the mutant) remained in the secondary structure analogous to the crystal geometry. (The same output was found for their selenenic acid and selenylsulfide forms.) However, in a few cases (Table 1), a flip of the side chain of Sec45 occurred with a concomitant shift of the last turn at the N-terminal end of an α-helix (Figure 4a). This conformational change is described in Figure 4b by the disruption of a hydrogen bond between the backbone carbonyl oxygen of Thr47 of the last turn and the amide hydrogen of Asp42 of the neighboring turn of the α-helix. The flip was only observed in one monomer unit in the simulations either with the empty active site or with one bound GSH and with neutral or ionized redox states of Sec45 (Table 1). It was not observed in the simulations with two reduced GSHs bound at the active site. It is not clear whether this indicates a biophysical process involved in the redox mechanism of GPx1 or is only a transient conformation of the enzyme.

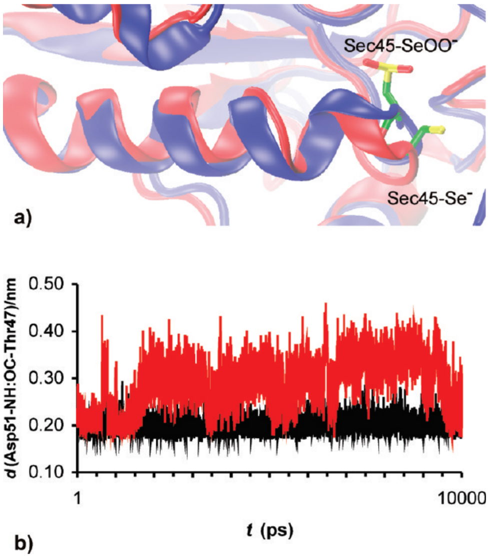


Figure 4. (a) A superposition of a trajectory snapshot (from MD10_SeSG, red) with the crystal structure of GPx1 (PDB ID: 1GP1, blue). A flip of Sec45(B) (green) was only observed in a few trajectories. (b) A flip of Sec45(B) described by changes in interatomic distances *d*[Asp51(B)−NH···O═C−Thr47(B)] (red) during 10 ns MD simulation (MD10_SeSG). The Sec45(A) maintains a starting conformation *d*[Asp51(A)−NH···O═C−Thr47(A)] (black).

The conformational analysis of the positions of the side chain of Arg177 indicates a possible catalytic role of this residue in the redox cycle of GPx1 (Figure 5). During simulations, the side chain of Arg177 exhibited high flexibility (Figure 5e, f, and g), and its guanidinium moiety frequently stayed in proximity of Sec45 (3−5 Å). It is known that the positive electrostatic potential of arginine or other residues, such as lysine, can stabilize the ionic form of ionizable residues and cause a decreasing pK_a value below 7 for residues, such as cysteine[38] (See also Figure 1d). Therefore, one catalytic role of Arg177 could be, besides the proper binding of GSH into the GPx1, the stabilization of the ionized selenolate of Sec45 and the thiolate state of GSH. Indeed, the side chain of Arg177 was often observed in the vicinity of the thiol group of GSH in our MD simulations (Figure 5). The importance of Arg residues in the reduction of hydrogen peroxide by a cysteine residue was recently demonstrated by QM calculations on the redox cycle of OxyR transcription factor.[38]

The His79(A) and His79(B) residues were found, both in the crystallographic measurements[3] and in our docking calculations, to interact with GSH. Their role is not clearly understood because they are not located in a reactive radius of Sec45(A) and Sec45(B) residues. They are part of the contact region of the monomer units, and the side chain of His79(A) is oriented toward Sec45(B), while His79(B) is toward Sec45(A). In 14 cases of 26 active sites analyzed,

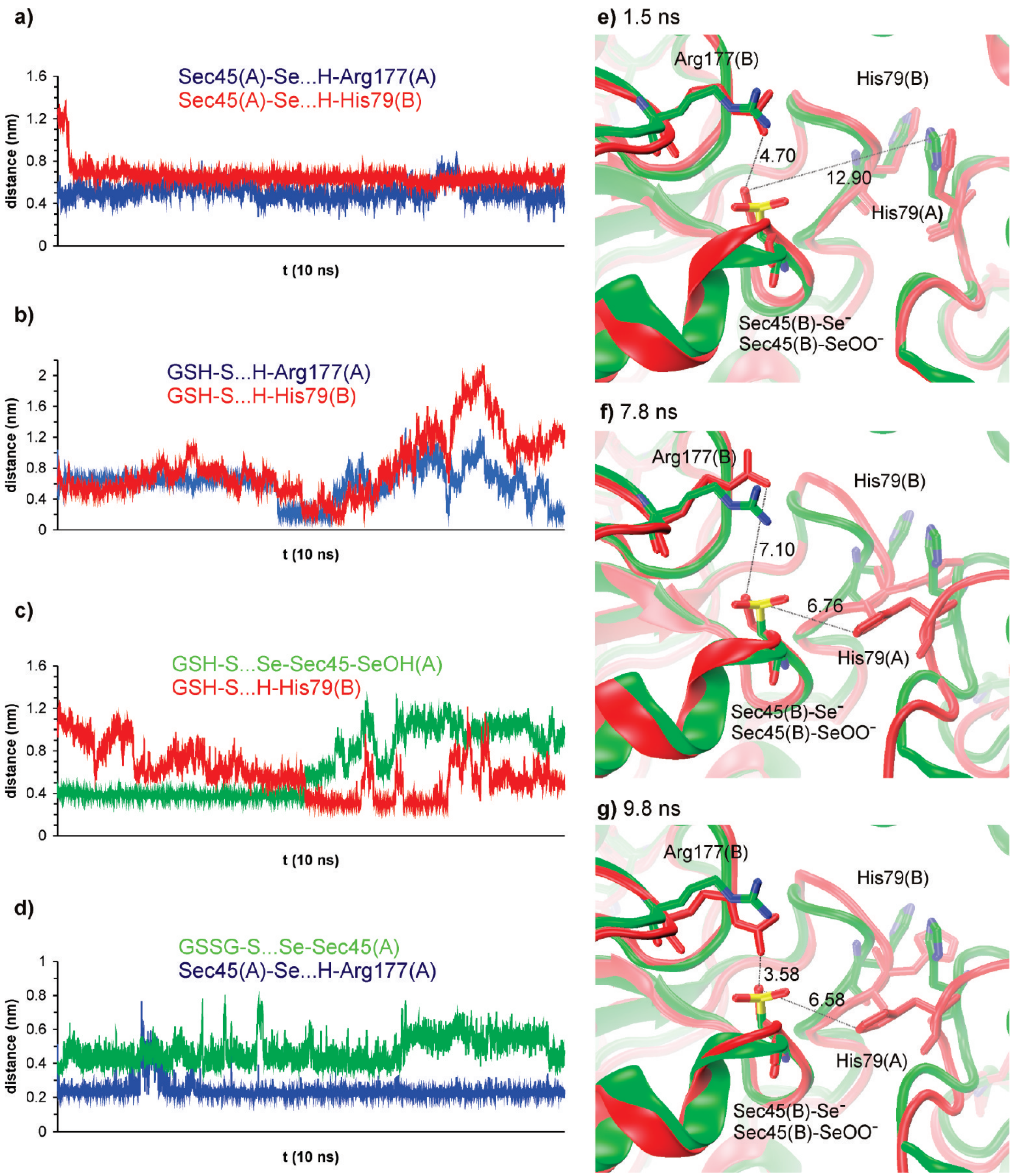


Figure 5. Interatomic distances between selected amino acid residues of GPx1 and GSH from the MD trajectories. (a) MD1_Se; (b) MD8_SeO_GSH_GS_HIP; (c) MD5_SeOH_GSH; and (d) MD12_GSSG. Snapshots: (e) 1.5, (f) 7.8, and (g) 9.8 ns from MD1_Se (red) are superimposed with the crystal structure of GPx1 (PDB ID: 1GP1, green). High mobilities of the side chains of Arg177(B) and His79(A) are evident.

side chains of His79(A) [or His79(B)] remained in conformations similar to those found in the crystal structures of GPx1 (Table 1). In the remaining active site, we observed the movement of either the side chains of these histidines (12 cases, Figure 5f and g) or the histidines within the conformational change of the Asn75−Gln81 loop (13 cases). This conformation change was observed in the simulations either with the empty active site or with bound GSH. It was not observed in the simulations with two reduced GSHs bound at the active site (monomer A in MD2_Se_GSH_GSH, MD7_SeO_GSH_GSH, MD8_SeO_GSH_GS_HIP, and MD9_SeOH_GSH_GS_HIP, Table 1). A detailed structural analysis has shown that His79(A) [or His79(B)] moves to the vicinity of Sec45 as a consequence of the conformational change of the Asn75−Gln81 backbone. However, it should be noted that our simulations were performed for the dimer structure of GPx1, which misses in vivo tetrameric contact region. In the tetramer structure, the Asn75−Gln81 sequence could be more stabilized in a position analogous to the crystal structure. Therefore, we also performed MD simulations on the large tetramer structure of GPx1 (a 10 ns simulation with Sec45 in the selenolate state without bound GSHs). The

Table 2. Comparison of Calculated and Experimental pK_a Values of Selenols[a]

compound	exptl.[b]	B3LYP[c]	MP2[c]	B3LYP[d]	B3LYP[e]	MP2[e]	B3LYP[f]
H_2Se	3.73	13.67	12.38	5.25	7.54	6.95	4.32
$(CH_3)_2N-(CH_2)-SeH$	4.74	19.31	17.25	9.51	11.36	10.24	9.21
$H_2N-CH_2-C(CH_3)_2-SeH$	5.21	19.76	17.80	10.84	12.43	10.40	9.52
$H_2N-(CH_2)_2-SeH$	5.01	18.67	17.08	10.27	12.36	10.90	9.53
$H_2N-(CH_2)_2-CH(CH_3)-SeH$	5.19	19.69	17.89	9.74	12.54	10.84	10.00
selenocysteine	5.24	18.26	16.38	8.41	10.42	9.27	8.07
r^2		0.83	0.84	0.74	0.75	0.71	0.79
k		0.23	0.26	0.25	0.26	0.33	0.24
d		0.66	0.63	2.61	1.95	1.65	2.80
SE		0.27	0.26	0.33	0.32	0.35	0.30
MAD		13.37	11.67	4.15	6.26	4.91	3.59

[a] SE is the standard error of the predicted pK_a; r^2 is the correlation coefficient; k is the slope; d is the intercept of correlation equation; and MAD is the mean value of ABS[pK_a(exp) − pK_a(calc)]. [b] Experimental pK_a values are from refs 2, 82, and 83. [c] IEF-PCM; 6-31G(d,p). [d] IEF-PCM; aug-cc-pVDZ. [e] IEF-PCM; 6-31+G(d,p). [f] CPCM-COSMO-RS; 6-31+G(d,p).

Table 3. Comparison of Calculated and Experimental pK_a Values of Selenenic Acids[a]

$2R^1$, $4R^2$−PheSeOH	exptl.[b]	B3LYP[c]	MP2[c]	B3LYP[d]	MP2[d]	B3LYP[e]	MP2[e]	B3LYP[f]
R^1=R^2=H	11.50	31.79	33.18	20.41	19.07	20.04	19.17	16.09
R^1=NO_2, R^2=H	10.45	27.82	31.13	17.37	17.55	17.07	18.43	13.07
R^1=NO_2, R^2=Cl	10.17	26.32	30.00	16.36	16.83	16.07	17.23	12.22
R^1=NO_2, R^2=CH_3	10.73	28.15	30.88	17.59	17.84	17.34	17.88	13.44
R^1=NO_2, R^2=CH_3O	10.83	28.13	31.40	17.57	−[g]	17.22	19.36	13.41
r^2		0.94	0.92	0.93	0.99	0.93	0.57	0.95
k		0.24	0.41	0.32	0.61	0.33	0.42	0.33
d		3.91	−2.12	5.05	−0.14	5.04	2.94	6.18
SE		0.14	0.16	0.15	0.08	0.15	0.38	0.13
MAD		17.71	20.58	7.12	7.11	6.81	7.68	2.19

[a] SE is the standard error of the predicted pK_a; r^2 is the correlation coefficient; k is the slope; d is the intercept of correlation equation; and MAD is the mean value of ABS[pK_a(exp) − pK_a(calc)]. [b] Experimental pK_a values from ref 85. [c] IEF-PCM; 6-31G(d,p). [d] IEF-PCM; aug-cc-pVDZ. [e] IEF-PCM; 6-31+G(d,p). [f] CPCM-COSMO-RS; 6-31+G(d,p). [g] Not converged.

Asn75−Gln81 loop in three subunits remained during the simulation in the position similar to the geometry of the crystal structure. In one subunit, a movement of the loop toward the active site, albeit in much less extent as it had been found in the simulations on the dimer of GPx1, was observed. This indicates: (i) the importance of the interactions between the two dimer units on the stability of GPx1; (ii) the observed conformational change of the Asn75−Gln81 loop is less frequent in the tetramer structure of GPx1 compared with our MD simulations on the dimer model of the enzyme; and (iii) the presence of two GSHs bound at the active site can probably stabilize the Asn75−Gln81 loop and His79 in a conformation similar to the crystal structure.

3.3. pK_a Calculations. Before calculating pK_a values for Sec45, calculations for a training set of organoselenium compounds, with available experimental pK_a values,[2,82−85] were performed to test the accuracy of the methodology used.

Absolute values of calculated pK_a substantially deviate from experimental data, trends for both selenols (Table 2) and selenenic acids (Table 3) are described with sufficient accuracy. Using larger basis sets augmented by diffuse functions significantly improved mean absolute deviations (MADs) by about 5−14 pK_a units, however, standard errors of the calculated pK_a values (SE = 0.26−0.27 pK_a units for RSeH and 0.14−0.16 for RSeOH) and r^2 (0.83−0.84 for RSeH and 0.92−0.94 for RSeOH) slightly impaired by less than 0.1 units for RSeH and 0.3 for RSeOH (Tables 2 and 3). The results do not change significantly when the DFT method is substituted by an ab initio method (MP2). Although using the CPCM-COSMO-RS solvation model instead of IEFPCM significantly lowers the MADs, slopes and correlation coefficients are hardly affected. Similarly, inclusion of diffuse functions in the basis set is essential to obtain reasonable MADs, however, correlation coefficients and slopes (k = 0.23−0.33 for both RSeH and RSeOH) are insensitive to either basis set or solvation model. As discussed by Kelly and co-workers,[63] implicit solvent models frequently give quite good correlations between calculated aqueous acid dissociation energies and experimental pK_a values. The slopes of the regression lines are generally too low (around 0.5−0.6), and the authors conclude that inclusion of explicit solvent molecules is necessary. For instance, Adam[87] has successively added water molecules until a slope of 0.88/RTln(10) for carboxylic acids could be attained. Kelly and co-workers[63] added a single explicit water molecule to anions with three or less atoms as well as to those with one or more oxygen atoms bearing a more negative partial charge than bare water. Table S2 in the Supporting Information presents the results obtained by adding one single H_2O to the selenolate anions. There is a slight improvement in the correlation coefficients, however, the slopes are virtually unchanged and the MADs increase. If we add a single H_2O only to those anions with a more negative charge on Se than on the water oxygen (data not shown), then even less agreement with experimental results is observed. A possible explanation for the low slopes might be the small spread of experimental pK_a values within both series of compounds. Combining the two sets into one single correlation equation results in somewhat larger slopes, e.g. k = 0.53 (IEFPCM-B3LYP/6-31G(d,p) (Table 4), comparable to those quoted

Table 4. Statistical Results Obtained for the Merged Training Set[a]

	IEF-PCM/ 6-31G(d,p)	IEF-PCM/ 6-31+G(d,p)	CPCM-COSMO-RS/ 6-31+G(d,p)
r^2	0.95	0.90	0.83
k	0.53	0.79	0.88
d	−4.62	−3.57	−1.93
SE	0.72	1.02	1.36
MAD	15.34	6.51	3.28

[a] SE is the standard error of the predicted pK_a; r^2 is the correlation coefficient; k is the slope; d is the intercept of correlation equation; MAD is the mean value of ABS[pK_a(exp) − pK_a(calc)].

Table 5. B3LYP Calculated Values of pK_a of Sec45 in Aqueous Solution Using the Correlations Coefficients from Table 4

	IEF-PCM/ 6-31G(d,p)	IEF-PCM/ 6-31+G(d,p)	CPCM-COSMO-RS/ 6-31+G(d,p)
Sec45−SeH	3.72 ± 0.72	3.07 ± 1.02	4.30 ± 1.36
Sec45−SeOH	11.43 ± 0.72	12.10 ± 1.02	11.71 ± 1.36

by Klamt et al.,[81] Chipman,[88] and Klicić et al.[89] Even larger slopes are obtained when including diffuse functions in the basis set [B3LYP/6-31+G(d,p)], 0.79 (IEFPCM) and 0.88 (CPCM-COSMO-RS). Note that these slopes are directly comparable to those quoted by Kelly et al. (0.76 without explicit solvent and 0.87 when one water molecule added to some of the anions).[63] However, it must be stressed that the main goal of this paper was not to calculate absolute pK_a values but rather to try to predict ionization states of Sec and its selenenic acid form in GPx1. Consequently, we used for the calculations of pK_a of Sec45 and its selenenic acid form the equation: pK_a(exp) = k × pK_a (calc) + d, with correlation coefficients developed for the merged training set of the organoselenium compounds (Table 4).

The pK_a values of the Sec45 residue were estimated at three levels, as they are compiled in Table 5. The estimated pK_a values of Sec45 in aqueous solution are significantly lower (ca. 3−4) compared with the calculated value for the selenenic acid form (ca. 11−12) or with an experimentally measured value of cysteine (8.3).[2,25] These values are in the range of experimentally measured pK_a values of selenocysteine in aqueous solution (5.3)[24] and the enzyme selenosubtilisin (pK_a < 4).[26] We should note that pK_a values calculated for Sec45 of GPx1 represent values for one enzyme conformation (see the Computational Details section). Due to the high mobility of the side chain of Arg177 as well as water molecules in a solvation shell of Sec45 observed in the MD simulations, the average pK_a values of Sec45 and its selenenic acid form could slightly deviate from the calculated values. However, these results allowed us to propose that at a physiological pH of GPx1[3] (with the pH optimum of 8.8 and 8.5 measured for bovine and human GPx[90,91]) ionized and neutral forms of Sec45 are in dynamic equilibrium, in which selenolate can be preferred for the redox reaction with hydroperoxides because of its higher reactivity as a nucleophile. In contrast, the selenenic acid form of Sec45 would prefer the neutral protonated state for the second redox reaction with GSH. While in the first reaction with hydroperoxides the selenium atom acts a nucleophilic center, in the second reaction with GSH, selenenic acid switches to electrophilic one. Consequently, the ionized form of the selenenic acid, i.e., its low pK_a value, would be undesirable to reach the electron-deficient selenium center necessary for attack of the GSH nucleophile.

Our results support the generally proposed redox mechanism of GPx1 (Figure 1a) that selenocysteine reacts with hydroperoxides as a selenolate and that a donor proton must be supplied. According to the MD simulations, the dyad Arg177−His79 could play the role of proton supplier in the redox cycle of GPx1 (Figure 1c).

Conclusion

Using three different tools of computational chemistry we have revealed some missing structural information concerning the redox cycle of GPx1. Based on pK_a calculations, the resting form of the catalytic Sec45 is allowed to be in selenolate (E−Se$^-$) at in vivo pH. This form is more reactive toward hydroperoxides,[17,18,92,93] thus increasing catalytic effectiveness of GPx1. Its transient selenenic acid form prefers the neutral state (E−SeOH) because Sec switches from nucleophile to electrophile in the second redox step of the GPx1 cycle. According to docking and MD simulations, Arg177, His79, Phe145, and primarily Trp158 seem to play a key role in the binding and proper orientation of the thiol group of GSH toward the selenocysteine reaction center. The MD simulations indicated that Arg177 and His79 could be involved in the catalytic cycle of GPx1, however, their exact chemical roles were not able to be deduced from the classical molecular mechanics (MM) simulations. We propose that Arg177 can maintain an increased positive electrostatic potential around the reaction center necessary for stabilizing the ionized states of Sec45 and GSH (Figure 1d), and together with His79 could be involved in proton transfer in the reaction steps of the GPx1 cycle (Figure 1c). To validate the redox mechanism of GPx1 with the Agr177−His79 catalytic dyad, further calculations at high quantum mechanics and QM/MM levels are needed.

Acknowledgment. Financial support for this research was granted by the Action Austria-Slovakia and the Centers of Excellence program of the Slovak Academy of Sciences (COMCHEM, Contract no. II/1/2007). J.K. acknowledges financial support from the Slovak Research and Development Agency, APVV (0117-06), and a VEGA grant 2/0176/09. We also thank to Dr. Otyepka from Palacký University in Olomouc for the support with programs for building and analyzing the structures from the ONIOM calculations.

Supporting Information Available: Gaussian input files of optimized structures, docked structures of the complexes GPx1:GSH:GSH, GPx1:GSH:GS, and GPx1:GSSG, an output structure from the PropKa calculations, force field parameters for the Sec residue, GSH and all their redox states. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Flohé, L.; Gunzler, W. A.; Schock, H. H. *FEBS Lett.* **1973**, *32*, 132.

(2) Johansson, L.; Gafvelin, G.; Arner, E. S. J. *Biochim. Biophys. Acta* **2005**, *1726*, 1.

(3) Epp, O.; Ladenstein, R.; Wendel, A. *Eur. J. Biochem.* **1983**, *133*, 51.

(4) Gettins, P.; Crews, B. C. *J. Biol. Chem.* **1991**, *266*, 4804.

(5) Maiorino, M.; Aumann, K. D.; Brigelius-Flohe, R.; Doria, D.; van den Heuvel, J.; McCarthy, J.; Roveri, A.; Ursini, F.; Flohe, L. *Z. Ernährungswiss.* **1998**, *37*, 118.

(6) Rocher, C.; Lalanne, J. L.; Chaudiere, J. *Eur. J. Biochem.* **1992**, *205*, 955.

(7) Ren, B.; Huang, W. H.; Akesson, B.; Ladenstein, R. *J. Mol. Biol.* **1997**, *268*, 869.

(8) Flohé, L.; Loschen, G.; Gunzler, W. A.; Eichele, E. *Hoppe-Seyler's Z. Physiol. Chem.* **1972**, *353*, 987.

(9) Mugesh, G.; du Mont, W. W.; Sies, H. *Chem. Rev.* **2001**, *101*, 2125.

(10) Prabhakar, R.; Vreven, T.; Morokuma, K.; Musaev, D. G. *Biochemistry* **2005**, *44*, 11864.

(11) Prabhakar, R.; Vreven, T.; Frisch, M. J.; Morokuma, K.; Musaev, D. G. *J. Phys. Chem. B* **2006**, *110*, 13608.

(12) Prabhakar, R.; Musaev, D. G.; Khavrutskii, I. V.; Morokuma, K. *J. Phys. Chem. B* **2004**, *108*, 12643.

(13) Bayse, C. A. *J. Phys. Chem. A* **2007**, *111*, 9070.

(14) Bayse, C. A.; Antony, S. *Main Group Chem.* **2007**, *6*, 185.

(15) Bayse, C. A.; Antony, S. *J. Phys. Chem. A* **2009**, *113*, 5780.

(16) Bayse, C. A. *J. Inorg. Biochem.* **2010**, *104*, 1.

(17) Cardey, B.; Enescu, M. *J. Phys. Chem. A* **2007**, *111*, 673.

(18) Cardey, B.; Enescu, M. *ChemPhysChem* **2005**, *6*, 1175.

(19) Bachrach, S. M.; Walker, C. J.; Lee, F.; Royce, S. *J. Org. Chem.* **2007**, *72*, 5174.

(20) Bachrach, S. M.; Demoin, D. W.; Luk, M.; Miller, J. V. *J. Phys. Chem. A* **2004**, *108*, 4040.

(21) Hayes, J. M.; Bachrach, S. M. *J. Phys. Chem. A* **2003**, *107*, 7952.

(22) Bachrach, S. M.; Woody, J. T.; Mulhearn, D. C. *J. Org. Chem.* **2002**, *67*, 8983.

(23) Bach, R. D.; Dmitrenko, O.; Thorpe, C. *J. Org. Chem.* **2008**, *73*, 12.

(24) Tan, K. S.; Arnold, A. P.; Rabenstein, D. L. *Can. J. Chem.* **1988**, *66*, 54.

(25) Huber, R. E.; Criddle, R. S. *Arch. Biochem. Biophys.* **1967**, *122*, 164.

(26) House, K. L.; Dunlap, R. B.; Odom, J. D.; Wu, Z. P.; Hilvert, D. *J. Am. Chem. Soc.* **1992**, *114*, 8573.

(27) Syed, R.; Wu, Z. P.; Hogle, J. M.; Hilvert, D. *Biochemistry* **1993**, *32*, 6157.

(28) Yang, Z.; Zhou, C. Z. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.* **2006**, *62*, 593.

(29) Scheerer, P.; Borchert, A.; Krauss, N.; Wessner, H.; Gerth, C.; Hohne, W.; Kuhn, H. *Biochemistry* **2007**, *46*, 9041.

(30) Claiborne, A.; Yeh, J. I.; Mallett, T. C.; Luba, J.; Crane, E. J.; Charrier, V.; Parsonage, D. *Biochemistry* **1999**, *38*, 15407.

(31) Poole, L. B.; Nelson, K. J. *Curr. Opin. Chem. Biol.* **2008**, *12*, 18.

(32) Salmeen, A.; Andersen, J. N.; Myers, M. P.; Meng, T. C.; Hinks, J. A.; Tonks, N. K.; Barford, D. *Nature* **2003**, *423*, 769.

(33) van Montfort, R. L. M.; Congreve, M.; Tisi, D.; Carr, R.; Jhoti, H. *Nature* **2003**, *423*, 773.

(34) Brandt, W.; Wessjohann, L. A. *ChemBioChem* **2005**, *6*, 386.

(35) Sarma, B. K.; Mugesh, G. *Inorg. Chem.* **2006**, *45*, 5307.

(36) Sandalova, T.; Zhong, L. W.; Lindqvist, Y.; Holmgren, A.; Schneider, G. *Proc. Nat. Acad. Sci. U.S.A.* **2001**, *98*, 9533.

(37) Roy, G.; Sarma, B. K.; Phadnis, P. P.; Mugesh, G. *J. Chem. Sci.* **2005**, *117*, 287.

(38) Kóňa, J.; Brinck, T. *Org. Biomol. Chem.* **2006**, *4*, 3468.

(39) Zheng, M.; Aslund, F.; Storz, G. *Science* **1998**, *279*, 1718.

(40) Kavanagh, K. L.; Johansson, C.; Smee, C.; Gileadi, O.; Von Delft, F.; Weigelt, C. J.; Sundstrom, M.; Edwards, A.; Oppermann, U. Crystal structure of the selenocysteine to glycine mutant of human glutathione peroxidase 1; The Research Collaboratory for Structural Bioinformatics (RCSB): RCSB-Rutgers, RCSB-San Diego Supercomputer Center, and RCSB-University of Wisconsin-Madison; http://www.rcsb.org/. Accessed May 22, 2009.

(41) *Glide*, version 4.5; Schrödinger, LLC: New York, NY, 2007.

(42) *MacroModel*, version 9.5; Schrödinger, LLC: New York, NY, 2007.

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

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

(45) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. *J. Med. Chem.* **2006**, *49*, 6177.

(46) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47*, 1739.

(47) *Maestro*, version 8.0; Schrödinger, LLC: New York, NY, 2007.

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

(49) van der Spoel, D.; Lindahl, E.; Hess, B.; van Buuren, A. R.; Apol, E.; Meulenhoff, P. J.; Tieleman, D. P.; Sijbers, A. L. T. M.; Feenstra, K. A.; van Drunen, R.; Berendsen, H. J. C. *Gromacs 3.3.1*; Department of Biophysical Chemistry, University of Groningen: Groningen, The Netherlands, 2004.

(50) Roothaan, C. C. J. *Rev. Mod. Phys.* **1951**, *23*, 69.

(51) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.;

Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision D.02 and E.01; Gaussian, Inc.: Wallingford, CT, 2005.

(52) Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431.

(53) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129.

(54) Bas, D. C.; Rogers, D. M.; Jensen, J. H. *Proteins* **2008**, *73*, 765.

(55) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704.

(56) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.

(57) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.

(58) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. *J. Comput. Chem.* **1997**, *18*, 1463.

(59) Schafmeister, C. E. A. F.; Ross, W. S.; Romanovski, V. *LEAP*; University of California: San Francisco, 1995.

(60) Case, D. A.; Darden, T. A.; Cheatham , T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *Amber 8*; University of California: San Francisco, 2004.

(61) Sadlej-Sosnowska, N. *Theor. Chem. Acc.* **2007**, *118*, 281.

(62) Wang, X. X.; Fu, H.; Du, D. M.; Zhou, Z. Y.; Zhang, A. G.; Su, C. F.; Ma, K. S. *Chem. Phys. Lett.* **2008**, *460*, 339.

(63) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 2493.

(64) Hudaky, P.; Perczel, A. *J. Phys. Chem. A* **2004**, *108*, 6195.

(65) Palascak, M. W.; Shields, G. C. *J. Phys. Chem. A* **2004**, *108*, 3692.

(66) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R. *J. Phys. Chem. A* **1998**, *102*, 7787.

(67) Vreven, T.; Byun, K. S.; Komáromi, I.; Dapprich, S.; Montgomery, J. A., Jr.; Morokuma, K.; Frisch, M. J. *J. Chem. Theory Comput.* **2006**, *2*, 815.

(68) Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *J. Mol. Struct. THEOCHEM* **1999**, *462*, 1.

(69) Tomasi, J.; Mennucci, B.; Cances, E. *J. Mol. Struct. THEOCHEM* **1999**, *464*, 211.

(70) Namazian, M.; Zakery, M.; Noorbala, M. R.; Coote, M. L. *Chem. Phys. Lett.* **2008**, *451*, 163.

(71) Lim, C.; Bashford, D.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 5610.

(72) Bryantsev, V. S.; Diallo, M. S.; Goddard, W. A. *J. Phys. Chem. A* **2007**, *111*, 4422.

(73) Schmidt am Busch, M.; Knapp, E. W. *ChemPhysChem* **2004**, *5*, 1513.

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

(75) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.

(76) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.

(77) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; von Rague Schleyer, P. *J. Comput. Chem.* **1983**, *4*, 294.

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

(79) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(80) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

(81) Klamt, A.; Jonas, V.; Burger, T.; Lohrenz, J. C. W. *J. Phys. Chem. A* **1998**, *102*, 5074.

(82) Sokolov, M. N.; Abramov, P. A.; Peresypkina, E. V.; Virovets, A. V.; Fedin, V. P. *Polyhedron* **2008**, *27*, 3259.

(83) Yokoyama, A.; Sakurai, H.; Tanaka, H. *Chem. Pharm. Bull.* **1970**, *19*, 1089.

(84) Kurz, J. L.; Harris, J. C. *J. Org. Chem.* **1970**, *35*, 3086.

(85) Kang, S. I.; Kice, J. L. *J. Org. Chem.* **1986**, *51*, 287.

(86) Sarma, B. K.; Mugesh, G. *Org. Biomol. Chem.* **2008**, *6*, 965.

(87) Adam, K. R. *J. Phys. Chem. A* **2002**, *106*, 11963.

(88) Chipman, D. M. *J. Phys. Chem. A* **2002**, *106*, 7413.

(89) Klicić, J. J.; Friesner, R. A.; Liu, S.-Y.; Guida, W. C. *J. Phys. Chem. A* **2002**, *106*, 1327.

(90) Wendel, A. *Methods Enzymol.* **1981**, *77*, 325.

(91) Rey, C.; Véricel, E.; Némoz, G.; Chen, W.; Chapuy, P.; Lagarde, M. *Biochim. Biophys. Acta* **1994**, *1226*, 219.

(92) Benková, Z.; Kóňa, J.; Gann, G.; Fabian, W. M. F. *Int. J. Quantum Chem.* **2002**, *90*, 555.

(93) Pearson, J. K.; Boyd, R. J. *J. Phys. Chem. A* **2007**, *111*, 3152.

CT9003355

JCTC Journal of Chemical Theory and Computation

A Statistical Framework for Hierarchical Methods in Molecular Simulation and Design

David F. Green*

Department of Applied Mathematics and Statistics and Graduate Program in Biochemistry and Structural Biology, Stony Brook University, Stony Brook, New York 11794-3600

Received August 24, 2009

Abstract: A statistical framework for performance analysis in hierarchical methods is described, with a focus on applications in molecular design. A theory is derived from statistical principles, describing the relationships between the results of each hierarchical level by a functional correlation and an error model for how values are distributed around the correlation curve. Two key measures are then defined for evaluating a hierarchical approach—completeness and excess cost—conceptually similar to the sensitivity and specificity of dichotomous prediction methods. We demonstrate the use of this method using a simple model problem in conformational search, refining the results of an *in vacuo* search of glucose conformations with a continuum solvent model. Second, we show the usefulness of this approach when structural hierarchies are used to efficiently make use of large rotamer libraries with the Dead-end Elimination and A* algorithms for protein design. The framework described is applicable not only to the specific examples given but to any problem in molecular simulation or design that involves a hierarchical approach.

1. Introduction

The ability to efficiently and robustly design molecules with particular characteristics is an objective targeted by many disciplines. Pharmaceutical development is largely a problem of designing a molecule that interacts specifically with a given protein target while maintaining additional characteristics (such as solubility in water, the ability to diffuse across plasma membranes, and stability to the varied environments of the body). Many applications in biotechnology similarly involve the development of proteins (most often through the modification of an existing sequence) that interact with specific targets or that catalyze particular reactions. Additional applications of molecular design abound, including materials engineering, nanotechnology, and catalyst development. While differing in the details of the design goal, all of these applications share the basic necessity of searching extremely large spaces of chemical and structural variation for individual molecular structures with the desired properties. In many cases, accurate computational models are available for the evaluation of individual molecules, but the immense size of the search spaces involved requires tradeoffs be made between computational efficiency and predictive accuracy.

Numerous biomedical and biotechnological applications have demonstrated a need for the ability to design new or modified proteins with particular stability and interaction properties.[1–3] It has been recognized that, by phrasing an inverse problem to the traditional protein-folding problem, great progress can be made in this area of practical protein design.[4,5] Typically in these approaches, a target backbone structure is chosen, and then a set of amino acid side-chain arrangements is selected to stabilize the target structure.[6] However, even given a fixed sequence length and backbone structure, the space of possible sequences to search is still immense. For a sequence of N residues, there are 20^N possible sequences—with as few as 10 variable residues, this is a space of over 10^{12} possibilities. When structural variability of each sequence is included (for example, by treating each residue as a set of commonly observed rotameric states), this complexity can easily grow to upward of 10^{26} (400 choices at 10 positions), and for larger designs, the search space can be greater than 10^{100}.

* Author phone: (631) 632-9344, fax: (631) 632-8490, e-mail: david.green@stonybrook.edu.

10.1021/ct9004504 © 2010 American Chemical Society
Published on Web 04/09/2010

Ligand docking—a key method for virtual high-throughput screening of pharmaceutical lead compounds—also involves a search over a huge structural space.[7,8] The size of the chemical space is essentially unbounded (even when "drug-like" restrictions are placed in molecular selection), and libraries of candidate molecules may easily contain hundreds of thousands of molecules. The docking process, which must be done on each molecule in a library, requires searching over six global degrees of freedom (three translational and three rotational) as well as any internal flexibility in the molecule and protein.[9,10] Again, even small systems can easily involve 10^{12} or more individual states.

Algorithmic advances, coupled with the rise in available computing power, have made these search problems tractable.[6,9] However, simplified descriptions of structural energetics are generally required for a number of reasons. In some cases, the large number of energetic evaluations required demands simplifications purely for computational tractability; for other methods, the algorithm itself requires an energetic description with particular properties (such as being decomposable into a sum of pairwise interactions between atoms or groups of atoms).

In parallel with developments in algorithms for searching large spaces of chemical and structural diversity, methodological advances have shown a remarkable ability to reproduce important experimental values, such as binding free energies and the effects of mutation on protein stability. However, these methods can be costly—free energy perturbation simulations using explicit solvent models being a perfect example.[11–13] Approximations can be made for efficiency, but with costs in accuracy. The frequently used Generalized-Born (GB) solvation model, for example, is orders of magnitude faster than explicit solvent simulation but suffers from well-characterized inaccuracies.[14–17]

Thus, while the power of computational approaches has developed strongly, there remains a fundamental trade-off that must often be made between the accuracy of the model used and the ability to sample an adequate space to solve the problem at hand. One solution to this dilemma is the use of a hierarchy of models. An inexpensive (but relatively inaccurate) model may be used for an initial search, and top ranking solutions from this search may be passed on to a more expensive, but more accurate, treatment. This procedure may be repeated, with successively increasing accuracy and expense in the models used. For most applications, a final level of the hierarchy is represented by experimental testing and validation.

Conceptually, the hierarchical approach is simple and has been applied in numerous applications.[18–21] However, there remains one important issue with a hierarchical approach—how do the cutoffs chosen at each level of the hierarchy affect the final results? Here, we present a statistical framework to help in answering this question. First, we describe the underlying statistics that describe the transfer of distributions with varying cutoffs. We use this framework to define two key descriptors of the efficacy of a hierarchical procedure: the completeness of the final set of results and the excess work done in calculations ultimately excluded from this set. The applications of this method is then outlined using a simple problem involving a conformational search with and without consideration of the solvent. Finally, these measures are used to consider the performance of hierarchical methods for an example application in protein design.

2. Theory

Here, we outline the fundamental statistical theory that can be used to characterize the performance of hierarchical methods. In this context, two levels of a hierarchy may be considered to be two energetic models for the same system; the levels may differ in the level of structural detail or in the Hamiltonian used. For example, levels may involve coarse-grained and fully atomistic descriptions of molecular structure, molecular mechanical and quantum mechanical Hamiltonians, or various treatments of the solvent. At each level, a "state" denotes an entity that can be associated with a single energetic value; these may be true structural microstates (such as a single conformation of a molecule) or an ensemble of microstates described by a free energy. A key requirement, however, is that the set of states at one level of the hierarchy be uniquely mapped to equivalent states at the next level.

We begin with a detailed outline of the theory; this section is purposefully constructed to be very general in scope and is thus somewhat abstract in its presentation. However, it may help the reader to consider that the ultimate goal in applying these methods will be to derive system-dependent functional relationships between hierarchical levels, and to use these to gain insight into the real-world performance of these approaches. Such applications are discussed in more detail in the Results and Discussion section. It should also be noted that this section is written using a formalism of continuous probability distributions; in many applications, including those presented as examples, the distributions will involve a finite number of discrete states. While a discrete variation of this theory can be easily derived (essentially by replacing all integrals by the appropriate sums), it is not clear that there is a significant motivation to do so.

Relating an Ensemble between Two Models. Consider an ensemble of states in a reference model, $\{\mathrm{E}^0\}$, and in a target model, $\{E^1\}$, where the distribution of states in the reference model is given by $P(E^0)$. If the energies in model 1 are correlated with those in model 0, then for all states with a given energy in model 0 ($\{E_i^0\}$), the energies of those states in model 1 ($\{E_i^1\}$) will distributed according to some error model, $P(E_i^1)$, about an expectation energy of $\bar{E}_i^1$. The expected energy in model 1 should be related to the energy in model 0 by $\bar{E}_i^1 = f(E_i^0)$, where $f(x)$ is a monotonically varying function. Similarly, with no loss of generality, we may assume an arbitrary error model, written as $P(E_i^1) = g(E_i^1, \bar{E}_i^1, \vec{\mu}_i)$, where $g(x, \bar{x}, \vec{\mu})$ is some general probability distribution, described by one or more parameters, $\vec{\mu}$. These parameters may include the standard deviation, higher-order moments of the distribution, or other descriptors. Since the expected energy varies with the reference energy, as may the parameters of the error model, we write:

$$P(E^1|E^0) = g(E^1, f(E^0), \overrightarrow{\mu}(E^0)) \qquad (1)$$

The probability of a given (E^1, E^0) pair is then $P(E^1|E^0)P(E^0)$ or

$$P(E^1, E^0) = g(E^1, f(E^0), \vec{\mu}(E^0))P(E^0) \tag{2}$$

and the probability of any E^1 over the full distribution of $\{E^0\}$ is then given by

$$P(E^1) = \int_{-\infty}^{\infty} g(E^1, f(E^0), \vec{\mu}(E^0))P(E^0)\,\mathrm{d}E^0 \tag{3}$$

requiring knowledge only of the reference distribution, $P(E^0)$, and the forms of $f(x)$, $g(x)$, and $\vec{\mu}(x)$. $P(E^0)$ may be considered to be either a normalized probability distribution or an unnormalized distribution of states. This choice does not affect the generality of the discussion but does fix the interpretation of $P(E^i)$ for all levels, i.

Multiple Levels of Propagation. Using the distribution $\{E^1\}$ as a reference for a target distribution $\{E^2\}$ and an analogous logic gives

$$\begin{aligned}P(E^2) &= \int_{-\infty}^{\infty} P(E^1)\, g_1(E^2, f_1(E^1), \vec{\mu}_1(E^1))\,\mathrm{d}E^1\\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} P(E^0)\, g_0(E^1, f_0(E^0), \vec{\mu}_0(E^0))\,\mathrm{d}E^0\right) \times\\ &\quad g_1(E^2, f_1(E^1), \vec{\mu}_1(E^1))\,\mathrm{d}E^1\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(E^0)\, g_0(E^1, f_0(E^0), \vec{\mu}_0(E^0)) \times\\ &\quad g_1(E^2, f_1(E^1), \vec{\mu}_1(E^1))\,\mathrm{d}E^0\,\mathrm{d}E^1\end{aligned} \tag{4}$$

This may be extended to any number of levels of propagation:

$$\begin{aligned}P(E^N) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} P(E^0) \times\\ &\quad \prod_{i=0}^{N-1} g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\,\mathrm{d}E^0\,\mathrm{d}E^1\ldots\mathrm{d}E^{N-1}\\ &= \int_V P(E^0)\prod_{i=0}^{N-1} g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\,\mathrm{d}V\end{aligned} \tag{5}$$

where the integral $\int_V \mathrm{d}V$ is taken from $(-\infty,\infty)$ over all dimensions E^i with i in the range $[0, N-1]$.

Defining Ensemble Completeness. Equation 5 defines the propagation of entire ensembles from one hierarchical level through another. In practice, however, only a sampling of each distribution will be passed on from one level to the next. One ensemble of particular interest is that of the low-energy states at the highest level of the hierarchy, $\{E^N|E^N < E^N_{\text{cut}}\}$. From the full distribution, $P(E^N)$, the size of this ensemble is given by

$$\begin{aligned}D(E^N_{\text{cut}}) &= \int_{-\infty}^{E^N_{\text{cut}}} P(E^N)\,\mathrm{d}E^N\\ &= \int_{-\infty}^{E^N_{\text{cut}}}\int_V P(E^0)\prod_{i=0}^{N-1} g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\,\mathrm{d}V\,\mathrm{d}E^N\end{aligned} \tag{6}$$

Now, consider what happens if each level of the hierarchy is truncated at some maximal value, E^i_{cut}. In this case, the propagation of the integrals is not done over the full space of $(-\infty,\infty)$ for each dimension. Rather, for each dimension i, the integral is taken over the range $(-\infty, E^i_{\text{cut}})$:

$$\begin{aligned}P'(E^N) &= \int_{-\infty}^{E^{N-1}_{\text{cut}}}\int_{-\infty}^{E^{N-2}_{\text{cut}}}\cdots\int_{-\infty}^{E^0_{\text{cut}}} P(E^0) \times\\ &\quad \prod_{i=0}^{N-1} g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\,\mathrm{d}E^0\,\mathrm{d}E^1\ldots\mathrm{d}E^{N-1}\\ &= \int_{V'} P(E^0)\prod_{i=0}^{N-1} g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\,\mathrm{d}V\end{aligned} \tag{7}$$

where V' represents the reduced space due to truncation of each distribution. Note that $P'(E^N)$ is not normalized, and thus its integral over all energies is less than that of $P(E^N)$. Given this distribution, the total size of the ensemble of interest is given by:

$$\begin{aligned}D'(E^N_{\text{cut}}) &= \int_{-\infty}^{E^N_{\text{cut}}} P'(E^N)\,\mathrm{d}E^N\\ &= \int_{-\infty}^{E^N_{\text{cut}}}\int_{V'} P(E^0) \times\\ &\quad \prod_{i=0}^{N-1} g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\,\mathrm{d}V\,\mathrm{d}E^N\end{aligned} \tag{8}$$

Given the size of the complete ensemble and that of the ensemble obtained through subsequent levels of truncation, we may define the ensemble completeness to be

$$C(E^N_{\text{cut}}) = \frac{D'(E^N_{\text{cut}})}{D(E^N_{\text{cut}})} \tag{9}$$

$C(E^N_{\text{cut}})$ gives the fraction of the complete low energy ensemble that is propagated through the truncated levels of the hierarchy. The completeness is analogous to the sensitivity (the true positive rate) of a dichotomous prediction method.

While $C(E^N_{\text{cut}})$ describes the completeness of the final ensemble, a second expression can be defined to describe the "excess work" required to obtain that level of completeness. At a given level in the hierarchy, the total size of the distribution carried through is given by

$$\begin{aligned}D''(E^N) &= \int_{-\infty}^{\infty} P'(E^N)\,\mathrm{d}E^N\\ &= \int_{-\infty}^{\infty}\int_{V'} P(E^0) \times\\ &\quad \prod_{i=0}^{N-1} g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\,\mathrm{d}V\,\mathrm{d}E^N\end{aligned} \tag{10}$$

However, only $D'(E^N_{\text{cut}})$ of this is valuable—either for carrying on to the next level of the hierarchy or for inclusion in the final ensemble. Thus, we define the "excess work" to be

$$X(E^N_{\text{cut}}) = \frac{D''(E^N) - D'(E^N_{\text{cut}})}{D'(E^N_{\text{cut}})} \tag{11}$$

That is, the excess work is the relative amount of time spent evaluating "kept" and "discarded" states. $X(E^N_{\text{cut}})$ may be arbitrarily large but will equal 0.0 if no false positives are carried along and will equal 1.0 if equal numbers of false and true positives are found. The excess work is related to the false-positive rate, or specificity, of a dichotomous prediction method but is scaled relative to the true-positive rate to account for cost. Note that this definition of excess work does not directly relate to a computational cost estimate, as there is no explicit consideration of the relative expense of each step. An ideal search method, applied directly to the final energetic ensemble, would require only evaluation of the energies of the lowest energy states; the excess work describes how much extra effort must be put into evaluation of energies at the higher hierarchical level, compared to this ideal. Note that this measure does not consider the cost of the search at the lower level.

Propagation in Normal Error Models. Consider the case where $\bar{E}^1$ is linearly correlated with E^0 ($\bar{E}^1 = m_1E^0 + b_1$), and the error distribution of $\{E_i^1\}$ about $\bar{E}_i^1$ is a normal distribution with constant variance ($\vec{\mu}_1 = \{\sigma_1\}$, independent of E^0). Thus, eq 1 becomes

$$P(E^1|E^0) = \frac{1}{\sqrt{2\pi}\sigma_1}\mathrm{e}^{\frac{-(E^1 - (m_1E^0 + b_1))^2}{2(\sigma_1)^2}} \tag{12}$$

If $\{E^0\}$ is normally distributed about E^0 with a standard deviation of σ_0, then eq 2 becomes:

$$P(E^1, E^0) = \frac{1}{(2\pi)\sigma_0\sigma_1}\mathrm{e}^{\frac{-(E^1 - (m_1E^0 + b_1))^2}{2(\sigma_1)^2}+\frac{-(E^0 - \bar{E}^0)^2}{2(\sigma_0)^2}} \tag{13}$$

The zero points of the energy distributions are arbitrary, and thus we may make the simplifying assumption of $b_1 = 0$ and $E^0 = 0.0$. Furthermore, a correlation with slope m and variance σ is equivalent to a correlation with unit slope and variance $\sigma' = \sigma/m$, allowing the further simplification of $m_1 = 1.0$. This gives

$$P(E^1, E^0) = \frac{1}{(2\pi)\sigma_0\sigma'_1}\mathrm{e}^{\frac{-(E^1 - E^0)^2}{2(\sigma'_1)^2}+\frac{-(E^0)^2}{2(\sigma_0)^2}} \tag{14}$$

which integrates (eq 3) to

$$P(E^1) = \frac{1}{\sqrt{2\pi}\sqrt{(\sigma_0)^2 + (\sigma'_1)^2}}\mathrm{e}^{\frac{-(E^1)^2}{2((\sigma_0)^2+(\sigma'_1)^2)}} \tag{15}$$

$P(\mathrm{E}^1)$ is a normal distribution, with variance $(\sigma_0)^2 + (\sigma'_1)^2$. This can be extended simply through multiple levels to give $P(E^N)$ as a normal distribution with variance $\sum_{i=0}^{N}(\sigma'_i)^2$. Truncating the integral at E^0_{cut} gives

$$P'(E^1) = \frac{1}{2\sqrt{2\pi}\sigma'_{01}}\mathrm{e}^{\frac{-(E^1)^2}{2(\sigma'_{01})^2}}\left[1 + \mathrm{erf}\left(\frac{E^0_{\mathrm{cut}}\sigma'_{01}}{\sqrt{2}\sigma_0\sigma'_1} - \frac{\sigma_0E^1}{\sigma'_{01}}\right)\right] \tag{16}$$

where $\sigma'_{01} = [(\sigma_0)^2 + (\sigma1)2]1/2$. The general form of $\int \mathrm{e}{-}x2\mathrm{erf}(ax + b)\,\mathrm{d}x$ can not be analytically determined, and thus propagation of the truncated set, as well as determination of D' and D'', must be done numerically.

Propagation of a Sampled Low-Energy Distribution. The above treatment was based on obtaining *all* members of the ensemble below a given E^{cut}. While a number of methods are designed to give this set deterministically, other methods yield a set that is enriched in low-energy states but is not guaranteed to give all states within any energy cutoff. The same definitions of completeness and excess cost can be applied to these methods. However, the description of how the ensembles of states are passed through the hierarchy is different. Consider a sampling algorithm that, given some uniform distribution over a variable x, yields a distribution $Q(x)$. The sampled distribution of $P'(E^i)$ will then be given by $P''(E^i) = P'(E^i)\,Q(E^i)$. The distribution obtained by propagating this distribution up a level of the hierarchy is given by integrating (over all energies) the product of this distribution with the correlation function:

$$\begin{aligned} P'(E^{i+1}) &= \int_{-\infty}^{\infty} P''(E^i)g_i(E^{i+1},f_1(E^i),\vec{\mu}_i(E^i))\,\mathrm{d}E^i \\ &= \int_{-\infty}^{\infty} P'(E^i)Q(E^i) \times \\ &\qquad g_i(E^{i+1},f_1(E^i),\vec{\mu}_i(E^i))\,\mathrm{d}E^i \end{aligned} \tag{17}$$

This may be extended to any number of levels of propagation in various ways. First, the sampled distribution may be passed on using another sampling-based algorithm (possibly the same one), in which case:

$$\begin{aligned} P'(E^N) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} P(E^0) \times \\ &\qquad \prod_{i=0}^{N-1} Q(E^i)g_i(E^{i+1},f_i(E^i),\vec{\mu}_i(E^i))\,\mathrm{d}E^0\,\mathrm{d}E^1\dots\mathrm{d}E^{N-1} \\ &= \int_V P(E^0)\prod_{i=0}^{N-1} Q(E^i)g_i(E^{i+1},f_i(E^i),\vec{\mu}_i(E^i))\,\mathrm{d}V \end{aligned} \tag{18}$$

where the integral $\int_V \mathrm{d}V$ is taken from $(-\infty, \infty)$ over all dimensions E^i with i in the range $[0, N - 1]$. The infinite domain of integration indicates that the entire sampled distribution is used as the input for each subsequent step. However, an alternative is to pass on only the lowest-energy states from the sampled distribution. In this case, the result is analogous to eq 7:

$$P'(E^{i+1}) = \int_{-\infty}^{E^i_{\mathrm{cut}}} P'(E^i)\,Q(E^i)g_i(E^{i+1},f_1(E^i),\vec{\mu}_i(E^i))\,\mathrm{d}E^i \tag{19}$$

$$\begin{aligned} P'(E^N) &= \int_{-\infty}^{E^{N-1}_{\mathrm{cut}}}\int_{-\infty}^{E^{N-2}_{\mathrm{cut}}}\dots\int_{-\infty}^{E^0_{\mathrm{cut}}} P(E^0)\prod_{i=0}^{N-1} Q(E^i) \times \\ &\qquad g_i(E^{i+1},f_i(E^i),\vec{\mu}_i(E^i))\,\mathrm{d}E^0\,\mathrm{d}E^1\dots\mathrm{d}E^{N-1} \\ &= \int_{V'} P(E^0)\prod_{i=0}^{N-1} Q(E^i)\,g_i(E^{i+1},f_i(E^i),\vec{\mu}_i(E^i))\,\mathrm{d}V' \end{aligned} \tag{20}$$

where V' represents the reduced space due to truncation of each distribution. The total sizes of the ensembles of interest are still given by $D'(E^N_{\mathrm{cut}}) = \int_{-\infty}^{E_{\mathrm{cut}}{}^N} P'(E^N)\,\mathrm{d}E^N$ and $D''(E^N) = \int_{-\infty}^{\infty} P'(E^N)\,\mathrm{d}E^N$. It should be noted that the results given earlier for the truncation of an enumerated system are simply a special case of eqs 18 and 20 when $Q(E^i) = H(E^i - E^i_{\mathrm{cut}})$, the Heaviside step function.

3. Methods

Conformational Analysis of Glucose. All calculations were done using the CHARMM computer program[22] with parameters from the Carbohydrate Solution Force Field

(CSFF).[23] D-β-glucopyranose was built in a standard geometry ($^4C_1$), and conformational states were generated by rotation about each of the hydroxyl dihedral angles (C_i-O_i, $i = 1-4, 6$) as well as about the exocyclic C_5-C_6 dihedral. An exhaustive enumeration of states was performed, sampling each dihedral at intervals of 60°, for a total of roughly 46 000 states. Energies were evaluated with no cutoffs, both in a vacuum (with a uniform dielectric constant of 1) and with the GBSW implementation of the Generalized-Born implicit solvent model.[24] GBSW calculations used an internal dielectric constant of 1 and an external dielectric constant of 80, with the dielectric boundary defined by a set of radii that have been optimized for this use.[25] The scaling coefficients were set to standard values of $a_0 = 1.2045$ and $a_1 = 0.1866$, the molecular surface was used, and a smoothing length of 0.2 Å was applied.

Protein Design. All calculations were done starting with the minimized average structure from the NMR structure of Calmodulin bound to the M13 peptide from rabbit skeletal muscle myosin light chain kinase (Protein Data Bank ID 2BBM).[26] Hydrogen atoms attached to carbons were removed for consistency with the PARAM19 parameter set used in the calculations. The positions of hydrogen atoms attached to heteroatoms were reoptimized using the HBUILD facility of the CHARMM computer program.

Sequences compatible with a low-energy complex structure were selected using a discrete structural search. The Dunbrack and Karplus rotamer library[27] was used, augmented by rotamers at ±10° of χ_1 and χ_2 for each rotamer. The selected set of positions consisted only of basic and acidic residues: lysine and arginine were varied to Asp, Glu, Asn, and Gln, and aspartate and glutamate were varied to Lys, Arg, His, Asn, and Gln. The three protonation states of histidine were each considered as individual choices.

Energies for the initial search were calculated using the CHARMM computer program[22] with the PARAM19 polar-hydrogen force field.[22] A distance-dependent dielectric of $4r$ was used for electrostatic interactions. All energies were calculated relative to isolated model compounds of the variable side chains. Software written by Tidor and colleagues was used for the search.

Different levels of structural detail were considered at various stages in the design. A rotameric structure refers to a specific choice of an amino acid conformer at each position in the protein. In many cases, similar rotamers make similar interactions—to prevent this from complicating the search, the fleximer model of Mendes et al. was used.[28] Here, all "sub-rotamers" derived from enhanced sampling of a Dunbrack and Karplus rotamer were grouped into a single entity denoted as a (DK-)fleximer. The energies of interaction between fleximers were taken as weighted averages of the interaction energies between the component rotamers. The same approach was used to group all rotamers of a single amino acid into an entity we term a sequence-mer,[29] or all rotamers of a given amino acid with the same χ_1 and χ_2 into a $\chi_{1,2}$-fleximer.

The structural search involved a hierarchical process over these levels. The Dead-End Elimination (DEE) and A* algorithms[30–32] were first used over the space of fleximers (or sequence-mers) to find all fleximeric states with an energy within a given cutoff of the global minimum. DEE and A* were then used to find the lowest rotameric state for each low-energy fleximer; at this stage this problem is one of side-chain placement for a single sequence, not of sequence design. A maximum of 10 unique DK-fleximers were processed for each sequence, and the final result was a single, minimum-energy rotameric state for each low-energy sequence.

4. Results and Discussion

Theoretical Framework. Hierarchical approaches to molecular design are not new; rather, it has long been recognized that the vast space of chemical and conformational variations necessitates the use of approximate models for computational tractability, but that more rigorous calculations are required to achieve reasonable predictive capability with respect to experimental results. While hierarchical filtering procedures have been used in many applications, these have generally been constructed in an *ad hoc* manner—it is often not clear whether false negatives arise due to inadequate sampling at any given stage, or whether significantly more computational expense was applied than necessary. A mathematical framework for assessing the effectiveness of different hierarchical approaches would allow for a more rigorous consideration of these and other issues.

The Theory section described precisely such a framework, based on a statistical description of state distributions and correlations between hierarchical levels. In particular, a number of key descriptors were outlined. First, it was shown how a distribution of selected, low-energy states at one level of a hierarchy is transformed on moving up the chain, as well as how to generate an estimate of the expected density of low-energy states at the highest hierarchical level. Two scalar quantities of performance assessment were also defined: the completeness of the final solution and the excess work required to achieve the final result. Completeness provides a quantitative metric for the success of a hierarchical solution, describing the fraction of low-energy solutions found relative to those expected. Excess work provides a measure of efficiency, describing the relative amount of effort spent on discarded solutions relative to those kept in the final solution. For a given application, one of these may be more important than the other, or a balance of the two may be sought. These metrics provide a quantitative basis on which to address this balance.

Conformational Analysis of Glucose. As an example of how these methods may be applied, we consider a conformational analysis of D-β-glucopyranose (Figure 1a). This molecule overwhelmingly prefers the $^4C_1$ ring conformation, which places all substituents in an equatorial arrangement, and thus the primary degrees of conformational flexibility are the rotations of the exocyclic hydroxyls. This is a six-dimensional space of finite volume (360° for each angle), and thus an exhaustive evaluation of conformational states is feasible (for a moderate sampling of each dihedral). While the number of states may not be beyond enumeration, the computational cost of the evaluation of each state must also be considered, and an accurate model of the conformational free energy surface must take into account the solvent; for

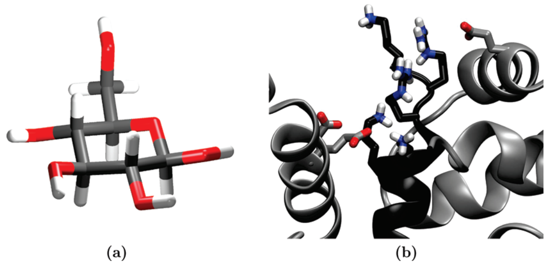


Figure 1. Model systems for application of statistical methods. (a) The minimum energy conformation of glucose found (in a Generalized-Born implicit solvent model) is displayed. (b) The residues of the CaM·M13 complex chosen for variation are displayed, using the minimized average solution structure. Calmodulin is shown in gray and M13 in black. Figures generated with VMD.[36]

most biological molecules, the environment of interest is that of an aqueous solution of moderate ionic strength. Among the most commonly used models for the inclusion of solvent effects are those based on continuum electrostatics: the Poisson–Boltzmann model and the Generalized-Born approximation.[33–35] As the computation of free energies in an implicit solvent model is significantly more costly than the corresponding calculation in a vacuum, one strategy for reducing the computation cost may be to screen the full space with a vacuum energy model and then refine the lowest energy states with a continuum model. However, the question arises as to how many low energy (vacuum) states need be considered in order to obtain an accurate description of the minimum-energy solvated states. This is precisely the question our technique aims to answer.

All six degrees of freedom were uniformly sampled at 60° intervals, giving a total of 46 656 distinct conformational states. The energy of each state was then evaluated with the CHARMM all-atom force field in a vacuum. These data follow a nearly perfect normal distribution ($R^2 = 0.9991$ for a nonlinear least-squares fit, see Supporting Information Figure S1) with a mean of 81.09 kcal/mol and a standard deviation of 6.42. All states with energies in the lowest 20 kcal/mol (8483 total, 18% of all states) were then selected for subsequent evaluation with a Generalized-Born (GB) solvent model;[24] the correlations of these two data sets are shown in Figure 2a. The two energies are correlated (although not strongly in a linear sense, $R^2 = 0.48$) and give a linear best fit with a slope of 0.62. However, for the application of the statistical analysis, we need to obtain an error model for the GB energies as a function of the vacuum energy.

The vacuum energies were divided into 1 kcal/mol bins, and the GB energies of all states in each bin were fit to a normal distribution (see Figure 2b and Supporting Information Figure S2). Several observations can be made that clearly demonstrate the applicability of the statistical model. First, in all cases, the fit to a normal error model was reasonable ($R^2 > 0.7$), and the fit was excellent ($R^2 > 0.96$) in every bin containing at least 200 states. Second, the mean GB energy in each bin is highly correlated with the vacuum energy, with an R^2 of over 0.99. Finally, for bins with a significant population, the standard deviation of GB energies in the bin is roughly constant, with a mean value of 1.97. These results motivate the use of a quite simple error model: (1) the expectation value of the GB energy varies linearly with the vacuum energy; (2) the distribution of GB energies around the expectation value follows a normal distribution of constant variance.

While in this case the full underlying distribution of vacuum energies is known, in many applications, it may not be. Thus, we additionally considered how well the full distribution may be fit using only those data within the lowest 20 kcal/mol. As clearly shown in Figure 2c and d, the fit is excellent, giving a mean of 81.00 (compared to the true mean of 81.09) and a standard deviation of 6.28 (true value, 6.42). Thus, strictly using the ensemble of low vacuum energies is reasonable.

Given the vacuum energy distribution and the error model for how GB and vacuum energies correlate, 10 000 model distributions were generated. As can be seen in Figure 2e, the distribution of model data matches the observed data (where known) very well. The model distributions were then used to estimate how well various low-GB-energy ensembles are captured; the lowest 2, 4, 6, 8, and 10 kcal/mol ensembles (relative to the lowest observed GB energy) were all considered. For each of these ensembles, the completeness and excess cost were computed as a function of the vacuum-energy cutoff used (see Figure 2f). With the actual cutoff applied to the data (20 kcal/mol), the 10 kcal/mol ensemble of GB states is found to have a completeness of roughly 90%, and the lower-energy ensembles are all near 100% complete.

In all these cases, however, the full 20 kcal/mol of low-vacuum-energy states had to be evaluated in the GB model, and many of these may not have been in the final low-energy ensemble; this is measured by the excess cost. For the 90% complete 10 kcal/mol ensemble, the excess cost is roughly 1, meaning half of the states that were evaluated were not part of the final low energy ensemble. On the other hand, the excess cost for the 2 kcal/mol ensemble is above 100; less than 1% of the evaluated states were part of the final set. This suggests that a cutoff of 20 kcal/mol in vacuum energy is inefficient if only the very lowest GB energies are desired; using a 15 kcal/mol cutoff would still give a near-perfect completeness, but with an excess cost 10-fold less.

To consider in more detail what this analysis provides, consider the middle of the ensembles considered (red curves); these are all states with GB energies within 6.0 kcal/mol of the global minimum and thus includes all states that would be populated more than 0.01% at room temperature. When only the lowest 6 kcal/mol of vacuum energies are considered, the completeness of this ensemble is only 7.6%; increasing this to the lowest 10 kcal/mol of vacuum energies increases the completeness to 39%, and thus many states are still missed. However, a 15 kcal/mol cutoff in vacuum energy gives a completeness of 91%, and the 20 kcal/mol cutoff gives a completeness of 99.89%.

To obtain this most complete ensemble required an excess cost of 16; for each state that was evaluated and “kept” in the low-energy ensemble, 16 were “discarded” as being outside the desired range. In other words, using the hierarchi-

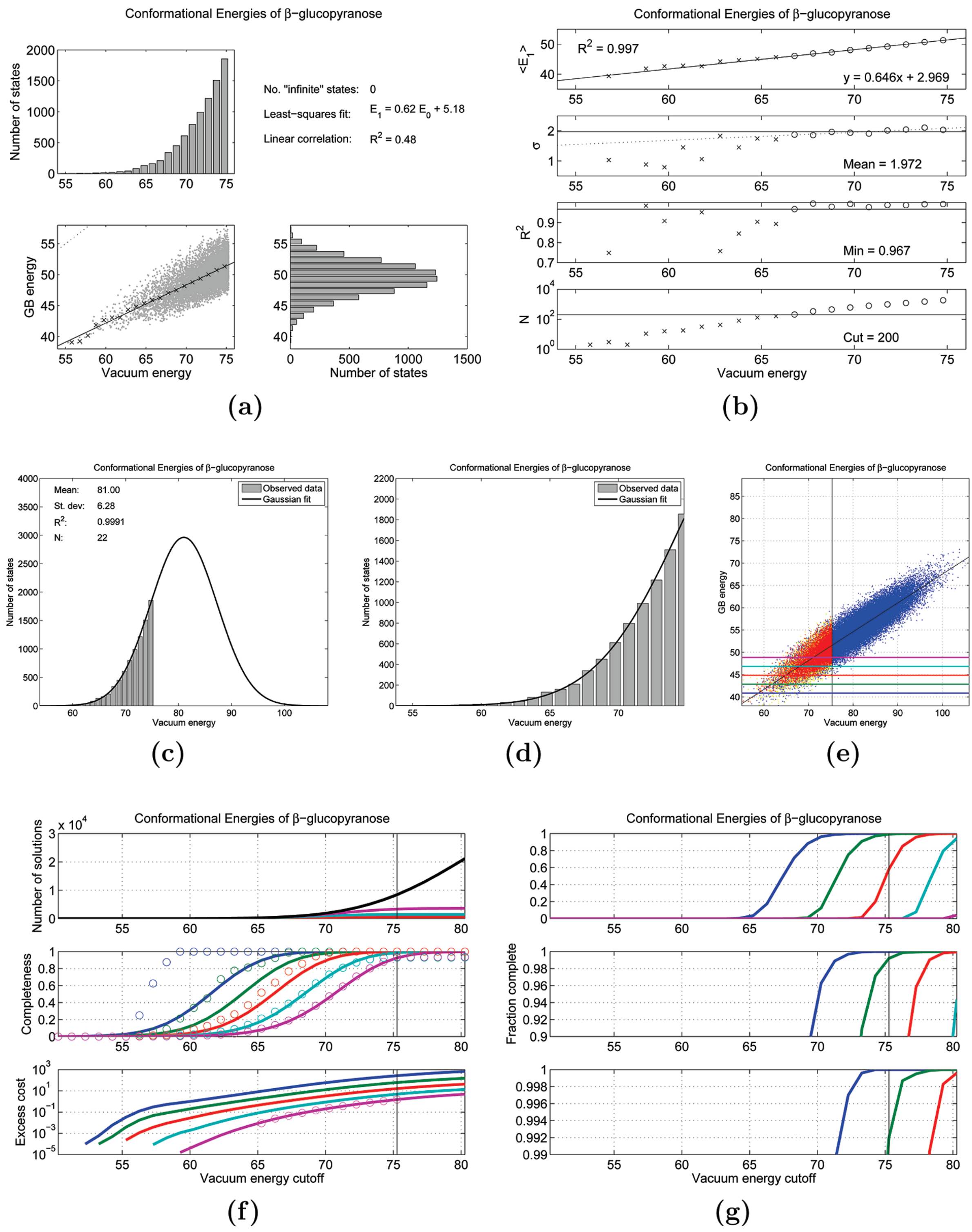


Figure 2. Conformational energetics of *β*-glucopyranose. Details of a conformational analysis of glucose are shown. (a) Correlations between Generalized-Born and vacuum energies for the lowest 20 kcal/mol of vacuum energy states. (b) Details of normal distributions fit to the GB energies of states having vacuum energies within 1.0 kcal/mol bins. $\langle E_1 \rangle$ is the mean GB energy, σ is the standard deviation, R^2 is the proportion of the variance described by the fit, and N is the number of data points in each bin. Bins of at least 200 points are indicated by open circles; these bins were used to evaluate the linear best fit equation of the mean, the mean standard deviation, and the minimum R^2 of the fits. (c, d) The distribution of low-vacuum energy states, fit to a normal distribution; d shows the same data with focused axes. (e) Model data, generated from the fit parameters of b and c. Blue points indicate model data and red, the actual data (yellow points are a model of GB energies from actual vacuum energies). The horizontal lines denote cutoffs in GB energy of 2, 4, 6, 8, and 10 kcal/mol, from the lowest actual value. (f, g) Metrics of performance for each GB-energy cutoff (colors match those of the cutoffs marked in e. In f, the total number of solutions, completeness, and excess cost from both model (lines) and actual (circles) data are shown. In g, the fraction of cases in which 100% completeness was achieved (out of 10 000 model calculations) is shown; each panel shows an increasingly focused *y*-axis range.

cal approach required evaluation of 16 times more states than were actually found. For the 91% complete ensemble, the excess cost drops to 3.6; 25% of the evaluated states are, in fact, part of the final solution.

Now, what does a 99.89% completeness really mean? In this case, we have 393 states in the low-GB-energy ensemble, and thus a 99.89% completeness suggests that we expect to have missed roughly "half a state". A complete enumeration of the space confirms that the ensemble is, in fact, 100% complete. The confidence in whether we found all states was assessed by considering the fraction of the 10 000 model distributions that led to a *perfectly* complete set of low GB energy states at a given vacuum energy cutoff (see Figure 2g); for a 20 kcal/mol vacuum cutoff, this value is 58.8% for the 6 kcal/mol GB ensemble. This can be interpreted as the confidence value that the ensemble is truly complete and thus provides one of the most important results of the statistical analysis. If a given level of confidence in the completeness of the solution is desired, the analysis also gives this: for a 96% confidence level, a vacuum energy cutoff of 22 kcal/mol should be used, and a solution that is complete to over 99.9% confidence can be found with a 25 kcal/mol cutoff.

A brief discussion of computational costs and the meaning of excess work is appropriate. As noted above, the excess work is not a direct measure of computational expense, but rather a measure of how many states selected at one level of the hierarchy (in this case, the vacuum energy) were not included in the final set of solutions (low-GB-energy states). The actual costs involve two primary contributions: (1) the cost of the first-level search and (2) the cost of re-evaluation in the second level. In an ideally performing system, the second step would only involve those states that are, in fact, low energy; the excess cost describes, in relative terms, how much more effort must be expended in the second step than this ideal bound.

In this case, the full conformational search over vacuum energies took 5.59 min, for a total of 0.0072 s per state, on a single 3.40 GHz Intel Xeon processor; due to software overhead, this is reduced to 0.0048 s per state when a large number of states are considered, for example by finer sampling. The GB calculations are slightly less than twice as costly, taking 0.013 s per state, and thus to perform the complete grid search requires 9.93 min. To evaluate the lowest 20 kcal/mol of vacuum energy states with GB, however, only requires 1.81 min, thus making the net time for the hierarchical search 7.40 min, or 75% of the exhaustive search time using GB. To achieve a 96% or 99.9% confidence in the completeness would require 8.32 or 10.05 min, respectively (84% or 101% of the exhaustive search cost).

Of course, these results suggest that there is minimal motivation for the use of a hierarchical method. Part of this arises from the relatively small cost differential of the two methods; in this small system, computing a GB energy is only fractionally more expensive than the corresponding vacuum energy. However, in more typical applications involving large biological macromolecules, GB is roughly 4-fold more costly. With this cost difference, the hierarchical approach would give 99.9% confidence in a complete low-energy GB ensemble at 70% of the cost of an exhaustive search; 96% confidence would be attained with 52% of the cost.

It should be noted that the model distributions give notably divergent results from the actual data for the very lowest GB-energy ensembles. There are only eight states in the lowest 2 kcal/mol, and only 80 in the lowest 4 kcal/mol. For samples of this small size, deviations must be expected. Additionally, the error model was fit primarily with data from a slightly higher energy range, where the density of states is larger; while this was done to reduce errors in the model from inadequate sampling, it could affect the accuracy of the error model in the lowest-energy regime. A comparison of how the two energies correlate across the full spectrum of energies (Supporting Information Figure S3) shows additional deviations from the model at high vacuum energies; as these states do not contribute to the low-GB-energy ensembles, these differences do not impact the analysis. The statistical analysis is most accurate for those data directly used in the derivation of the model.

Applications to Protein Design. A significant motivating force behind the development of this framework was for direct application to protein design. Thus, it is informative to consider a problem in this application space. We have recently described initial progress toward the development of variant Calmodulin−M13 peptide complexes with altered specificity.[21] In that work, we applied a hierarchical technique to the protein design problem at a number of focused sites. The same system is used here as an example with which we may evaluate the theory developed here.

Eight residues (five basic residues on M13 and three acidic groups on CaM) at a surface-exposed site at one end of the CaM−M13 binding site were varied to evaluate the viability of charge-reversal mutants at these positions (Figure 1b shows the design site). Each positive group was allowed to vary to the acidic amino acids and the amides, and each negative group was allowed to vary to the basic amino acids (including His) and the amides. As the three protonation states of histidine were considered individually, this corresponds to 1.6×10^6 possible sequences; with structural flexibility considered, there were 1.6×10^{26} individual structures under consideration. A significant number of rotameric states led to easily detected clashes with the fixed portion of the protein. After removal of these, the total structural search space was 8.8×10^{23}.

This space must be then be searched for low energy structures; the Dead-End Elimination (DEE) and A* algorithms may be used to enumerate the lowest-energy states in a guaranteed manner.[30,32] Rather than a single global minimum sequence and structure, we aim to find all sequences within a given cutoff of the global minimum, for several reasons. First, the DEE/A* approach requires a pairwise decomposable (in terms of individual side-chain positions) energy function. Thus, approximations to fundamentally non-pairwise energetic components (such as solvation free energies) must be made. Finding a number of low-energy states allows these to be reranked with more accurate energy functions, and to thereby obtain a better

Table 1. Number of States Found in Initial Search

E_0^{cut}	0	1	2	3	4	5	10	15	20
fine states	1	28	337	2352					
fine sequences	1	2	7	18					
coarse states	1	10	73	309	1095	3402	336 142	11 928 271	221 817 700
coarse sequences	1	4	21	51	99	188	2080	10 353	

Table 2. Rank of Full Search Global Minimum in Initial Search

	rotamer	DK-fleximer	$\chi_{1,2}$-fleximer	sequence-mer
rank	1	3	272	148
ΔE	0.0	0.20	7.88	9.71

estimate of the “true” lowest-energy sequences. Second, the design algorithm works on a single target energy function, while in reality several energies may need to be simultaneously optimized. For example, in optimizing the binding free energy between a pair of proteins, it is important to additionally maintain the stability of each individual structure. This can be addressed by optimizing on a single term that is a prerequisite for satisfying all others; in this case, optimizing the total complex energy (the sum of folding and binding free energies) satisfies the requirement. Given an ensemble of sequences with low complex energies, some will have higher affinity, and some higher stability; as we are particularly interested in high-affinity complexes, the low-energy set can be subsequently screened for this criterion. Finally, current models for protein energetics are still not ideal, and thus for experimental testing, a set of possible variants is desired.

A Fine Rotamer Library Makes Enumeration Infeasible. When DEE/A* is used to enumerate low-energy states of the full space, a problem becomes readily apparent: due to the large density of states, many with similar energies, it is infeasible to enumerate states beyond 3 kcal/mol above the global minimum (all computations beyond this level required beyond 8 GB of internal memory and many days of computational expense). While over 2000 structural states are found in this range, these states correspond to only 18 distinct sequences (see Table 1). This is a result of two features: the relative solvent exposure of the site and the size of the rotamer library used in the search. An augmented version of the Dunbrack−Karplus library was used in this search, with χ_1 and χ_2 sampled at ±10° around each standard rotamer. Since the design site is fairly exposed, it is reasonable to consider using the unaugmented library (the finer sampling is often needed in buried sites to allow reasonable packings to be found). Using this coarser library allows a much more extensive sampling of sequences (see Table 1). For example, 2080 sequences can be found within 10 kcal/mol of the global minimum, from a total of over 300 000 structural states.

However, when the results of the two calculations are compared, there is little correlation. Of the 18 sequences within 3 kcal/mol of the global minimum with the fine library, only four are within the same cutoff with the coarse library. While all 18 are found within 10 kcal/mol of the coarse global minimum, the sequence ranking third with the fine library ranks 595 with the coarse library. In terms of energies, the top fine library sequences are roughly 20 kcal/mol more favorable than those from the coarse library, and essentially no correlation between the two values is seen.

A Fleximer Model Makes the Search Tractable. These issues have been recognized previously, and the “fleximer” model of Mendes et al. is an elegant solution.[28] Briefly, in this approach, pairwise energies are computed for all rotameric states in the fine library, but the discretization of the coarse library is used in the search. When computing energies for the search, the interactions of any parent rotamer at a given position is given by the Boltzmann-weighted average of the interactions of all substates of that rotamer. The results of such a search do not correspond to any single structural state, and thus a second step is required, in which the minimum energy state for a given set of fleximers is found. As this involves a search over only roughly nine choices at each position, this evaluation is very fast. This approach is very successful; using the fleximer model, the true, fine-library global minimum is ranked third, with a difference in energy of only 0.2 kcal/mol from the fleximer global minimum (see Table 2). Of the 18 top fine-library sequences, 11 are found within the same 3 kcal/mol of the fleximer global minimum, and 17 are found within the top 5 kcal/mol.

The search with the fleximer model is much more efficient than that with the fine library, and thus it is feasible to enumerate as high as 20 kcal/mol from the global minimum (in which range there are over 40 million fleximer states, and 11 000 distinct sequences). When collapsed into a single rotameric state, the energy obtained for a given fleximer is identical to that found in the fine-library search. However, since the search is performed on the fleximer energy, it is possible that true low-energy sequences are not found in the fleximer-based search. This limitation depends directly on how many fleximer states are enumerated, and how many true low-energy states are desired. For example, we have seen that 5 kcal/mol of fleximer states are required to find 17 of the 18 sequences within 3 kcal/mol of the fine-rotamer minimum. How far must the fleximer energetic landscape be explored to find all sequences within a given cutoff in fine-library energy? Again, this question can be directly addressed with the statistical framework presently here.

Defining Correlations between Rotamer and Fleximer Energies. In order to address this question, it is necessary to know both the degree of correlation between the fleximer and rotamer energies and the distribution of energetic states in the fleximer model. As can be seen in Table 3, the number of states increases exponentially with increasing distance from the global minimum. This is consistent with an ensemble of states that is normally distributed: the extremes of a normal distribution are essentially exponential, and only 0.7% of the total sequence space is sampled in the lowest

Table 3. Number of States Found in Initial Search

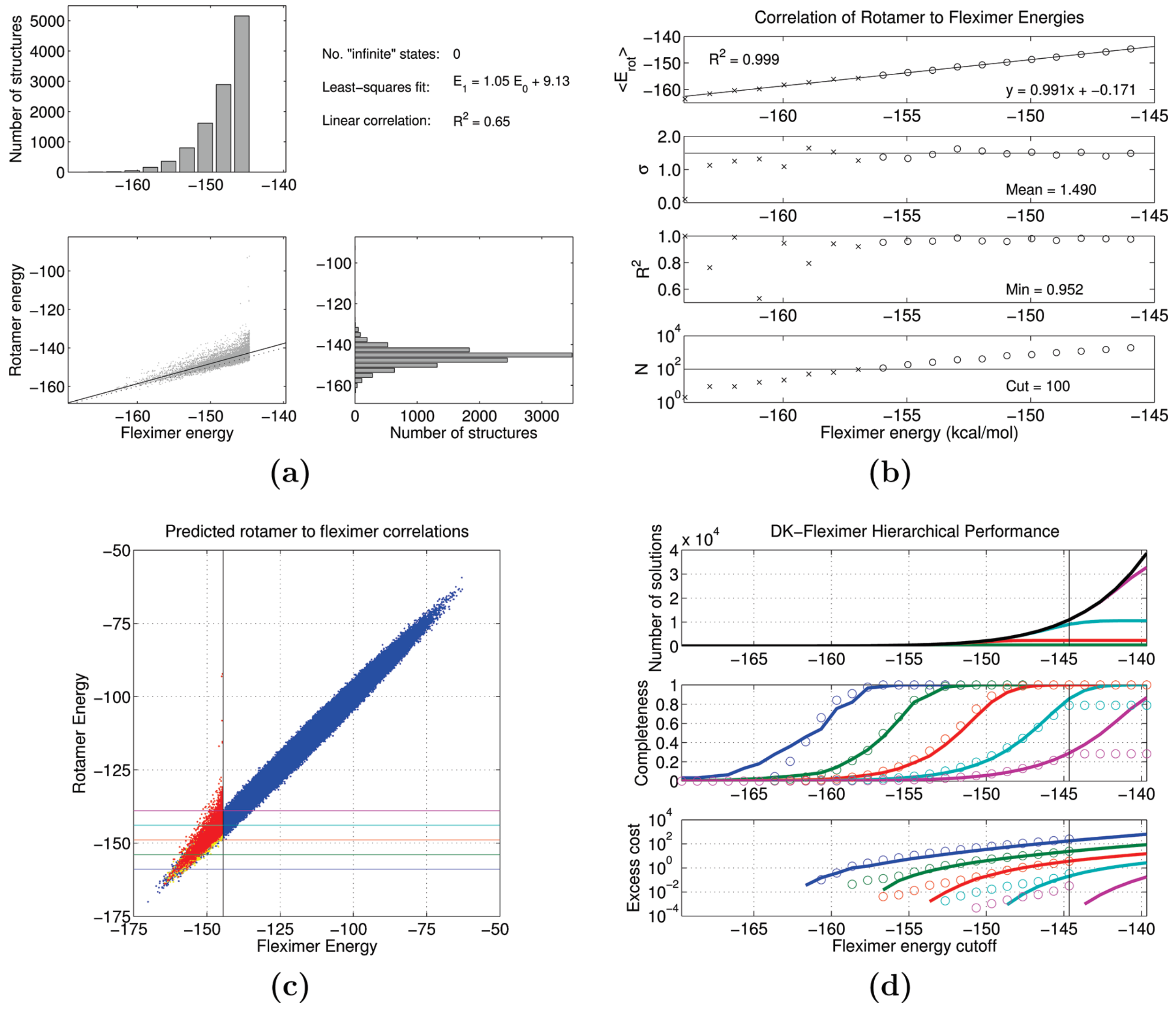
	fine rotamer		DK fleximer		$\chi_{1,2}$-fleximer		sequence-mer
E_0^{cut}	states	seq.	states	seq.	states	seq.	sequences
0	1	1	1	1	1	1	1
1	28	2	7	3	2	1	1
2	337	7	47	15	3	1	3
3	2352	18	156	32	7	4	11
4			481	45	19	9	17
5			1316	69	35	11	25
10			77466	632	975	145	161
15			2205651	3136	13591	788	743
20			41998715	11030	129939	3198	2576
25					971173	9585	7441
30					5773982	24501	18310
40					116352240		78564
50							229332
75							1019658
100							1538405
150							1600000
total	1.49×10^{26}		3.58×10^{18}		1.42×10^{13}		1.60×10^{6}

20 kcal/mol. As was done in the conformational analysis of glucose, the fleximer energies were grouped into bins, and the distributions of rotamer energies within each bin were characterized (Figure 3a and b). While the distribution of rotamer energies in each bin deviates somewhat from normality, the bulk of each distribution is, in fact, well fit by a normal curve—the R^2 of the fit was greater than 90% in all cases, except in the lowest-energy bins, where there were few data. Samples of the fits to individual bins may be found in the Supporting Information (Figure S4). The deviations are most notable at low energies, where the observed number of states is less than would be expected from a normal curve, and at high energies, where a greater number of states is observed. This is expected, as there is a lower bound on the rotamer energy for a given fleximer energy, but no upper bound. While it may be possible to define an alternate distribution that is a slightly better fit for the tails, we have chosen to use the normal distribution for further analysis.

The details of a linear fit to the mean are given in Table 4 and Figure 3b. The means are essentially perfectly correlated ($R^2 = 0.999$), and the best-fit line is very near the $y = x$ line, with a slope of 0.99 and an intercept of -0.17. The standard deviation in each bin is on average 1.49 kcal/mol. This fit was done with all sequences within 20 kcal/mol of the global minimum, including data from all bins (1 kcal/mol in width) with a population of at least 100; this provides 11 points for the fitting. However, it is clear that the fit also matches well those data from low-energy bins with lower sampling. Eleven-thousand sequences were considered in order to reach this result, with 42 million individual structures enumerated. Thus, a natural question is whether fewer data would suffice. Table 4 additionally shows the results of fitting only data within 10 or 15 kcal/mol of the global fleximer minimum; in these cases, all bins with at least 50 points were included, so as provide a reasonable number of bins for fitting. The results are very similar—the slope of the best-fit line ranges from 0.97 to 0.99, and the average standard deviation is between 1.5 and 1.6 kcal/mol. Thus, it is clear that very similar results are obtained even when only the very lowest-energy fleximer states are considered; the 10 kcal/mol ensemble contains only 0.04% of the total sequence space, with less than 80 000 structures, while the 20 kcal/mol ensemble contains 0.7% of the sequence space (and 42 million structures).

Model Distributions Capture Observed Behaviors in the Hierarchy. In addition to the correlation between energetic levels, the theoretical framework outlined above requires a knowledge of the distribution of states in the lowest (fleximer) level. The DEE/A* methodology gives a rigorous enumeration of the lowest-energy states, and we additionally have prior knowledge of the total number of sequences in the space. Thus, it is possible to perform a nonlinear least-squares fit of a normal distribution to the available data. The fit (see Supporting Information Figure S5) is excellent in the region where there are data, with an R^2 of 98.8%, although these data only occupy the region from -4.2 to -2.5 standard deviations from the mean. Thus, this should be considered an estimate, and it should be expected that there will be significant deviations between this estimate and the true distribution.

This estimated fleximer distribution was then combined with the observed correlation and error model for transferring between fleximer and rotamer energies to provide an estimate of the complete distribution of minimum rotamer energies. This was then used to compute expected completeness and excess-cost curves, as a function of fleximer-energy cutoff, for different low-rotamer-energy ensembles (see Figure 3c and d). Given the 20 kcal/mol cutoff that was used, near 100% completeness is expected for ensembles up to 15 kcal/mol of the global minimum in rotamer energy, roughly 80% completeness is expected for rotamer energies within the same 20 kcal/mol cutoff, and about 35% of sequences in the lowest 25 kcal/mol of rotamer energies are expected to have been found. Comparing to the observed data, the agreement with the completeness estimates is remarkable. For the lowest energy ensembles (5, 10, and 15 kcal/mol), the 100% completeness is strongly supported by the observation that the ensemble is converged with increasing fleximer energy. For the higher-energy rotamer ensembles (20 and 25 kcal/mol), the number of observed states matches very closely the number of states predicted by the model; the expected total of number of states can be used with the



(a)

(b)

(c)

(d)

Figure 3. Correlations of rotamer to fleximer energies. (a) The distribution of fleximer energies within 20 kcal/mol of the global minimum are shown, along with the distribution of minimum rotamer energies for the same ensemble and the correlation between the two. A linear least-squares fit gives a slope near unity, with a modest correlation ($R^2 = 0.65$). (b) The results of fitting a Gaussian to the distribution of rotamer energies within 1.0 kcal/mol bins of fleximer energies are shown. The mean ($\langle E_{\mathrm{rot}} \rangle$) shows strong linear correlation. The standard deviation (σ) is uniform, with a value of 1.5 kcal/mol in the nearly all bins, and the fit to a normal distribution is excellent ($R^2 > 0.9$) in all cases with 100 data points (N) or higher, shown as open circles. (c) The correlation of rotamer to fleximer energies simulated using the distribution of fleximer states and correlations of rotamer to fleximer energies computed from low energy states (blue points) are shown, along with the observed data (red points). Yellow points indicate the simulated distribution of rotamer energies given the actual (low-energy) fleximer energies. (d) The computed performance metrics are shown for the simulated data (solid lines), and for the observed data (open circles). Colors correspond to different rotamer energy cutoffs (from 5 to 25 kcal/mol, in increments of 5 kcal/mol), indicated by horizontal lines of the same color in the right panel. The black line in the number of solutions indicates the total number of solutions at the fleximer level.

Table 4. Statistics of Rotamer to Fleximer Fit

	DK fleximer			χ_1, χ_2 fleximer			sequence-mer		
E_0^{cut}	slope	intercept	$\langle\sigma\rangle$	slope	intercept	$\langle\sigma\rangle$	slope	intercept	$\langle\sigma\rangle$
30[a]				0.83	−20.28	3.73	0.66	−41.70	4.05
20[a]	0.99	−0.17	1.49						
30[b]				0.79	−26.14	3.74	0.66	−42.80	3.99
25[b]				0.79	−25.34	3.88	0.65	−44.09	3.86
20[b]	0.98	−1.93	1.48	0.84	−17.38	4.01	0.65	−43.04	3.72
15[b]	0.99	−0.58	1.51	1.07	19.25	4.42	0.62	−48.63	3.59
10[b]	0.97	−2.70	1.57	0.62	−49.67	4.68			

[a] Fit included all bins with at least 100 values. [b] Fit included all bins with at least 50 values. In all cases, bins were of 1 kcal/mol width.

number of observed states to compute a completeness measure that agrees with prediction.

Excess-cost estimates suggest that 80% complete ensembles can be obtained at very little excess cost (10%), and 100% complete ensembles can be realized with an excess cost of roughly 1 (equal number of kept and discarded states). Additionally, it can be seen that a 100% complete ensemble of the top 10 kcal/mol (rotamer energy) should be attainable with enumerating only up to 15 kcal/mol in rotamer energy, and that the top 5 kcal/mol of rotamer energy sequences can be fully determined with a fleximer cutoff of 10 kcal/mol. Considering the observed data, there is good agreement in the regime of near-complete sampling (completeness greater than about 75%), but the observed excess work is significantly larger than the model would predict for less-complete sampling. This is not surprising, as the true distribution of rotamer energies for a given fleximer energy has a longer positive tail than the model normal distribution. Thus, there are a larger number of discarded states than expected by the model, which leads to higher excess cost. This is most dramatic when there are few low-energy states (low completeness), and less apparent when there are many states.

Statistical Guarantees for a Hierarchical Approach. We thus have a very important result—while with a direct search using a fine library only the top 3 kcal/mol could be enumerated (giving 18 sequences), using the fleximer model allows complete enumeration of 15 kcal/mol of low-energy states (2388 sequences). While the completeness of this ensemble is not algorithmically guaranteed, as is the case when a space is directly searched with DEE/A*, a *statistical guarantee* has been provided; that is, we can rigorously define a confidence value that all solutions have been found. Averaging over 10 000 model distributions, the completeness of the 15 kcal/mol ensemble was found to be 99.992%. Given the size of this ensemble, this leads to perfect completeness in 83% of the cases, and in an additional 15%, a single sequence was not found. The ensemble may thus be described as perfectly complete with greater than 80% confidence, and there is greater than 98% confidence that no more than a single sequence has been omitted.

Improving Efficiency with Alternative Hierarchies. While the above approach allowed for complete sampling of the top 15 kcal of energies, a great deal of computational expense was involved. In particular, 42 million fleximer states had to be enumerated, and about 100 000 fleximer states expanded to a unique rotameric state (up to 10 per sequence). This involved roughly one week of computation on a single AMD Opteron 250 (2.4 GHz) processor. Could this expense be reduced by creating coarser-grained models for the initial search? To test this, the fleximer model was applied to alternate groupings of rotameric states. In the first, all rotamers with the same χ_1 and χ_2 angles (from the Dunbrack−Karplus library) were grouped in a single fleximer; the finer samplings of ±10° were included in the fleximer defined by the parent angles. In the second, all rotamers of a given amino acid were grouped into a single "sequence-mer". Thus, whereas the Dunbrack−Karplus-based fleximer sampled 282 fleximer states at each acidic position, and 174 at each basic position, the $\chi_{1,2}$-fleximer model samples 54 and 39 per acidic or basic position, respectively, and the sequence-mer model samples eight and five states per position. This leads to a dramatic reduction in the overall size of the search space—from 3.6×10^{18} for the Dunbrack−Karplus fleximer to 1.4×10^{13} for the $\chi_{1,2}$-fleximer and 1.6×10^{6} for the sequence-mer.

For each fleximer model, all low-energy states were enumerated with DEE/A*, and up to 10 states per sequence were expanded into single rotamer structures, as discussed above (see Table 3). For the $\chi_{1,2}$-fleximer model, there were 130 000 states corresponding to 3200 sequences in the top 20 kcal/mol of fleximer energies; for the sequence-mer model, there were 2600 sequences in the same range. This is, of course, much smaller than the number of sequences found in the top 20 kcal/mol of the initial (Dunbrack−Karplus, DK) fleximer model, and thus larger cutoffs were considered. In the top 30 kcal/mol, 24 500 sequences (almost 6 million states) were found for the $\chi_{1,2}$ model, and 18 300 sequences were found with the sequence-mer model.

As above, these data were binned according to fleximer energy, and the relationships between rotamer and fleximer energy were determined (see Table 4 and Figure 4). As for the DK-fleximer model, the mean rotamer energy is linearly correlated with both coarser models, with R^2 values of greater than 99%. However, the slope of the correlation is below unity in both cases—0.83 for the $\chi_{1,2}$-fleximer and 0.66 for the sequence-mer. The distributions of the rotamer energies around the mean are also fit well by a Gaussian, and the standard deviation is constant across the observed range. Not surprisingly, however, the distributions are much broader than was the case for the Dunbrack−Karplus fleximer (3.7 kcal/mol for $\chi_{1,2}$, 4.0 kcal/mol for sequence-mer). While these seem quite similar, it is important to note that, if the data in each set were scaled to give a slope of unity in the correlation of the means, the standard deviation about the mean would scale by the reciprocal of the original slope. Thus, the normalized standard deviation for the $\chi_{1,2}$ distribution is 4.5 kcal/mol, and that for the sequence-mer distribution is 6.1 kcal/mol. This compares with 1.5 kcal/mol for the Dunbrack−Karplus fleximer.

These linear correlations were then combined with best-Gaussian fits to the original fleximer energy distribution (Table 5) to give the expected completeness and excess-cost for varying energetic cutoffs (see Figure 5). The completeness curves for the $\chi_{1,2}$-fleximer model transition more sharply than those for the sequence-mer model; this is expected, as the slope of the transition depends on the accuracy of the correlation between models. However, at the highest cutoff in fleximer-energy considered (30 kcal/mol), both models give similar overall completeness. Curiously, the completeness for the 20 kcal ensemble in rotamer energy shows roughly the same degree of completeness (80%) with both these models (and a 30 kcal/mol fleximer cutoff) as with the DK-fleximer model using a 20 kcal/mol cutoff. This is coincidental, but allows for an interesting observation to be made concerning the ensembles of slightly higher and lower rotamer energy. The 15 kcal/mol (rotamer energy) ensemble was 99.992% complete when searching with the DK-fleximer model, and the completeness of the 25 kcal/mol ensemble was 35%. The completeness of the 15 kcal/mol ensemble is somewhat reduced when the search is performed with the coarser fleximer models (97.0% for the $\chi_{1,2}$-fleximer and 97.7% for the sequence-mer). This is a

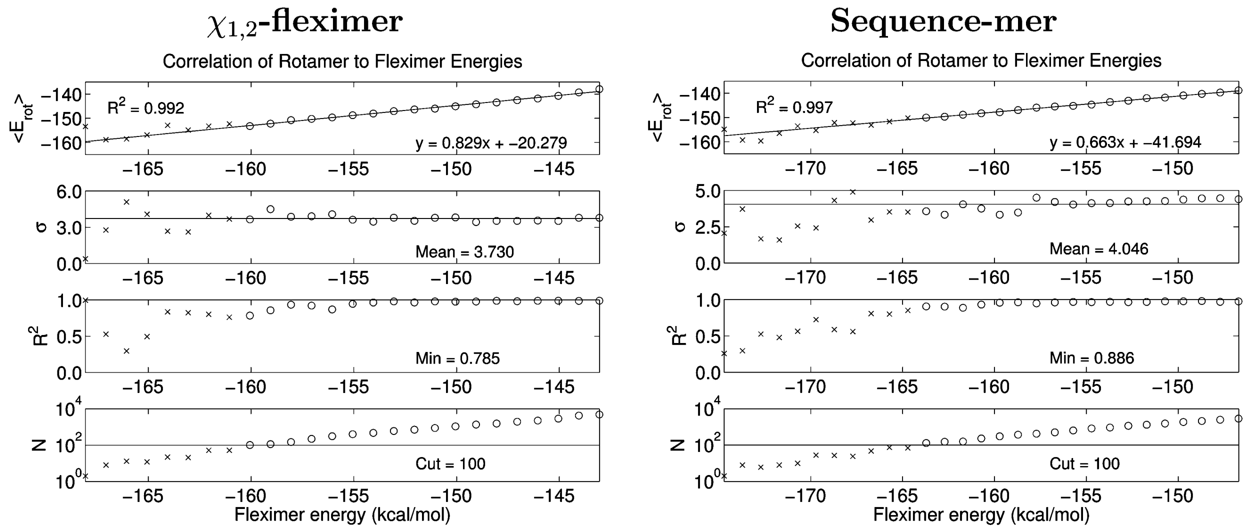


Figure 4. Correlations of rotamer to $\chi_{1,2}$-fleximer and sequence-mer energies. The results of fitting a Gaussian to the distribution of rotamer energies within 1.0 kcal/mol bins of $\chi_{1,2}$-fleximer (left) and sequence-mer (right) energies are shown. In both cases, the mean ($\langle E_{\mathrm{rot}} \rangle$) shows strong linear correlation, the standard deviation (σ) is uniform, and the fit to a normal distribution (R^2) is very good in all cases with 100 data points (N) or higher, shown as open circles.

Table 5. Statistics of Gaussian Fit to Low-Energy Sequences

	mean	σ	R^2	bins
DK-fleximer	−119.58	10.16	0.9876	21
$\chi_{1,2}$-fleximer	−122.59	9.23	0.9529	27
sequence-mer	−110.78	15.61	0.9992	31

result of the broader transition to completeness and thus is not unexpected. However, the broadness of the transition also contributes to a *higher* completeness of the less-fully sampled ensembles—the 25 kcal/mol (rotamer energy) ensemble is estimated to be 44% complete when the $\chi_{1,2}$-fleximer model is used, and more than 52% complete when the sequence-mer approximation is used.

The more coarsely sampled models naturally have a larger excess-cost; achieving near 100% completeness (for example, in the 15 kcal/mol ensemble) requires evaluation of 10 times as many states as are found in the final ensemble. While this may initially seem like a large drawback compared with the DK-fleximer model (which only required evaluation of double the number of kept states), it is important to additionally consider the computational cost of each step. For all cases, the full pairwise energy matrix for the fine library must be computed; this took roughly 50 min on a single AMD Opteron 250 (2.4 GHz) processor. To enumerate the top 20 kcal/mol of states in the DK-fleximer model (41 million states and 11 030 sequences) took 6.5 days on a single CPU, while to enumerate the top 30 kcal/mol of the $\chi_{1,2}$-fleximer model (5.8 million states, 24 501 sequences) took just under 1 h, and to enumerate the 18 310 sequences in the top 30 kcal/mol of the sequence-mer model took less than 1 min. These dramatic differences in the initial search are somewhat offset by the need to do additional calculations to find the true low-energy structures corresponding to each fleximeric state; this process is more costly for fleximers containing more members. For the DK-fleximer model, the cost of this stage was negligible, roughly 1 h, while it took roughly 1.5 days for the $\chi_{1,2}$-fleximer model and 2 days for the sequence-mer model. However, when all steps are considered, both coarser fleximer models (requiring about 1.5−2 days total time) are significantly more cost-efficient than the DK-fleximer model (requiring roughly 1 week total CPU time).

While slightly more cost-effective overall than the sequence-mer model, the $\chi_{1,2}$-fleximer model gives a somewhat poorer best-Gaussian fit to the original distribution ($R^2 = 95.2\%$) than either the DK-fleximer ($R^2 = 98.8\%$) or the sequence-mer model ($R^2 = 99.9\%$) and noticeably underestimates the true density of states at low fleximer energies (Supporting Information Figure S5). The reason for this is not entirely clear, and work on additional systems will be needed to determine whether this is a broader issue.

The Underlying Sequence-mer Distribution Is Normally Distributed. The above analysis was based on an assumption that the underlying (lowest level) energy distribution can be reasonably estimated by a normal distribution fit to the lowest energy states. In most cases, the vast number of states precludes directly assessing this. However, as our model system consists of only 1.6 million distinct sequences, we can, in fact, enumerate the sequence-mer distribution. Figure 6 shows this full distribution, along with a Gaussian fit to all the data, and a Gaussian fit to only the lowest 30 kcal/mol of sequences (2% of the total).

Considering these data, we may make two observations. First, it is notable that a normal distribution models the full distribution of energies nearly perfectly, and thus the initial assumption is validated in this case. Second, while the low-energy fit is not perfect, the estimated distribution matches the actual data remarkably well; the fit mean is shifted to slightly lower energy, and the fit variance is slight smaller. Thus, the number of moderate-energy states will be overestimated and the number of high energy sequences underestimated. As the most significant deviations are at high energies, which contribute little to the low-energy states at higher hierarchical levels, the errors from this approach will

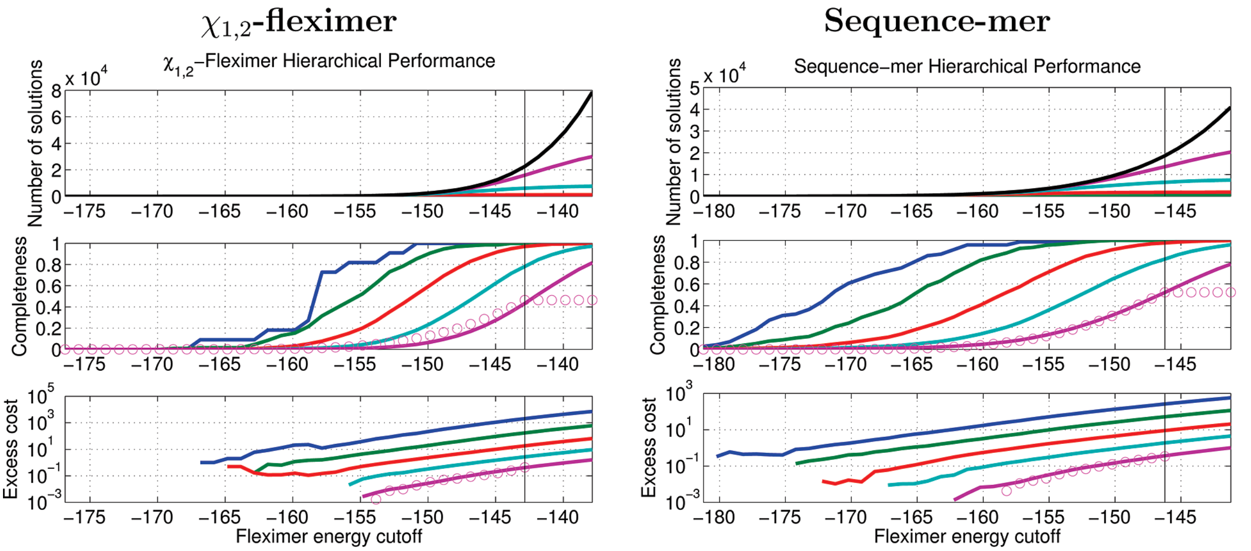


Figure 5. Performance metrics of rotamer to $\chi_{1,2}$-fleximer and sequence-mer hierarchies. The computed performance metrics for $\chi_{1,2}$-fleximer (left) and sequence-mer (right) are displayed. Solid lines of different colors correspond to simulated data for different rotamer energy cutoffs (from 5 to 25 kcal/mol, in increments of 5 kcal/mol, as in Figure 3). Open circles correspond to the actual data for highest rotamer-energy cutoff (25 kcal/mol). The black line in the number of solutions indicates the total number of solutions at the fleximer level.

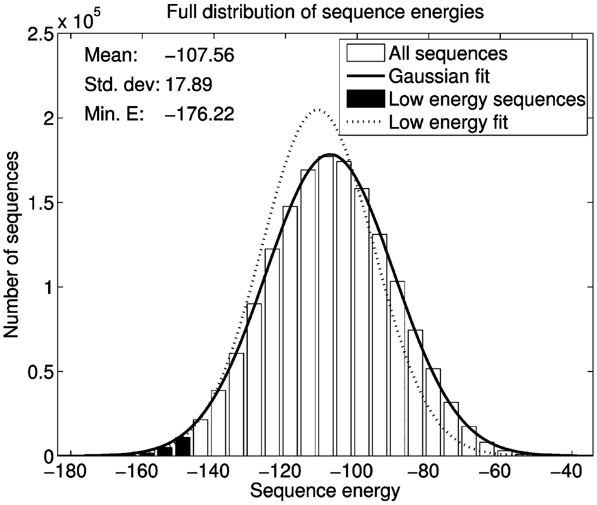


Figure 6. Complete histogram of sequence-level energies. The distribution of the enumerated set of sequence energies for the design site are shown. The dark curve is a best-fit normal distribution to all the given data. The dotted curve is the best-fit Gaussian curve, using only the lowest 30 kcal/mol of sequences (denoted by black bars).

be further reduced. As a result, reasonable results may be obtained using a predicted sequence-mer distribution.

Applications of the Statistical Framework. The previous sections detailed applications for the theory developed here in conformational search and in protein design. However, the methods may be applied with equal ease to any of a large number of applications in molecular simulation and design. The statistical framework provides a number of key benefits. One of the most significant of these is the ability to assess completeness, which can be used to provide a level of confidence that the global minimum has been found. It may also be used to answer a challenging problem: if no satisfactory solution is found, is it the result of incomplete sampling or due to the true absence of a satisfactory solution? These questions are important ones in protein design, in ligand docking, and in other areas of molecular design.

An additional application is for problems where an ensemble of states is necessary. Important problems in protein evolution can be addressed by determining the set of all sequences compatible with a given protein structure; protein-design methods can be applied to this problem, but completeness measures are essential. Conformational search methods are also often used to generate ensembles of states from which ensemble-averaged properties, such as free energy and entropy, may be computed. The challenge in these approaches, though, is that these properties are only accurate with adequate sampling of the low-energy regions of phase space; completeness provides a direct means of evaluating such sampling.

It should be noted that there is an important distinction that can separate the commonly-used search algorithms. Some methods are designed to enumerate all low-energy structures (these include exhaustive search approaches and tree-search-based methods such as Dead-End Elimination and A*), while other methods produce a *sampling* of the low-energy states (genetic algorithms, Monte Carlo, and simulated annealing are among these). While in an ideal situation, the sampling produced by the latter methods will be complete (or nearly so), this cannot be guaranteed with finite resources; the first class of methods can provide guarantees that *all* low-energy states (of the set considered) have been found. The statistical theory described here is compatible with both types of algorithm so long as a functional description of the expected sampling produced by the search algorithm can be given—DEE/A* gives a sampling distribution described by a step-function, while Monte Carlo gives a sampling distribution equivalent to the Boltzmann distribution at a given temperature.

Possible Limitations. The utility of the approach outlined here is fundamentally limited by the accuracy to which the underlying (lowest level) probability distribution and the correlations between hierarchical levels can be estimated. In general, these will not be known *a priori* and thus, as discussed above, must be derived *a posteriori* from sampled data. If there is a systematic error which leads to a dramatic difference in energy for only a subset of states, an error model derived solely from states that are low energy in the reference model may not be representative of the full set of states. For example, consider two models in which a particular class of states (e.g., molecules with a net charge of $-1e$) are destabilized in the lower-level model but stabilized in the higher-level model, relative to all other states. The low-energy states from the first model may not include any molecules of this class, and thus this bias would be excluded from the error model. As a result, the completeness could be estimated to be very high, even though a significant number of lower-energy states (at the higher-level) were missed.

A related limitation is the need to be able to sample enough of the space in the lower-level model to derive a reasonable error model. In a system where the density of states very close to the global minimum is large, methods which aim to enumerate low-energy states may not be able to sample a wide enough range of energies for a reasonable model to be obtained. This should be less of an issue for sampling methods which include some higher energy states (such as Monte Carlo), although cases that remain problematic in this regime could still be constructed.

These caveats are important to be aware of, and care should be taken to evaluate how accurate the error models are expected to be. It should also be noted that it is possible to decouple the derivation of the error model from the search. For example, a random (or less-biased) search over states in the lower-level model would give a broad sampling of states that may be used to derive an error model across a wide energy range; the model could then be used to evaluate the performance of more targeted search strategies. Although there will always be *some* possibility of error, the use of a statistical model explicitly considers this; terms like the completeness are estimated to a certain level of confidence, rather than given as precise predictions.

5. Conclusion

We have outlined a statistical framework for performance analysis in hierarchical methods, with a particular focus on applications in molecular design. The theory is derived from fundamental statistical principles, presuming that the relationship between the results of each hierarchical level may be described by some functional correlation (linear or not), and an error model for how values are distributed around the correlation curve. Two example problems—one in conformational search and one in protein design—clearly show the usefulness of this approach; measures of completeness of the final ensemble can be computed, providing a level of confidence that adequate sampling of low-energy states has been achieved.

The framework we have described here is applicable not only to specific examples presented here but to any problem in molecular design that involves a hierarchical approach. Perhaps the most common of these is that of protein−ligand docking and virtual high-throughput screening, and we look forward to seeing this framework applied to these problems.

Acknowledgment. D.F.G. would like to thank Bruce Tidor for making the protein-design software available, as well as for helpful discussions on an early version of this manuscript. D.F.G. also thanks the State University of New York, the College of Engineering and Applied Sciences and the Department of Applied Mathematics and Statistics for financial support.

Supporting Information Available: Figures demonstrating the various fits in more detail are presented. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Szymkowski, D. E. Creating the next generation of protein therapeutics through rational drug design. *Curr. Opin. Drug Discovery Dev.* **2005**, *8*, 590.

(2) Rosenberg, M.; Goldblum, A. Computational protein design: A novel path to future protein drugs. *Curr. Pharm. Des.* **2006**, *12*, 3973.

(3) Razeghifard, R.; Wallance, B. B.; Pace, R. J.; Wydrzynski, T. Creating functional artificial proteins. *Curr. Protein Pept. Sci.* **2007**, *8*, 3.

(4) Drexler, K. E. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 5275.

(5) Pabo, C. O. Molecular technology: Designing proteins and peptides. *Nature* **1981**, *301*, 200.

(6) Park, S.; Yang, X.; Saven, J. G. Advances in computational protein design. *Curr. Opin. Struct. Biol.* **2004**, *14*, 487.

(7) Anderson, A. C. The process of structure-based drug design. *Chem. Biol.* **2003**, *10*, 787.

(8) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein-ligand docking: Current status and future challenges. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 15.

(9) Halperin, I.; Ma, B. Y.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 409.

(10) Bonvin, A. M. Flexible protein-protein docking. *Curr. Opin. Struct. Biol.* **2006**, *16*, 194.

(11) McCarrick, M. A.; Kollman, P. A. Predicting relative binding affinities of non-peptide HIV protease inhibitors with free energy perturbations calculations. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 109.

(12) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. Absolute binding free energies: A quantitative approach for their calculation. *J. Phys. Chem. B* **2003**, *107*, 9535.

(13) Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C. D.; Shirts, M. R.; Sorin, E. J.; Pande, V. S. Direct calculation of the binding free energies of FKBP ligands. *J. Chem. Phys.* **2005**, *123*, 084108 .

(14) Mardis, K. L.; Luo, R.; Gilson, M. G. Interpreting trends in the binding of cyclic ureas to HIV-1 protease. *J. Mol. Biol.* **2001**, *309*, 507.

(15) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III. Assessing scoring functions for protein−ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032.

(16) Gohlke, H.; Case, D. A. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem.* **2004**, *25*, 238.

(17) Jaramillo, A.; Wodak, S. J. Computational protein design is a challenge for implicit solvation models. *Biophys. J.* **2005**, *88*, 156.

(18) Given, J. A.; Gilson, M. K. A hierarchical method for generating low-energy conformers of a protein-ligand complex. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 475.

(19) Grüneberg, S.; Stubbs, M. T.; Klebe, G. Successful virtual screening for novel inhibitors of human carbonic anhydrase: Strategy and experimental confirmation. *J. Med. Chem.* **2002**, *45*, 3588.

(20) Floriano, W. B.; Vaidehi, N.; Zamanakos, G.; Goddard, W. A., III. HierVLS Hierarchical docking protocol for virtual ligand screening of large-molecule databases. *J. Med. Chem.* **2004**, *47*, 56.

(21) Green, D. F.; Dennis, A. T.; Fam, P. S.; Tidor, B.; Jasanoff, A. Rational design of new binding specificity by simultaneous mutagenesis of calmodulin and a target peptide. *Biochemistry* **2006**, *45*, 12547.

(22) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187.

(23) Kuttel, M.; Brady, J. W.; Naidoo, K. J. Carbohydrate solution simulations: Producing a force field with experimentally consistent primary alcohol rotational frequencies and populations. *J. Comput. Chem.* **2002**, *23*, 1236.

(24) Im, W.; Lee, M. S.; Brooks, C. L., III. Generalized born model with a simple smoothing function. *J. Comput. Chem.* **2003**, *24*, 1691.

(25) Green, D. F. Optimized parameters for continuum solvation calculations with carbohydrates. *J. Phys. Chem. B* **2008**, *112*, 5238.

(26) Ikura, M.; Clore, G. M.; Gronenborn, A. M.; Zhu, G.; Klee, C. B.; Ad, B. Solution structure of a Calmodulin-target peptide complex by multidimensional NMR. *Science* **1992**, *256*, 632.

(27) Dunbrack, R. L., Jr; Karplus, M. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J. Mol. Biol.* **1993**, *230*, 543.

(28) Mendes, J.; Baptista, A. M.; Arménia Carrondo, M.; Soares, C. M. Improved modeling of side-chains in proteins with rotamer-based methods: A flexible rotamer model. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 530.

(29) Hanf, K. J. M. *Protein design with hierarchical treatment of solvation and electrostatics*, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, September 2002.

(30) Desmet, J.; De Maeyer, M.; Hazes, B.; Lasters, I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **1992**, *356*, 539.

(31) Gordon, D. B.; Mayo, S. L. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J. Comput. Chem.* **1998**, *19*, 1505.

(32) Leach, A. R.; Lemon, A. P. Exploring the conformational space of protein side chains using dead-end elimination and the A * algorithm. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 227.

(33) Warwicker, J.; Watson, H. C. Calculation of the electric potential in the active site cleft due to α-helix dipoles. *J. Mol. Biol.* **1982**, *157*, 671.

(34) Gilson, M. K.; Honig, B. H. Calculation of electrostatic potentials in an enzyme active site. *Nature* **1987**, *330*, 84.

(35) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(36) Humphrey, W.; Dalke, A.; Schulten, K. VMD − Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33.

CT9004504

JCTC Journal of Chemical Theory and Computation

Analysis of Hydrogen Tunneling in an Enzyme Active Site Using von Neumann Measurements

Isaiah Sumner and Srinivasan S. Iyengar*

Department of Chemistry and Department of Physics, Indiana University, 800 E. Kirkwood Ave, Bloomington, Indiana 47405

Received November 25, 2009

Abstract: We build on our earlier quantum wavepacket study of hydrogen transfer in the biological enzyme soybean lipoxygenase-1 by using von Neumann quantum measurement theory to gain qualitative insights into the transfer event. We treat the enzyme active site as a measurement device which acts on the tunneling hydrogen nucleus via the potential it exerts at each configuration. A series of changing active site geometries during the tunneling process effects a sequential projection of the initial, reactant state onto the final, product state. We study this process using several different kinds of von Neumann measurements and show how a discrete sequence of such measurements not only progressively increases the projection of the hydrogen nuclear wavepacket onto the product side but also favors proton over deuteron transfer. Several qualitative features of the hydrogen tunneling problem found in wavepacket dynamics studies are also recovered here. These include the shift in the "transition state" toward the reactant as a result of nuclear quantization, greater participation of excited states in the case of deuterium, and the presence of critical points along the reaction coordinate that facilitate hydrogen and deuterium transfer and coincide with surface crossings. To further "tailor" the dynamics, we construct a perturbation to the sequence of measurements, that is a perturbation to the dynamical sequence of active site geometry evolution, which leads us to insight on the existence of sensitive regions of the reaction profile where subtle changes to the dynamics of the active site can have an effect on the hydrogen and deuterium transfer process.

I. Introduction

Hydrogen transfer reactions[1–3] play a significant role in many organic[2–5] and biological[6–10] reactions. Due to the de Broglie wavelength of the transferring hydrogen atom, the role of quantum nuclear effects in such reactions has been one focus area of study.[2–27] Experimentally, an important indication of quantum nuclear effects including tunneling is the appearance of an unexpectedly large primary kinetic isotope effect (KIE), which has been noted in many lipoxygenases. For example, the room temperature rate constant for hydrogen nuclear transfer (k_H) catalyzed by the enzyme soybean lipoxygenase-1 (SLO-1)[13,16–18,23–26,28–32] is a factor of 81 larger than that for deuterium nuclear transfer (k_D).[29] Human lipoxygenase was noted to have a similar behavior.[33] Quantum mechanical tunneling has been proposed to have a central role in this phenomenon,[16,23,24,26] since this observation cannot be explained using classical rate theories. Temperature dependence of primary and secondary isotope effects is another set of experimentally measurable parameters that directly probe the extent of quantum nuclear effects. The proper description of nuclear quantum effects for hydrogen-transfer reactions, including the role of tunneling, is a challenging and an often actively debated area of study.[6,9,10,19,22,27,34–36]

A few of the approaches that adequately treat the quantum nuclear effect and have been used to study hydrogen transfer in enzymes are as follows. [This paragraph is not an exhaustive review of all treatments of quantum nuclear effects in enzymes but only highlights some of the prevailing studies. For a detailed overview of the methods employed, see refs 6, 7, 9, 10, and 22 and publications cited within these references.] Klinman and co-workers experimentally[11]

* Corresponding author e-mail: iyengar@indiana.edu.

10.1021/ct900630n
Published on Web 03/31/2010

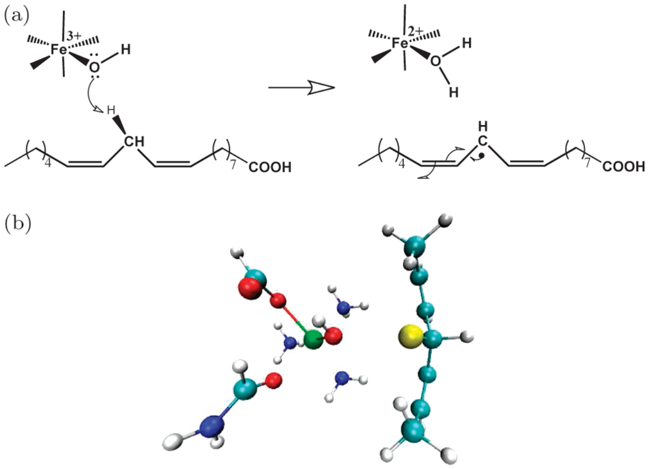


Figure 1. (a) The rate-determining step in SLO-1 and (b) one of the two active site models used in ref 26 with pruned representations of the active site residues and the substrate. The transferring hydrogen is enlarged and shown in yellow.

noted the effect of tunneling on the hydrogen transfer steps in biological enzyme catalysis. They have subsequently computed associated rates with a vibrationally nonadiabatic methodology[37,38] that employs Franck−Condon-like overlaps based on one-dimensional potentials. Warshel and co-workers[14–17,36] used Feynman path integral approaches[39–41] to describe the trajectory of the quantized hydrogen nucleus, which moves on an enzyme potential surface computed from empirical valence bond (EVB) theory.[7,42–44] Additionally, calculations on the uncatalyzed reaction in a reference solution [usually water] allowed them to explore enzyme-specific contributions to catalysis.[16–21] Truhlar, Gao and co-workers[22,32] have utilized a multidimensional tunneling correction to variational transition state theory,[22,34] where the potential energy surfaces are generally obtained from QM/MM calculations.[45–48] Hammes-Schiffer and co-workers[23,49] implemented a vibronically nonadiabatic formalism to treat proton-coupled electron transfers. This method is based on EVB[7,42–44] and includes quantum mechanical treatment of one electron and one proton that undergo proton-coupled electron transfer. The protein is treated through classical molecular dynamics simulations.[7,42–44] Schwartz and co-workers[10] utilized a semiclassical description based on the Langevin equation. A classical dynamics simulation was conducted with a Hamiltonian that includes parametrized, analytical potentials and environmental interactions. The trajectory determined a friction kernel, which was used to calculate the quantum mechanical rate constant using the flux operator formalism.[50] Siebrand and Smedarchina[25] applied time-dependent perturbation theory with a one-dimensional potential surface.

In a recent publication,[26] we explored the hydrogen and deuterium nuclear tunneling process involved in the rate-determining step in the catalytic cycle of the enzyme SLO-1. This step [see Figure 1a] involves the abstraction of a hydrogen atom from the substrate [linoleic acid] by the octahedral $Fe^{3+}-OH$ complex present deep in the active site.[13,16,17,23–25,28–31] The reaction displays a large KIE [k_H/k_D] of 81 at room temperature under certain mutations.[29] In ref 26, we computed the hydrogen tunneling probabilities for a model system constructed from the active site atoms in close proximity to the iron cofactor in SLO-1 [Figure 1b]. This simplification of the active site is based on the assumption that only the immediate environment exerts an electronic influence on the hydrogen nuclear transfer. We described the tunneling hydrogen nucleus [proton or deuteron] as a three-dimensional quantum wavepacket[26,51–55] coupled to the change in electronic structure which was computed using hybrid density functional theory, benchmarked with MP2 post-Hartree−Fock theory. At each step of the quantum dynamics, the potential surface was computed by *including all electrons* in our model system. As a result, our method is not restricted to a specific mode of transfer such as proton coupled electron transfer,[23,24] proton transfer, hydrogen transfer, or hydride transfer. Also, since the transferring nuclear wavepacket is propagated via the time-dependent Schrödinger equation, using an efficient and accurate "distributed approximating functional" propagator,[26,51,52,56,57] all quantum effects pertaining to the quantized H/D nucleus as well as those arising from the electronic degrees of freedom within the model are included. However, it must be noted that the main goal of ref 26 was to evaluate the quantum nuclear contribution to the hydrogen transfer step. This aspect was studied through rigorous quantum dynamics conducted on surfaces created from electronic structure theory. Hence, the exact nature of large-scale rearrangements of the protein that may facilitate gating modes and the contribution of nuclear quantum effects to catalysis were not explicitly probed. Therefore, only reduced active site models [such as in Figure 1a] were considered. Similar models have been used in previous studies on metalloenzymes.[31,58]

The kinetic isotope effect was computed using the transmission amplitude of the wavepacket, and the experimental value was reproduced. Some physical insights gleaned from our studies in ref 26 are as follows: (a) Tunneling for both hydrogen and deuterium occurs through the existence of distorted, spherical "s"-type hydrogen nuclear wavefunctions and "p"-type polarized hydrogen nuclear wavefunctions for transfer along the donor−acceptor axis. (b) There is also a significant population transfer through distorted "p"-type hydrogen nuclear wavefunctions directed perpendicular to the donor−acceptor axis [via intervening "π"-type interactions] which underlines the three-dimensional nature of the tunneling process. The quantum dynamical evolution indicates a significant contribution from tunneling processes both along the donor−acceptor axis and along directions perpendicular to the donor−acceptor axis. (c) The hydrogen nuclear wavefunctions display curve-crossings, in a fashion similar to electronic states. The tunneling process is vibrationally nonadiabatic and is facilitated by these curve-crossings. In our calculations, multiple proton and deuteron excited states (greater than five) were shown to contribute to tunneling. (d) The inclusion of nuclear quantization shifted the transition-state toward the reactants. The precise location of the shifted transition state, however, depends on the populations of each hydrogen and deuterium eigenstate during dynamics.[26]

In this publication, we inspect the hydrogen transfer problem in SLO-1 using the concept of measurement-driven quantum evolution. The enzyme active site is treated as a measurement device. The effect it has on the hydrogen transfer process is represented using the potential energy surfaces computed in ref 26. Thus, while the enzyme active site is not included in an atomistic fashion, its effect is accounted for as stated above. We use this analysis to probe whether the action of the enzyme active site during the hydrogen transfer step of the catalysis process can be described using a measurement paradigm. As we find here, these ideas have utility in providing a qualitative description of the hydrogen transfer step, and we find that such a measurement can accelerate the hydrogen nuclear transfer process as compared to the deuterium transfer process. However, a detailed quantitative description requires the use of quantum dynamics, such as that performed in ref 26. It is important to note that this study focuses on the hydrogen transfer step and hence cannot elucidate the role of measurement on the overall catalytic process.

To facilitate the discussion, we provide a brief overview of the basic ideas of measurement in section II. As will become clear, the mathematical formalisms presented are influenced by those utilized in the fields of quantum information theory and optimal control.[59–61] In this section, we also outline a set of measurement criteria that are used later in section III to make connections to the proton transfer event in SLO-1. In section III, we consider alterations to the measurement and active site atomic evolution sequence to tailor the H/D transfer probability. Conclusions are presented in section IV.

II. Measurement-Induced Control of Quantum Processes

In quantum theory, measuring a system can fundamentally alter its state. Perhaps the most familiar examples of this phenomenon are the sequential Stern-Gerlach experiments, which measure the spin of silver atoms.[62] The Stern-Gerlach experiments pertain to population transfer between $|S_z^{\pm}\rangle$ spinor states through the application of an external magnetic field. As an illustrative example, suppose an ensemble of atoms was prepared that only had the “spin down” component along the z axis, i.e., $|S_z^{-}\rangle$. Next, these atoms are subjected to a magnetic field along the x or y directions, which leads to a “spin measurement.” If the spin along the z axis were again measured, through application of a magnetic field along the z axis, half of the atoms would be spin down, $|S_z^{-}\rangle$, like our original system, but the other half would now be spin up, $|S_z^{+}\rangle$. In other words, the intermediate measurement projected [or altered] the state of the original system such that half of the $|S_z^{-}\rangle$ population is now in a different, orthogonal state, $|S_z^{+}\rangle$. Generally, a Stern-Gerlach experiment is treated as an instantaneous, von Neumann measurement.[63]

In a von Neumann measurement, the initial state is projected onto the eigenstates of the measurement operator. If the outcome of the measurement is recorded, the wavefunction collapses onto a specific measurement operator eigenstate, like $|S_z^{+}\rangle$ in the Stern-Gerlach example. von Neumann measurements have been used in studies detailing how quantum measurements can drive an initial system to a specific set of target states which are orthogonal to the original state.[59,64–68] A quantum system coupled to a classical system or bath can also be interpreted as a quantum measurement process.[59,69–75] [This perspective can be rationalized with the Stern-Gerlach example as well, since the Stern-Gerlach magnet (measurement device) is treated as a classical object, whereas the silver atom spin states are quantum-mechanical.] It is these two properties of quantum measurement theory that we exploit in this study.

We inspect the hydrogen transfer in SLO-1, by invoking the idea that the initial state of the hydrogen nucleus [the donor state] is driven to a final, orthogonal acceptor state [or a finite set of acceptor states] by a series of measurements. Here, the active site in SLO-1, treated as a classical system, constructs a series of measurements on the hydrogen nucleus, a quantum system. Furthermore, our results indicate that the series of measurements enacted by SLO-1 along the reaction path accelerate proton transfer over deuteron transfer. We explore three types of von Neumann measurements, which are described in the following subsections. This tiered set spans a wide range of perceivable measurement-induced perturbations of the quantized hydrogen nucleus due to the active site atoms. They differ from each other and from standard unitary evolution through the discrete elimination of off-diagonal matrix elements or coherences of the density matrix.

In other words, let us first consider the time-evolution of a wavepacket $|\chi(t)\rangle = U(t)|\chi(0)\rangle$, or the density matrix, $U(t)\,\rho_0 U(t)^{\dagger}$, where $U(t)$ is the time-evolution operator appropriate for a reduced-dimensional Hamiltonian, $H(t)$, which depends on an effective time-variable, t. In ref 26, $H(t) = -(\hbar^2/2m_{\mathrm{H}})\nabla_{R_{\mathrm{H}}}{}^2 + V^{\mathrm{DFT}}(R_{\mathrm{H}};\{\mathbf{R_{as}}\};t)$, where R_{H} represents the position of the tunneling proton or deuteron and $V^{\mathrm{DFT}}(R_{\mathrm{H}};\{\mathbf{R_{as}}\};t)$ is the density-functional potential at R_{H}, that also depends on the active site geometry, $\mathbf{R_{as}}$, as seen in Figure 3. It is important to note that, when the Hamiltonian is time-independent, the projected probability of a propagated wavepacket onto a final state, $|\langle f|\chi(t)\rangle|^2$, is generally an oscillatory (periodic) function of time. The periodic nature is defeated to obtain a nearly monotonic form of such a final state projection in ref 26 through the time-dependence of the reduced dimensional Hamiltonian. We treated the tunneling phenomenon in SLO-1 in ref 26 using (a) unitary propagation of a wavepacket on potentials described by the local geometry of the enzyme active site and (b) adaptation of the propagator to the change in the active site geometry. However, as we will see in the next subsections, a similar qualitative effect on the projected probabilities can also be achieved through a measurement operator paradigm, where the measurement operators are determined from the active site geometries and induce H-transfer. Although this type of measurement-induced control has also been analyzed by others[67,68] by including unitary propagation interspersed between a finite number of measurements, in the current publication, we aim to study the effect of measurement [i.e., the active site evolution] alone on the H/D transfer phenomenon.

A. Filtered Measurements. Consider a Hilbert space comprised of the orthogonal kets $\{|D\rangle;|A_m\rangle\}$, where m enumerates the kets comprising a N_A-dimensional subspace. In addition, let $|D\rangle$ denote the initial state of the system, or more generally, $\rho_0 \equiv |D\rangle\langle D|$, and let $\{|A_m\rangle\}$ be the target subspace. In the case of SLO-1, $|D\rangle$ may be regarded as the donor state for the transferring proton wavefunction, whereas $\{|A_m\rangle\}$ is a set of acceptor states. [Note that, while we consider only one $|D\rangle$ state, the treatment is easily generalized.] We wish to drive the $|D\rangle$ state population to the $\{|A_m\rangle\}$ subspace via a series of intermediate measurements, $\{\hat{\mathscr{M}}_I\}$. Thus, the measurements in some sense take the role of active site motions. In the next section, we make this connection more explicit. The dyadic representation of the intermediate measurement operators is $\hat{\mathscr{M}}_I = \sum_j^{N_{DM}(I)} M_j^I |M_j^I\rangle\langle M_j^I| = \sum_j M_j^I P_I^j$, where $\{M_j^I\}$ and $\{|M_j\rangle\}$ are the eigenvalues and eigenvectors of $\hat{\mathscr{M}}_I$, $N_{DM}(I)$ is the dimensionality of the Ith measurement space, and P_I^j is the jth projector of the Ith measurement operator. The projectors resolve the identity, i.e., $\sum_j P_I^j = \mathbb{I}$, where $\mathbb{I}$ is the identity matrix, and are idempotent, i.e., $P_I^j P_I^k = \delta_{j,k} P_I^j$, where $\delta_{j,k}$ is the Kronecker delta. However, for two sets of measurements, $I \neq J$, $P_I^j P_J^k = \langle M_j^I | M_k^J\rangle |M_j^I\rangle\langle M_k^J|$, where $\langle M_j^I | M_k^J\rangle$ is not necessarily $\delta_{j,k}$. That is, the measurement operators do not commute with each other in general and do not have simultaneous eigenstates.

In a filtered measurement scheme, which is also referred to as a selective measurement scheme,[67] the original state, represented as the density matrix ρ_0, is measured by $\hat{\mathscr{M}}_I$, resulting in a new state $\rho_I = \sum_j^{N_{DM}} P_I^j \rho_0 P_I^j$. [To simplify the notation, we have chosen to drop the dependence of N_{DM} on the measurement operator, I.] The $|D\rangle$ and $\{|A_m\rangle\}$ populations of ρ_I are then also observed, i.e., measured. Filtered measurements have been studied for possible use in the field of quantum computation.[64] The probability of finding the system in the $\{|A_m\rangle\}$ subspace after the measurement $\hat{\mathscr{M}}_I$ on ρ_0 is

$$\sum_m^{N_A} \sum_j^{N_{DM}} \langle A_m | P_I^j | D\rangle\langle D | P_I^j | A_m\rangle = \sum_m^{N_A} \sum_j^{N_{DM}} |\langle A_m | M_j^I\rangle\langle M_j^I | D\rangle|^2 \tag{1}$$

Thus, the probability density not in $\{|A_m\rangle\}$, i.e., the probability density remaining in $|D\rangle$, is

$$1 - \sum_m^{N_A} \sum_j^{N_{DM}} |\langle A_m | M_j^I\rangle\langle M_j^I | D\rangle|^2 \tag{2}$$

If the net probability in eq 2 is nonzero, additional measurements may further drive the population from $|D\rangle$ to $\{|A_m\rangle\}$. After a sequence of N_I such measurements, the accumulated probability in the $\{|A_m\rangle\}$ subspace is given by[64]

$$\left\{1 - \prod_{I=1}^{N_I} \left[1 - \sum_m^{N_A} \sum_j^{N_{DM}} |\langle A_m | M_j^I\rangle\langle M_j^I | D\rangle|^2\right]\right\} \tag{3}$$

For the special case of a two-dimensional Hilbert space, comprised of $\{|D\rangle;|A\rangle\}$, eq 3 reduces to

$$\left\{1 - \prod_{I=1}^{N_I} \left[1 - \sum_{j=1}^{2} |\langle A | M_j^I\rangle\langle M_j^I | D\rangle|^2\right]\right\} = \left\{1 - \prod_{I=1}^{N_I} \left[1 - \frac{1}{2}\sin^2 2\theta_I\right]\right\} \tag{4}$$

where $\langle M_1^I | D\rangle \equiv \cos\theta_I$. In further discussions, we refer to this process as a "filtered measurement" process since the component of the state in the $\{|A_m\rangle\}$ subspace is filtered out at each measurement step [eq 2], and therefore, probability only moves in the forward, $|D\rangle \rightarrow \{|A_m\rangle\}$, direction. This is the case, for example, in chemical reactions where the products, once formed, are not available for back-reaction. [A similar process, described by a different realization of measurement theory, is presented in ref 69.] This point becomes clear when one understands eq 1 to be a discrete path integral in Hilbert space that contains only $|D\rangle \rightarrow \{|M_j^I\rangle\} \rightarrow \{|A_m\rangle\}$ paths. Therefore, if a measurement drives $|D\rangle$ to $\{|A_m\rangle\}$, the transfer is complete and further measurements cause no change.

B. Unfiltered Measurements. Unlike the filtered measurement process discussed in the previous section, the "unfiltered measurement", or nonselective measurement process,[67] does not subject the system to a $\{|D\rangle;|A_m\rangle\}$ interrogation after each intermediate measurement. The consequences of this distinction will be explained in the sections below, where we discuss two separate kinds of "unfiltered measurements". The "complete space unfiltered measurement" process is discussed in section II.B.1, and the "reduced space unfiltered measurement" process is discussed in section II.B.2.

1. Complete Space. In the complete space, unfiltered measurement formalism, we represent the system's state after a measurement $\hat{\mathscr{M}}_I$ on ρ_0 as the density matrix

$$\rho_I = \sum_j P_I^j \rho_0 P_I^j = \sum_j |\langle M_j^I | D\rangle|^2 |M_j^I\rangle\langle M_j^I| = \sum_j \rho_j^I P_I^j \tag{5}$$

where ρ_j^I is the probability associated with state $|M_j^I\rangle$. It is convenient, but not necessary, to express the density matrix in the basis of the eigenstates of the measurement operator, $\hat{\mathscr{M}}_I$. However, we no longer enforce a $\{|D\rangle;|A_m\rangle\}$ interrogation in ρ_I, as done in eq 1 of section II.A. The use of unfiltered measurements as an augmentation to optimal control experiments has been studied.[59] Using this notation, it can be shown[59,67] that the result of a sequence of such unfiltered measurements acting on a system is calculated from the recursion relation

$$\rho_{I+1} = \sum_j^{N_{DM}} P_{I+1}^j \rho_I P_{I+1}^j = \sum_{j,i}^{N_{DM}} \rho_i^I |\langle M_j^{I+1} | M_i^I\rangle|^2 P_{I+1}^j \tag{6}$$

After N_I such measurements, the population in the $\{|A_m\rangle\}$ subspace is a discrete sum over paths of the form

$$\sum_m^{N_A} \langle A_m | \rho_{N_I} | A_m\rangle = \sum_m^{N_A} \sum_{j_1,j_2,\cdots,j_{N_I}}^{N_{DM}} |\langle A_m | M_{j_{N_I}}^{N_I}\rangle \cdots \langle M_{j_3}^3 | M_{j_2}^2\rangle\langle M_{j_2}^2 | M_{j_1}^1\rangle\langle M_{j_1}^1 | D\rangle|^2 \tag{7}$$

The above, discrete path integral formalism reflects that all possible paths from $|D\rangle$ to $\{|A_m\rangle\}$ are allowed. Thus, unlike the filtered process in section II.A, the probability is not constrained to flow in one direction in the $\{|D\rangle;|A_m\rangle\}$ space. Thus, unfiltered measurements share characteristics of microscopic reversibility in chemical reactions. This distinction is more explicit if we rewrite the previous equation and only consider two intermediate measurements for simplicity. The summand in eq 7 now takes the form

$$\begin{aligned} |\langle A_m|M^2_{j_2}\rangle\langle M^2_{j_2}|M^1_{j_1}\rangle\langle M^1_{j_1}|D\rangle|^2 = \\ |\langle A_m|M^2_{j_2}\rangle\langle M^2_{j_2}|D\rangle\langle D|M^1_{j_1}\rangle\langle M^1_{j_1}|D\rangle + \\ \langle A_m|M^2_{j_2}\rangle\langle M^2_{j_2}|\hat{\mathscr{R}}|M^1_{j_1}\rangle\langle M^1_{j_1}|D\rangle|^2 \end{aligned} \tag{8}$$

where $\hat{\mathscr{R}} = \mathbb{I} - |D\rangle\langle D|$. The first term in eq 8 contains $|D\rangle \rightarrow |D\rangle$ flow arising from the action of $|M^1_{j_1}\rangle$. Also, a comparison of eqs 7 and 3 reveals that the order of the measurements is only significant in the unfiltered scheme, since the $\{|D\rangle;|A_m|\rangle\}$ filter in the operation described in section II.A disconnects one measurement from the next.

In the current scenario, all measurement eigenstates are explicitly observed. In the following section, we describe a process in which only a portion of the measurement space is observed.

2. *Reduced Space.* In the second unfiltered measurement procedure, the result of a sequence of measurements is obtained from the recursion relation[59,67]

$$\begin{aligned} \rho_{I+1} = \left[\sum_{\{k[I+1]\}} (P^{k[I+1]}_{I+1} \rho_I P^{k[I+1]}_{I+1}) \right] + \\ \left(\mathbb{I} - \sum_{\{k[I+1]\}} P^{k[I+1]}_{I+1}\right) \rho_I \left(\mathbb{I} - \sum_{\{k[I+1]\}} P^{k[I+1]}_{I+1}\right) \end{aligned} \tag{9}$$

Here, the first summation is over the elements in the chosen $\{P^{k[I]}_I\}$ subspace, the elements of which may depend on the measurement index I. Notice further that, while this process does not allow off-diagonal blocks, i.e., $P^{k[I+1]}_{I+1}\rho_{I+1}[\mathbb{I} - P^{k[I+1]}_{I+1}] = 0$, ρ_{I+1} is not necessarily diagonal inside the $[\mathbb{I} - P^{k[I+1]}_{I+1}]$ subspace, as noted by the second term in eq 9. This is an important distinction between the two "unfiltered measurement" processes considered here. In the scheme described in section II.B.1, all off-diagonal elements of ρ in the measurement space are completely eliminated, since all eigenstates are explicitly measured. In this case only the $P^{k[I]}_I$ subspace is explicitly monitored. The remaining projectors are present in the orthogonal complement, $[\mathbb{I} - \sum_{\{k[1]\}} P^{k[I]}_I]$, and the density matrix is no longer diagonal inside the $\{|M^I_i\rangle\}$ basis [compare eqs 5 and 9.] Like the complete, unfiltered formalism described in section II.B.1, probability flow in the $\{|D\rangle;|A_m|\rangle\}$ space is unrestricted. This process creates the $|D\rangle \rightarrow \{|A_m\rangle\}$ flow by measuring only the $\{k[I]\}$ subspace. Thus, it can be seen as a probe of the effectiveness of the $\{k[I]\}$ subspace to effect a measurement-induced population transfer.

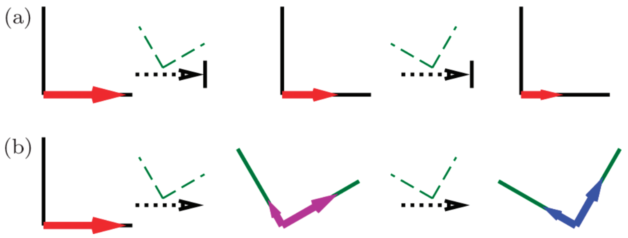
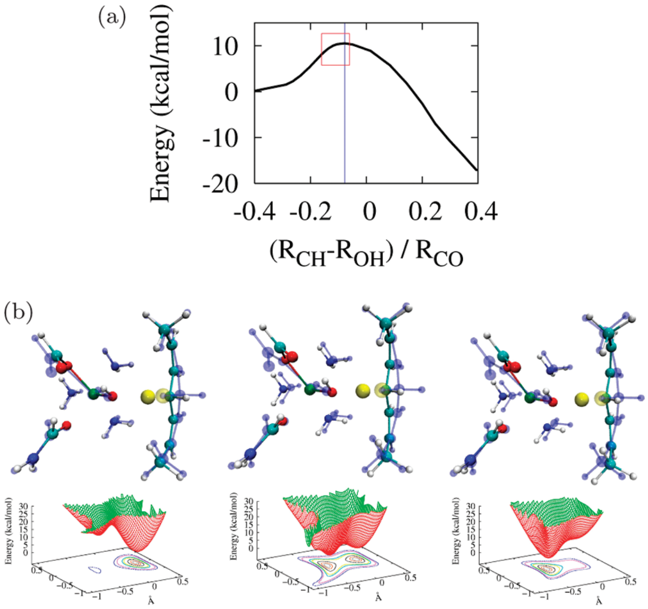


Figure 2. Illustration of filtered (a) and unfiltered (b) measurements. The initial density matrix population is shown using a red, horizontal arrow. The measurements are represented by the horizontal, dotted arrows and the eigenstates of the measurement operators are shown using dark green, dashed lines above these arrows. For example, the eigenstates of the first measurement are oriented at 30° and 120°, while those for the second are oriented at 60° and 150°. The vertical line after the arrow in (a) is the $\{|D\rangle;|A_m\rangle\}$ filter, which removes the vertical, $|A\rangle\langle A|$ component and repopulates the $|D\rangle\langle D|$ dyad. In the unfiltered case, the measurements populate the measurement eigenstates [eq 7], as can be seen from the purple and blue arrows in part (b). The population driven to the $|A\rangle\langle A|$ dyad after both measurements is 0.61 for (a) and 0.56 for (b). [This set of measurements is optimal for (b), as discussed in section III.C.2, whereas the optimal measurement for (a) should be taken at 45° each time.[64]]

A few comments are now in order with respect to the different measurement techniques discussed above. Let us start with the case where an intermediate set of operators $\{\hat{\mathscr{M}}_I\}$ exists, but at each stage there is no measurement process. In other words, $\rho_{I+1} = \sum_{i,j}\rho^{I+1}_{i,j}|M^{I+1}_i\rangle\langle M^{I+1}_j| = (\sum_i|M^{I+1}_i\rangle\langle M^{I+1}_i|)\rho_I(\sum_j|M^{I+1}_j\rangle\langle M^{I+1}_j|)$, and all off-diagonal elements in the measurement operator basis are retained at each stage. This is not a measurement and, in fact, will not evolve ρ_I, since $\sum_i|M^{I+1}_i\rangle\langle M^{I+1}_i| = \mathbb{I}$. This can be distinguished from the case discussed in section II.B.2, where some of the coherences are eliminated through the measurement of a restricted subspace, whereas in section II.B.1, all of the coherences are eliminated. Finally, in section II.A, the process above is further enhanced by an additional measurement or interrogation of the system in the $\{|D\rangle;|A_m\rangle\}$ subspace. These measurements are pictorially represented in Figure 2. A Bloch vector formalism based on the Feynman−Vernon−Hellwarth theory of interaction with strong fields can also be used to depict this within a two-dimensional Hilbert space.[67,76,77] [We note in the limit of a two-dimensional Hilbert space, the complete and reduced unfiltered schemes are equivalent.] One can develop additional evolution schemes where all of the above are combined, and these will be considered as part of future studies.

Finally, it is perceivable that certain sequences of intermediate, noncommuting measurements can maximize $\{\sum_m\langle A_m|\rho_{N_I}|A_m\rangle\}$ and drive $|D\rangle$ to $\{|A_m\rangle\}$ more efficiently than others.[60] For example, one might treat eqs 4, 7, and 9 as multidimensional optimization problems with respect to the unknowns θ_I, $\langle M^I_{j_I}|M^{I-1}_{j_{I-1}}\rangle$, and $\langle M^I_{j_I}|M^{I-1}_{i_{I-1}}\rangle$. Additionally, the measurement operators can be optimized, which will also affect the filtered measurement process. The goal of optimal control experiments,[59,67,68] which make use of measurement-driven evolution, is to find the set of parameters which allow $\{\hat{\mathscr{M}}_I\}$ to maximally drive the $|D\rangle$ state to the $\{|A_m\rangle\}$ states for a given number of measurements. In this paper, we view the enzyme as generating a control field, and we examine how the parameters already "chosen" by SLO-1 affect the proton transfer reaction, utilizing the measurement schemes described above. An exploration of the field parameters is explored in section III.D by changing the measurement sequence.



$\frac{R_{CH}-R_{OH}}{R_{CO}} = -0.204$ $\frac{R_{CH}-R_{OH}}{R_{CO}} = -0.121$ $\frac{R_{CH}-R_{OH}}{R_{CO}} = -0.078$

Figure 3. (a) The minimum energy reaction profile for the rate determining step [Figure 1a] in SLO-1. The critical tunneling region is highlighted with a box, and the classical transition state is shown using a vertical line. The horizontal axis is a reduced coordinate that only includes the distances between the donor carbon, the tunneling proton, and acceptor oxygen. However, it is important to note that, *for every point along this reaction coordinate, the entire active site geometry changes*. That the quantity $(R_{CH} - R_{OH})/R_{CO}$ only indicates a measure of this effect is shown in the upper portions of b, which depict the entire active site model at different points along the reaction coordinate. The transferring hydrogen is enlarged and highlighted in yellow, and the blue transparent system represents the initial, reactant geometry. Finally, the bottom panels of b depict two-dimensional cuts of the three-dimensional, hydrogen nuclear potential surface $[V_I^{DFT}(R_H; \{\mathbf{R_{as}}\})$ of eq 11] at a reactant, tunneling region, and classical transition state active site geometry, respectively. [The exact placement on the reaction coordinate is indicated by the captions.] The *x* and *y* coordinates correspond to the hydrogen nuclear position at each active site geometry.

III. Measurement-Induced Control as Applicable to Hydrogen Nuclear Tunneling in SLO-1

A. Definition of $\{\hat{\mathscr{M}}_I\}$, $|D\rangle$, and $\{|A_m\rangle\}$. In this section, we apply the methods described in section II to the hydrogen transfer in SLO-1. The enzyme active site is treated here as a measurement device. To achieve this, we utilize the potential energy surfaces computed in ref 26, where, as discussed earlier, we studied active site models of SLO-1 using quantum wavepacket dynamical treatment of the transferring H/D nucleus along with treatment of electrons at the level of DFT, benchmarked through MP2 calculations. To maintain correspondence between the active site geometries and the measurement operator, we first define our intermediate measurement operators, $\{\hat{\mathscr{M}}_I\}$, as the hydrogen nuclear Hamiltonians generated by each active site geometry along the reaction path depicted in Figure 3a. It is important to note that, while Figure 3a displays a simplified reaction coordinate, in fact the entire active site geometry [see Figure 3b] changes at each point[26] along the axis. Figure 3b displays a set of selected active site geometries encountered as one moves along the direction indicated by the horizontal axis in Figure 3a. The measurement operators are the effective Hamiltonian operators that describe the dynamics of the tunneling hydrogen or deuterium atom, under the influence of the active site.

$$\hat{\mathscr{M}}_I \equiv \hat{H}_I = \sum_j \varepsilon_j^I |\varepsilon_j^I\rangle\langle\varepsilon_j^I| = \sum_j \varepsilon_j^I P_I^j \tag{10}$$

Here, ε_j^I and $|\varepsilon_j^I\rangle$ are the eigenenergies and eigenvectors of the Hamiltonian $\hat{H}_I$:

$$\hat{\mathscr{M}}_I \equiv \hat{H}_I = -\frac{\hbar^2}{2m_H}\nabla_{R_H}{}^2 + V_I^{DFT}(R_H;\{\mathbf{R_{as}}\}) \tag{11}$$

where, as noted earlier, R_H represents the position of the proton or deuteron and $V_I^{DFT}(R_H; \{\mathbf{R_{as}}\})$ is the density functional potential at R_H, that also depends on the active site geometry, $\{\mathbf{R_{as}}\}$. [The eigenfunctions, eigenenergies, and potential surfaces of each operator were calculated in ref 26 using the Arnoldi diagonalization scheme.[78,79]] Figure 3b depicts the change in potential, $V_I^{DFT}(R_H; \{\mathbf{R_{as}}\})$, in the critical portion of the measurement (Hamiltonian) operators along the direction of the horizontal axis in Figure 3a. Thus, each measurement operator depends on the corresponding active site geometry and electronic structure via the potential energy term in the Hamiltonian, $\hat{H}_I$. [This is similar to the magnetic field in the Stern-Gerlach experiments.] We envision that the state of the proton at each point along the reaction path is influenced by the measurement apparatus [active site geometry].

In the following sections, we examine how these measurement operators may drive the proton from the donor state, identified as $|D\rangle$, to the subspace of acceptor states, $\{|A_m\rangle\}$. To accomplish this, we only consider the tunneling region which was determined in ref 26 to be $-0.\ 2 \le (R_{CH} - R_{OH})/R_{CO} \le -0.06$ and is shown in the boxed region in Figure 3a. The states $|D\rangle$ and $\{|A_m\rangle\}$ were chosen on the basis of how well the proton and deuteron eigenstates are localized in their respective donor carbon and acceptor oxygen basins,[26] and on the donor−acceptor orthogonality condition. Some properties of the chosen states are described in Table 1. [Although only the ground acceptor state is featured, the first five states have similar characteristics to those displayed in the table, e.g., $\cos^{-1}(\langle D|A_m\rangle) = 90.0°$ for $m = 1-5$ for both hydrogen and deuterium.]

B. Filtered Measurements by the SLO-1 Active Site. *1. Single Measurement.* Here, using eq 1, we examine the effect of a single measurement acting on the donor state, to identify critical regions along the reaction coordinate based on the effectiveness of the active site in driving population to the acceptor states. The index I in eq 1 now refers to a point along the reaction coordinate shown in Figure 3a. The results are shown in the middle panels of Figure 4 for (i) one acceptor state and two-dimensional measurements [$N_A = 1$ and $N_{DM} = 2$ in eq 1] and (ii) a subspace of acceptor

Table 1. Summary of the Donor, $|D\rangle$, and Acceptor, $|A_m\rangle$, Subspaces Utilized in the Measurement Process[a]

	$\frac{(R_{CH} - R_{OH})}{R_{CO}}$	$R_{C\langle H\rangle}$ (Å)[b]	$R_{O\langle H\rangle}$ (Å)[c]	σ_{DA} (Å)[d]	θ_{TS}[e]	θ_A[f]
$\|D\rangle$	−0.201	1.11	1.71	0.0738 (0.0612)	89.8 (90.0)	90.0 (90.0)
$\|TS\rangle$	−0.0778	1.68	0.98	0.0822 (0.0652)	0	7.0 (8.1)
$\|A_1\rangle$	−0.0586	1.69	0.98	0.0765 (0.0643)	7.0 (8.1)	0

[a] The values within parentheses are for deuteron. [b] Distance between the $|D\rangle$ (donor), $|TS\rangle$ (classical transition state), and $|A_1\rangle$ (ground acceptor state) centroids and the donor carbon. [c] Distance between the $|D\rangle$, $|TS\rangle$, and $|A_1\rangle$ centroids and the acceptor oxygen. [d] $\langle(R_H^{DA} - \langle R_H^{DA}\rangle)^2\rangle^{1/2}$, where R_H^{DA} is the proton (deuteron is R_D^{DA}) coordinate parallel to the donor−acceptor axis. [e] Angle formed with the classical transition state vector, $|TS\rangle$ in degrees. [$\equiv \cos^{-1}(\langle X|TS\rangle)$], where $|X\rangle$ is $|D\rangle$, $|TS\rangle$ or $|A_1\rangle$. [f] Angle formed with the acceptor ground state vector, $|A_1\rangle$, in degrees. [$\equiv \cos^{-1}(\langle X|A_1\rangle)$].

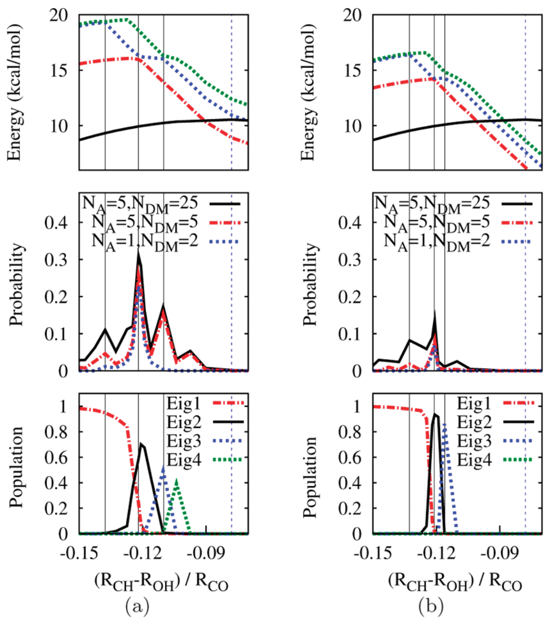


Figure 4. (Top) The classical reaction profile [black curve] and the first three (a) hydrogen and (b) deuterium eigenenergies, along the reaction coordinate. These energies drop below the reaction path surface since the eigenstates close to the top of the barrier are localized on the product side and hydrogen tunneling shifts the "transition state" toward the reactant. The middle panels depict the probability of driving the donor state to the acceptor subspace after one measurement for (a) proton and (b) deuteron taken at the corresponding point on the reaction coordinate. The dimensionalities of the measurements are indicated by N_A and N_{DM} [see eq 3], and the vertical axes' scales are maintained to exemplify the effect of measurement-driven transfer probability for the proton and deuteron. The bottom panels are the instantaneous wavepacket components along the time-dependent eigenstates for the (a) proton and (b) deuteron. Nuclear excited state contributions are critical in both cases. The curve crossings between eigenstates 2 and 3 and the avoided crossings between eigenstates 1 and 2, and 2 and 3 (left to right) are indicated by the gray, solid, and vertical lines. All panels have the same horizontal axis. The position of the classical transition state is also shown using the vertical, dotted, blue line.

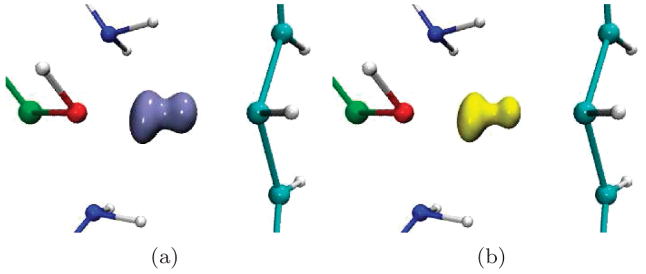


Figure 5. Ground eigenstates for the (a) proton and (b) deuteron at the maximum zero-point corrected energy point, $(R_{CH} - R_{OH})/R_{CO} = -0.121$. The spread along the donor−acceptor direction for the states displayed is $\langle(R_H^{DA} - \langle R_H^{DA}\rangle)^2\rangle^{1/2}$ = 0.240 Å for hydrogen and equal to 0.173 Å for deuterium; i.e., the hydrogen nuclear state is only slightly more delocalized. Additionally, the included angles of these states with the donor ($|D\rangle$) and acceptor ($|A_1\rangle$) states in Table 1 are $\langle H_{-0.121}|D\rangle = \cos(68.1°)$, $\langle D_{-0.121}|D\rangle = \cos(78.4°)$ and $\langle H_{-0.121}|A_1\rangle = \cos(39.5°)$, $\langle D_{-0.121}|A_1\rangle = \cos(28.5°)$. These data represent the fact that both states (in parts a and b) are relatively closer to the acceptor state than to the donor state, which is also clear from the larger left lobe in both figures.

states [$N_A = 5$] and multidimensional measurement operators [N_{DM} = 5 and 25]. The horizontal axis in Figure 4 is the reduced reaction coordinate of Figure 3, and each point corresponds to a measurement operator. The results for hydrogen are presented in Figure 4a, while those for deuterium are presented in Figure 4b.

The top panels in Figure 4 are the eigenstate energy profiles, i.e., ε_j^I from eq 10, as a function of I, and the bottom panels are the dynamical eigenstate populations as calculated from the quantum wavepacket dynamical studies in ref 26. From Figure 4, we note that first, quantum mechanical transfer of population occurs over a range. This range is N_{DM} dependent and is wider for the proton than for the deuteron. [This aspect is also noted for the sequential measurements discussed in section III.B.3 and is further discussed at the end of section III.C.] We also note from the middle panels that the probability for proton transfer into an acceptor state, as a result of the single measurement, is always greater than for deuteron transfer. Furthermore, for each dimensionality studied, a single measurement at $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$ has maximal transition promoting effect. The origin of the measurement operator situated at $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$ coincides with the appearance of an avoided crossing between the ground and first excited hydrogen nuclear eigenfunction as noted in the top panels in Figure 4a and b. This also coincides with the point on the bottom panel in Figure 4a and b where a transfer of population occurs between the ground hydrogen nuclear state to the first excited hydrogen nuclear state. Therefore, the measurement at $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$ may be interpreted as being germane to both proton as well as deuteron transfer. However, this is to be expected on the basis of the delocalized nature of the ground hydrogen nuclear state at $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$ shown in Figure 5. Note further that the point $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$ is not the classical transition state but is situated on the reactant side as seen in Figure 3a. Thus, the measurement at $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$ leads to hydrogen/deuterium tunneling, where hydrogen tunneling is clearly more probable.

Additionally, it is apparent from Figure 4a and b that all measurement operators that have an effect on population transfer from a donor state to an acceptor state are supported by curve crossings in the nuclear eigenstate manifold [top panels of Figure 4]. It is also clear that the curve crossings do not occur at the same points along the proton and deuteron reaction surfaces, which distinguishes the process of hydrogen tunneling from deuterium tunneling from a single measurement perspective. In fact, even the avoided crossing between the ground and first excited state and the associated maximum point in the middle panels of Figure 4 discussed above are not at the same point, although they are very close. This implies that different measurement operators may play a role in promoting H or D, but it appears that the geometries close to −0.121 are of fundamental importance for both transfer processes. This aspect will be probed to a greater extent in section III.D when we study measurement-assisted control of the H/D transfer process, with a greater focus on this region of the reaction coordinate.

The above analysis also shows that curve crossings not only allow nonadiabatic population transfers, as shown in the bottom panels in Figure 4, but also facilitate transfer of population from donor to acceptor states. This is consistent with the relatively large angles between the hydrogen-nuclear eigenstates in this region and the acceptor states, as might be clear upon inspection of Figure 5. *Hence, the idea that the tunneling process may be measurement-driven, where the measurement is constructed by the active site dynamics and its interaction with the tunneling nucleus, is an appealing consequence of this analysis.* It is important to underline the fact that the analysis depicted through the middle panel in Figure 4 only includes the computation of probabilities as dictated by eq 1, for a chosen set of $|D\rangle$ and $\{|A_m\rangle\}$ states and a single measurement operator [$N_I = 1$ in eq 3].

2. Commutators as a Metric to Probe Sensitive Regions on the Reaction Surface. Once a measurement has been made by operator $\hat{\mathscr{M}}_I$, the incremental disturbance due to measurement $\hat{\mathscr{M}}_{I+1}$ is zero if the measurement operators $\hat{\mathscr{M}}_I$ and $\hat{\mathscr{M}}_{I+1}$ commute, since this implies that the operators have simultaneous eigenstates. Therefore, we can quantify the perturbation to the system caused by subsequent measurements by computing the magnitude of $[\hat{\mathscr{M}}_{I+1}, \hat{\mathscr{M}}_I]_\beta \equiv [H_{I+1}, H_I]_\beta \equiv [H_2, H_1]_\beta$, where the subscript β is the inverse temperature and is included to filter out unphysical (i.e., high energy) eigenstates through Boltzmann weighting. In other words, we compute

$$\|[\hat{\mathscr{M}}_{I+1}, \hat{\mathscr{M}}_I]_\beta\|_F \equiv \|[H_{I+1}, H_I]_\beta\|_F \approx \left\|\left[\sum_k \exp[-\beta\varepsilon_k^{I+1}]\varepsilon_k^{I+1}P_{I+1}^k, \sum_j \exp[-\beta\varepsilon_j^I]\varepsilon_j^I P_I^j\right]\right\|_F \tag{12}$$

where "$\|...\|_F$" indicates the Frobenius norm[79] of the commutator.

The evolution of eq 12 for both hydrogen and deuterium at several temperatures is provided in Figure 6. We note from this figure that, in the vicinity of $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$, there is a large spike for the hydrogen and smaller spike for the deuterium commutators at low temperatures. This large change is an indication of the importance of this region, as already established in the previous sections. However, this spike becomes less significant as more excited states are included [i.e., as temperature increases]. As outlined above, eq 12 is a Boltzmannized self-similarity metric of subsequent measurement operators. Clearly, Figure 6 indicates that the self-similarity of subsequent measurement operators in the vicinity of $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$ is greater for D than for H at all temperatures, and quite pronounced at lower temperatures, which in turn leads to the result that the ground states of D, along $(R_{CH} - R_{OH})/R_{CO}$, are more self-similar than H. This has an important effect on the transfer process.

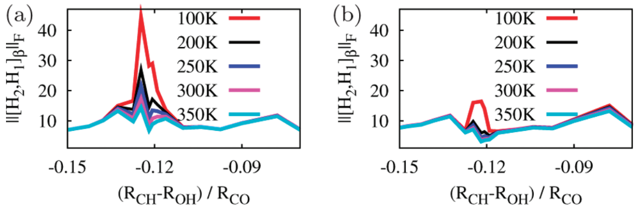


Figure 6. The Frobenius norm of the thermally reduced Boltzmanized subspace commutator [eq 12] at different temperatures depicted for (a) hydrogen and (b) deuterium.

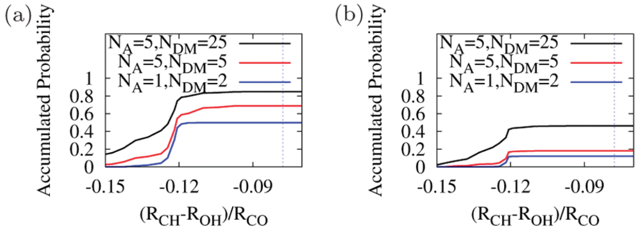


Figure 7. The accumulated probability of driving the donor state to the acceptor subspace after N_I measurements for the (a) proton and (b) deuteron. The dimensionality of the measuring process is indicated by N_A and N_{DM} [see eq 3]. The classical transition state is the vertical, dotted blue line.

3. Sequential Measurements. Next, we consider a sequence of filtered measurements as described by eq 3. Here, the probability of projecting the initial state onto the $\{|A_m\rangle\}$ subspace is accumulated over a set of measurements along the reaction coordinate. This calculation examines the cumulative effectiveness of the concatenated sequence of measurements [i.e., the active site geometric evolution] in promoting the transfer of the hydrogen nucleus. Each intermediate measurement is followed by an interrogation of the donor/acceptor space as described in section II.A. We have utilized $N_I = 29$ in eq 3 [that is, 29 measurements in the tunneling region highlighted within the box in Figure 3a] for the results displayed in Figure 7. Again, we notice that the tunneling from donor to acceptor takes place over an N_{DM}-dependent range of active site geometries similar to those already noted in Figure 4. In addition, the proton transfer probability is always greater than the deuteron probability. The evolution of transfer probabilities follows a sigmoid-like behavior, where the inflection point occurs near $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$, which corresponds to the point in Figure 4 displaying maximal transfer probability from donor to acceptor. Furthermore, we note that H-transfer appears to occur over a wider range in Figure 7, as is

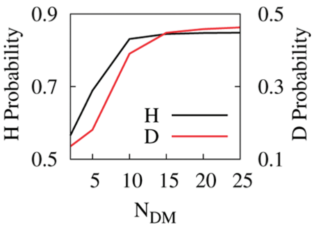


Figure 8. The accumulated probability in the vicinity of the classical transition state as a function of N_{DM} for protium and deuterium. Note that the hydrogen curve plateaus at $N_{DM} \approx$ 10−15, whereas the deuterium still has not fully plateaued at N_{DM} = 25. [Note that the *y*-axis ranges are the same for protium and deuterium.]

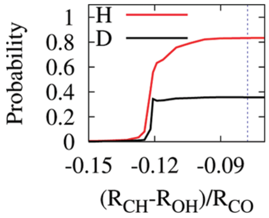


Figure 9. Probability of driving the donor [$|D\rangle$] population to the acceptor [$\{|A_m\rangle\}$] as a function of the reaction coordinate, calculated from eq 7. Here, N_{DM} = 25 and N_A = 5. The blue, dotted, vertical line is the classical transition state.

expected on the basis of the wider range of single measurements that contribute to donor→acceptor probability transfer in Figure 4.

Upon inspection of the converged transfer probabilities in the vicinity of the classical transition state [see Figure 7], we note that deuterium transfer is more sensitive to N_{DM} than protium transfer. Since N_{DM} represents the size of the measurement operator subspace, this implies that excited states participate to a greater extent in deuterium transfer than protium transfer. This property is confirmed through Figure 8, which displays a slower convergence of the transfer probability as a function of N_{DM} for deuterium. A similar result was obtained in ref 26 using quantum wavepacket dynamics.

C. Unfiltered Measurements by the SLO-1 Active Site. *1. Complete Space.* We now consider an unfiltered measurement process over the expanded measurement space (N_{DM} = 25 in section II.B.1), since the expanded space is required to fully describe the transfer in this scheme. The probability of reaching the acceptor space along the reaction coordinate is shown in Figure 9. Although this process favors proton over deuteron transfer and the proton transfer width is greater, a comparison of Figure 9 to Figure 7 shows that the proton and deuteron transfer range is tighter here. In other words, the reaction coordinate range with a growing population in the acceptor states is narrower here since the process depicted in Figure 9 allows for probability to flow freely between the donor and acceptor spaces, whereas the procedure depicted in Figure 7 prevents backflow. Figure 9 shows that the $\{|A_m\rangle\}$ population grows rapidly near $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$, which coincides with the appearance of an avoided crossing between the ground and first excited proton eigenfunctions as discussed in section III.B.1. Again, the proton and deuteron transfer probability become significant before the classical transition state is reached.

2. Reduced Space. Before analyzing the reduced space measurements, we must decide on the $\{P^k\}$ subspace. Henceforth, the explicit dependence of k on I is dropped for simplicity [see eq 9 in section II.B.2]. To gain an understanding of the role of this subspace, we consider $\{P^k\}$ with a span of one [i.e., each $\{P^k\}$ consists of one vector]. We also want to determine the conditions under which this subspace is the dominant path for $|D|\rangle \rightarrow \{|A_m\rangle\}$ transfer. In other words, we wish to choose our subspace to maximize transfer through the path $|\langle A_m|\prod_I P_I^k|D\rangle|^2$, where $P_I^k = |M_k^I\rangle\langle M_k^I|$. Equivalently, we optimize $\prod_I \cos\,\theta_I$, where $\cos\,\theta_I = \langle M_k^{I+1}|M_k^I\rangle$. We place an additional constraint on this optimization problem by requiring that the angles, θ_I, sum to $\phi = \pi/2$. This equality holds if the $\{|D\rangle;|A_m\rangle\}$ space spans the $\{P^k\}$ space. Therefore, we solve

$$\frac{\partial}{\partial\theta_J}\left\{\prod_I \cos\,\theta_I - \lambda\left(\sum_I \theta_I - \phi\right)\right\} = 0 \tag{13}$$

where λ is a Lagrange multiplier equivalent to $-\sin\,\theta_J \prod_{I\neq J} \cos\,\theta_I$. This expression then leads to

$$\frac{\tan\,\theta_J}{\tan\,\theta_K} = 1 \tag{14}$$

for all J and K or

$$\theta_J = \theta_K + n\pi \tag{15}$$

where n is an integer. Thus, if the span of the subspace $\{P^k\}$ is unity, then it is obviously true that all angles, θ_I, must be equal. But if the subspace size is greater than one, the constraint above is modified in that there must exist at least one path $\{\theta_I^{K(I)}\}_{\forall I}$ such that $\sum_I \theta_I^{K(I)} \geq \pi/2$. The equal angle relation may not hold then between eigenstates of consecutive measurement operators. The equal angle result is identical to that derived by Pechen and co-workers in ref 67.

In our case, the ground state projectors, $\{P_I^1\}$, approximately satisfy the equal-angle relation almost everywhere along the reaction profile except in the vicinity of $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$. This central region is complicated by the presence of the avoided crossing and is characterized by a rapid change in wavepacket morphology. Hence, we infer that, while the evolution obeys a two-dimensional Hilbert-space paradigm away from the high-interaction region, the existence of a large number of avoided crossings in the vicinity of $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$ couples the multiple proton vibrational states. This result is consistent with previous studies and strongly suggests the nonadiabatic nature of hydrogen transfer in the vicinity of $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$.

Working from this proposition, matrix elements of ρ_I in eq 9 may be written as

$$\rho_{i,j}^I \equiv \langle M_i^I|\rho_I|M_j^I\rangle = \prod_{k'\in k} |1 - \{\delta_{i,k'} + \delta_{j,k'}\}| \times \sum_{l,n}^{N_{DM}} \rho_{l,n}^{I-1} \cos\,\theta_{M_i^I,\,M_l^{I-1}} \cos\,\theta_{M_j^I,\,M_n^{I-1}} \tag{16}$$

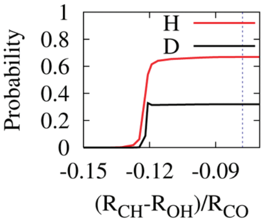


Figure 10. Probability of driving the donor [$|D\rangle$] population to the acceptor [$\{|A_m\rangle\}$] as a function of the reaction coordinate, calculated from eq 16. Here, $N_{DM} = 25$ and $N_A = 5$. The blue, dotted, vertical line is the classical transition state.

Therefore, the density matrix has the property that, after its measurement, only the ground state is diagonal in the $\hat{\mathscr{M}}_I$ representation. This is an important distinction between the measurement paradigm chosen in this section as opposed to that in the previous sections. That is, in this case the measurement does not affect the $[\mathbb{I} - \sum_{\{k\}} P_I^k]\rho_I[\mathbb{I} - \sum_{\{k\}} P_I^k]$ block of the density matrix, since $[\mathbb{I} - \sum_{\{k\}} P_I^k]$ is a reduced-dimensional identity operator and the transfer occurs only due to the elimination of the $P_I^k\rho_I[\mathbb{I} - P_I^k]$ coherences. Therefore, the transfer probabilities are determined by the morphology of the measurement operator eigenstates in the $\{P^k\}$ subspace.

The accumulated transfer probabilities, as seen in Figure 10, have very tight, classical-like proton transfer widths. Furthermore, the transfer probability becomes significant in the vicinity of $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$, which is characterized by a change in morphology of the ground state. Thus, it appears that one set of deviations from classical-like transfers arises when multiple states are measured; i.e., the nonlocality of the transfer paths in the energy representation results in nonlocality of the hydrogen nuclear transfer over the reaction coordinate. [Another reason for deviations from classical behavior is already noted to be due to a shift in the critical transfer region from the classical transition state to $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$, and both of these deviations are consistent with those already seen in the wavepacket dynamics studies in ref 26.] We further find that the final transfer probability is smaller in Figure 10 compared to Figure 9, since an additional 3% of the probability lies beyond the first 25 proton eigenstates at the end of the reaction coordinate and 5% lies beyond the first 25 deuteron eigenstates.

Finally, a few comments are warranted about the results given in the previous sections. First, several of the qualitative features found in the wavepacket dynamics studies in ref 26, such as (a) the shift in the “transition state” toward the reactant as a result of nuclear quantization, (b) greater participation of excited states in the case of deuterium, and (c) the presence of critical points along the reaction coordinate that facilitate hydrogen and deuterium transfer and coincide with surface crossings, are also recovered using the measurement perspective. But, the transfer probabilities obtained in the sections above indicate that measurement alone is not the only factor to consider when looking at the proton transfer in SLO-1. For example, the unitary evolution of the wavepacket on changing potential energy surfaces, which was utilized to reproduce the experimental KIE,[26] is not utilized in the process depicted here. This suggests that the hydrogen transfer process in SLO-1 may be interpreted as a combination of measurement-driven and unitary evolution. Such a combination has been found to accelerate processes in several other studies.[65,67]

Another important factor that should be discussed relates to the role of donor−acceptor distance in tunneling. This aspect has received considerable attention in the literature.[6,9,10,18−21,25,32,80,81] In our quantum dynamics study of ref 26, we found that, as the donor−acceptor distance increased, the H-transfer probability was much greater than the D-transfer probability [see the dotted line in Figure 13 of ref 26]. This result is to be expected, since for a given potential surface, a particle with a larger de Broglie wavelength tunnels through larger distances. This implies that H-tunneling will occur over a larger donor−acceptor distance, and this result is consistent with what we see in ref 26 and also with that obtained from other groups.[18,19,21,25] One can make a similar qualitative deduction on the basis of the broader nature of hydrogen transfer curves in Figures 4 [central panel], 7, 9, and 10, which indicate that hydrogen transfer does occur over a broader donor−acceptor distance. These results are, of course, computed without quantum dynamical evolution, whereas those in ref 26 include quantum dynamical evolution.

D. Measurement Assisted Control. In this section, we describe a numerical experiment we performed that illustrates the control the active site measurement device may exert over the tunneling hydrogen or deuterium nucleus. In this experiment, we explored how a small perturbation to the sequence of measurements [i.e., active site dynamics] might affect the transfer probabilities. *In vivo*, mutations to the amino acid sequence might be responsible for a similar perturbation to the overall dynamics of the active site. [Certain mutations have been shown to noticeably affect transfer properties in SLO-1.[80,82,83]] Our perturbation consists of permuting the order of active site measurements in small regions along the reaction coordinate. In particular, we chose sets of six consecutive measurements, each set from a different region of the reaction coordinate, and permuted the order in each set, one region at a time. For example, while retaining the original sequence in other areas of the reaction coordinate, we permuted all three measurements [or active site geometries] in the range $-0.125 \geq (R_{CH} - R_{OH})/R_{CO} \geq -0.121$ and combined them with permutations of all three geometries in the subsequent range of $-0.119 \geq (R_{CH} - R_{OH})/R_{CO} \geq -0.110$. From the discussion in earlier sections, the transfer probabilities associated with these regions may be expected to be significant. We then used the new order to recalculate eq 7 as an illustration, since this measurement scheme is order sensitive and it describes the entire transfer event with the first 25 proton/deuteron eigenstates. Permuting the measurement operators [and the corresponding active site geometries] is expected to provide an alternative approach to probe the role of active site reorganization on the hydrogen transfer process. For instance, in regions of the reaction coordinate where the active site atoms actively facilitate the H-transfer, one would expect a larger effect from permutation. In an enzyme, suitable active site mutations can give

Table 2. Magnitude and Effect of Perturbation to the Active Site Dynamics Sequence

$(R_{CH} - R_{OH})/R_{CO}$ range	$\langle R_{DA}\rangle^a$	$\Delta R_{DA}{}^b$	$\Delta R_{as}{}^c$	$\Delta\theta^d$	$\rho_{N_I}{}^e$
−0.197 to −0.180	2.78	3.88×10^{-3}	8.54×10^{-5}	2.3 (2.6)	82.4 (35.4)
−0.152 to −0.127	2.69	2.51×10^{-3}	2.43×10^{-4}	5.4 (4.9)	79.3 (35.9)
−0.125 to −0.110	**2.67**	$\mathbf{1.10 \times 10^{-3}}$	$\mathbf{1.57 \times 10^{-4}}$	**12.2 (32.0)**	**69.1 (39.5)**
−0.110 to −0.0671	2.66	6.82×10^{-4}	5.32×10^{-4}	4.8 (5.1)	79.2 (35.2)

[a] The average donor−acceptor distance in Å computed using only geometries in the given reaction coordinate range. [b] Change in the donor−acceptor distance sequence between the perturbed and unperturbed measurement sets: $1/6(\sum_i^6(R_{DA_i} - R_{DA_i}^{\mathrm{pert}})^2)^{1/2}$, in Å. The factor of 6 occurs since each range contains six geometries, all of which are involved in the permutation. [c] Change in the active site geometry sequence between the perturbed and unperturbed measurement sets. This is computed using the active site distance matrices. [d] Angle between the ground proton (deuteron) eigenstate sequence in degrees. [e] Final transfer probability for the proton (deuteron) calculated using eq 7. The original transfer probability is 83.4 (35.6).

rise to a similar effect through structural [electronic and steric] as well as dynamical [fluctuations in active site structure] influence.

We found the final transfer probabilities change most significantly when measurement operators were reordered in the critical range of $-0.124 \geq (R_{CH} - R_{OH})/R_{CO} \geq -0.110$ according to

$$\{\{M1, M2, M3\};\{M4, M5, M6\}\} \rightarrow \{\{M1, M3, M2\};\{M5, M6, M4\}\} \quad (17)$$

where the left-hand side represents the original sequence of measurement operators and active site geometry evolution. The fact that this perturbation has an important impact can also be gauged from the fact that $M3$ above was at $(R_{CH} - R_{OH})/R_{CO} = -0.121$. Thus, the perturbation above has the effect of modifying the active site dynamics in the vicinity of $(R_{CH} - R_{OH})/R_{CO} = -0.121$ according to $\{M2, M3, M4\} \rightarrow \{M3, M2, ..., ..., M4\}$. That is, the ordering of active site dynamics in the vicinity of this critical point is completely changed.

The extent of the perturbation in eq 17 in four regions spanning the reaction coordinate is displayed in Table 2, where the critical region is bold. The perturbations give rise to a combined electronic and structural effect and are quantified as follows: Since altering the dynamical sequence changes the time evolution of donor−acceptor distances, we present a measure of this change, ΔR_{DA}, in the third column. The average donor−acceptor distance inside each perturbed reaction coordinate range is presented in the second column. We also provide a measure of the perturbation to the sequence of active site geometries, ΔR_{as}, in column four and the perturbation to the ground eigenstates, $\Delta\theta$, in column five. The combined [structural and electronic] effect of these perturbations on the final population transfer is presented under the sixth column, labeled ρ_{N_I}.

The following aspects become apparent upon inspection of Table 2. First, we notice that the perturbation in the critical tunneling range (third row) has the largest effect on the transfer probabilities. Changes to this region decrease the proton transfer probability by 14% and increase the deuteron transfer probability by 4%. The original transfer probabilities are 83% for the proton and 36% for the deuteron. Also, the proton transfer probability is more sensitive to the measurement order than the deuteron, and no perturbation that we explored increased the proton transfer probability, although increases in the deuteron transfer probability did occur.

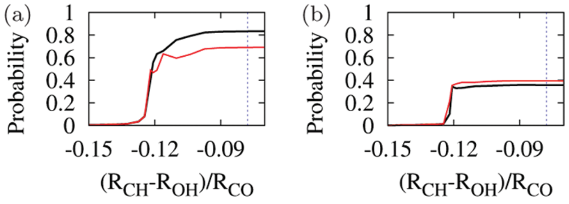


Figure 11. Probability of driving the donor [$|D\rangle$] population to the acceptor [$\{|A_m\rangle\}$] as a function of the reaction coordinate, calculated from eq 7 for (a) the proton and (b) the deuteron. Here, $N_{DM} = 25$ and $N_A = 5$. The black curves represent the unperturbed transfer, whereas the red curves are the transfer when the perturbation described in the text is applied to the critical $-0.125 \geq (R_{CH} - R_{OH})/R_{CO} \geq -0.110$ region. The blue, dotted, vertical line is the classical transition state.

The values in Table 2 again indicate the importance of the critical −0.125 to −0.110 region, and in particular the importance of the active site sequence. Although the size of the perturbation to the active site structure, as measured by ΔR_{DA} and ΔR_{as}, may be relatively larger in other areas, this does not translate into a large effect on the transfer probability. However, the localization of the perturbation to the critical region results in a large difference. The reason for this is seen by examining $\Delta\theta$, which indicates how the potential surfaces affect the hydrogen nuclear eigenstates near the zero-point region. Clearly, the sensitivity of the proton and deuteron eigenstate shape to the changes in the underlying potential energy surfaces, which are double-well in this region [see Figure 3b], is responsible for the large effect on the probability transfer.

The transfer probability curves for the perturbation discussed in this section applied to the critical region are depicted in Figure 11. These figures indicate that the permutation results in a region of net backflow for proton transfer as seen from the reduction in transfer probability of the red curve in Figure 11a to the right side of −0.12. Such backflow can be understood from a transition state theory perspective. If we imagine a dividing surface between $|D\rangle$ and $\{|A_m\rangle\}$, measurements that result in a higher transfer probability are indicative of more forward than backward crossings [i.e., a higher ratio of productive Hilbert space paths in eq 7 to unproductive paths], whereas measurements that result in a lower transfer probability are indicative of the opposite. Thus, these permutations can be seen to move the dividing surface, which results in a changing ratio of forward and backward crossings or productive and unproductive paths.

IV. Conclusions

In a previous publication,[26] we examined the properties of the hydrogen transfer reaction in soybean-lipoxygenase-1 (SLO-1) by computing three-dimensional hydrogen nuclear potential energy surfaces at points along the reaction coordinate using *ab initio* electronic structure methods. From these surfaces, we were able to generate proton and deuteron eigenstates. On the basis of these calculations, we explore a rather fascinating concept in this publication, where the active site motion in SLO-1 projects the hydrogen nuclear state onto intermediate energy eigenstates, which depend on a time-dependent potential. This process may cause measurement-driven evolution of the quantized hydrogen and deuterium atoms. Thus, in this publication, we viewed the SLO-1 active site as a "measurement device" that alters the quantum state of the hydrogen nuclear wavefunction.

Three possible mechanisms for this process were proposed and explored. From these different schemes, we were able to reproduce many of the *qualitative* features found using quantum wavepacket dynamical studies.[26] For instance, we note that the proton and deuteron begin to have significant population transfer near $(R_{CH} - R_{OH})/R_{CO} \approx -0.121$, which occurs *before the classical transition state*. Our results also indicate that excited states play a more important role in deuteron transfer than in proton transfer, since the deuteron transfer probability has a stronger dependence on the measurement operator dimensionality.

The measurement theory paradigm also provides us with new insights. For example, we note in section III.B.1 that the eigenstates located at curve crossings maximize transfer from donor to acceptor subspaces, whereas the avoided crossings in the unitary evolution picture allow for transfer between eigenstates. This result is related to the work of Modi and Shaji,[84] where they show that an experimentally observed[85] anti-Zeno effect occurs only due to the existence of an intermediate state between the ground/bound (donor) and vacuum/decay (acceptor) states. Likewise, at our avoided crossings, there are two, nearly degenerate states which mediate between the donor and acceptor states. Furthermore, we explored the properties of a novel metric for the self-similarity between the consecutive, active-site geometry-dependent Hamiltonians. This measure grows more in the critical transfer region for a proton than a deuteron. This behavior is indicative of an active site sequence that transfers a proton more efficiently than a deuteron.

Finally, we note that perturbations to the order of active site dynamics can have an important effect on the transfer probabilities. In conclusion, the measurement paradigm captures some of the qualitative ideas seen earlier from full quantum wavepacket dynamical studies[26] but does not quantitatively describe the effect of the SLO-1 enzyme active site during the hydrogen transfer step. A quantitative description requires the use of quantum dynamical evolution as discussed in ref 26.

Acknowledgment. This research is supported by the Arnold and Mabel Beckman Foundation (SSI), the National Science Foundation, grant CHE-0750326 (SSI), and the National Institute of Health, grant GM-087475 (SSI). I.S. acknowledges support from the Kraft Fellowship and thanks G. A. Ferguson for helpful discussions.

References

(1) Hynes, J. T.; Klinman, J. P.; Limbach, H.-H.; Schowen, R. L. *Hydrogen-Transfer Reactions*; Wiley-VCH: New York, 2007.

(2) Isaacs, N. *Physical Organic Chemistry*; Wiley & Sons: New York, 1995.

(3) Sheridan, R. *Quantum Mechanical Tunneling in Organic Reactive Intermediates*; Wiley-Interscience: Hoboken, NJ, 2007.

(4) Bell, R. P. *The Proton in Chemistry*; Cornell University Press: Ithaca, NY, 1973.

(5) Brunton, G.; Griller, D.; Barclay, L. R. C.; Ingold, K. U. *J. Am. Chem. Soc.* **1976**, *98*, 6803.

(6) Nagel, Z.; Klinman, J. *Chem. Rev.* **2006**, *106* (8), 3095–3118.

(7) Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes and Solutions*; John Wiley & Sons, Inc.: New York, 1997.

(8) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. *Science* **2004**, *303*, 5655.

(9) Hammes-Schiffer, S.; Benkovic, S. J. *Ann. Rev. Biochem.* **2006**, *75*, 519.

(10) Antoniou, D.; Basner, J.; Nunez, S.; Schwartz, S. *Chem. Rev.* **2006**, *106* (8), 3170–3187.

(11) Cha, Y.; Murray, C. J.; Klinman, J. P. *Science* **1989**, *243*, 1325.

(12) Bahnson, B. J.; Colby, T. D.; Chin, J. K.; Goldstein, B. M.; Klinman, J. P. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 12797.

(13) Liang, Z. X.; Klinman, J. P. *Curr. Opin. Struct. Biol.* **2004**, *14*, 648.

(14) Hwang, J.; Chu, Z.; Yadav, A.; Warshel, A. *J. Phys. Chem.* **1991**, *95* (22), 8445−8448.

(15) Hwang, J. K.; Warshel, A. *J. Am. Chem. Soc.* **1996**, *118* (47), 11745−11751.

(16) Olsson, M. H. M.; Siegbahn, P. E. M.; Warshel, A. *J. Am. Chem. Soc.* **2004**, *126*, 2820–2828.

(17) Olsson, M. H. M.; Siegbahn, P. E. M.; Warshel, A. *J. Biol. Inorg. Chem.* **2004**, *9*, 96–99.

(18) Mavri, J.; Liu, H.; Olsson, M. H. M.; Warshel, A. *J. Phys. Chem. B* **2008**, *112* (19), 5950–5954.

(19) Olsson, M. H. M.; Mavri, J.; Warshel, A. *Phil. Tran. R. Soc. B* **2006**, *361*, 1417–1432.

(20) Olsson, M. H. M.; Parson, W. W.; Warshel, A. *Chem. Rev.* **2006**, *106*, 1737–1756.

(21) Liu, H.; Warshel, A. *J. Phys. Chem. B* **2007**, *111*, 7852–7861.

(22) Pu, J. Z.; Gao, J. L.; Truhlar, D. G. *Chem. Rev.* **2006**, *106*, 3140–3169.

(23) Hatcher, E.; Soudackov, A. V.; Hammes-Schiffer, S. *J. Am. Chem. Soc.* **2007**, *129*, 187 .

(24) Hatcher, E.; Soudackov, A. V.; Hammes-Schiffer, S. *J. Am. Chem. Soc.* **2004**, *126*, 5763–5775.

(25) Siebrand, W.; Smedarchina, Z. *J. Phys. Chem. B* **2004**, *108*, 4185.

(26) Iyengar, S. S.; Sumner, I.; Jakowski, J. *J. Phys. Chem. B* **2008**, *112*, 7601.

(27) Dutton, P. L.; Munro, A. W.; Scrutton, N. S.; Sutcliffe, M. J. *Phil. Trans. R. Soc. London, Ser. B* **2006**, *361* (1472), 1293–1294.

(28) Klinman, J. P. *Pure Appl. Chem.* **2003**, *75*, 601.

(29) Glickman, M. H.; Wiseman, J. S.; Klinman, J. P. *J. Am. Chem. Soc.* **1994**, *116*, 793–794.

(30) Antoniou, D.; Schwartz, S. D. *Proc. Natl. Acad. Sci.* **1997**, *94*, 12360–12365.

(31) Lehnert, N.; Solomon, E. I. *J. Biol. Inorg. Chem.* **2003**, *8*, 294.

(32) Tejero, I.; Garcia-Viloca, M.; Gonzalez-Lafont, A.; Lluch, J. M.; York, D. M. *J. Phys. Chem. B* **2006**, *110*, 24708.

(33) Segraves, E. N.; Holman, T. R. *Biochemistry* **2003**, *42*, 5236–5243.

(34) Garcia-Viloca, M.; Alhambra, C.; Truhlar, D. G.; Gao, J. L. *J. Comput. Chem.* **2003**, *24*, 177–190.

(35) Billeter, S. R.; Webb, S. P.; Agarwal, P. K.; Iordanov, T.; Hammes-Schiffer, S. *J. Am. Chem. Soc.* **2001**, *123*, 11262–11272.

(36) Warshel, A.; Sharma, P.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. *Chem. Rev.* **2006**, *106* (8), 3210–3235.

(37) Kuznetsov, A. M.; Ulstrup, J. *Can. J. Chem.—Rev. Can. Chim.* **1999**, *77*, 1085.

(38) Meyer, M. P.; Klinman, J. P. *Chem. Phys.* **2005**, *319*, 283.

(39) Gillan, M. J. *J. Phys. C* **1987**, *20*, 3621.

(40) Voth, G. A.; Chandler, D.; Miller, W. H. *J. Chem. Phys.* **1989**, *91*, 7749.

(41) Warshel, A.; Chu, Z. T. *J. Chem. Phys.* **1990**, *93*, 4003.

(42) Warshel, A.; Weiss, R. M. *J. Am. Chem. Soc.* **1980**, *102*, 6218.

(43) Chang, Y.-T.; Miller, W. H. *J. Phys. Chem.* **1990**, *94*, 5884.

(44) Day, T. J. F.; Soudachov, A. V.; Cuma, M.; Schmidt, U. W.; Voth, G. A. *J. Chem. Phys.* **2002**, *117*, 5839.

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

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

(47) Singh, B.; Kollman, P. A. *J. Comput. Chem.* **1986**, *7*, 718.

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

(49) Benkovic, S. J.; Hammes-Schiffer, S. *Science* **2006**, *312*, 208.

(50) Miller, W. H.; Schwartz, S. D.; Tromp, J. W. *J. Chem. Phys.* **1983**, *79*, 4889.

(51) Iyengar, S. S.; Jakowski, J. *J. Chem. Phys.* **2005**, *122*, 114105.

(52) Iyengar, S. S. *Theo. Chem. Accts.* **2006**, *116*, 326.

(53) Jakowski, J.; Sumner, I.; Iyengar, S. S. *J. Chem. Theory Comput.* **2006**, *2*, 1203–1219.

(54) Sumner, I.; Iyengar, S. S. *J. Phys. Chem. A* **2007**, *111*, 10313–10324.

(55) Sumner, I.; Iyengar, S. S. *J. Chem. Phys.* **2008**, *129*, 054109.

(56) Hoffman, D. K.; Nayar, N.; Sharafeddin, O. A.; Kouri, D. J. *J. Phys. Chem.* **1991**, *95*, 8299.

(57) Kouri, D. J.; Huang, Y.; Hoffman, D. K. *Phys. Rev. Lett.* **1995**, *75*, 49–52.

(58) Blomberg, M.; Siegbahn, P. *J. Phys. Chem. B* **2001**, *105* (39), 9375–9386.

(59) Shuang, F.; Pechen, A.; Ho, T.; Rabitz, H. *J. Chem. Phys.* **2007**, *126*, 134303.

(60) Tannor, D. J.; Rice, S. A. *J. Chem. Phys.* **1985**, *83* (10), 5013–5018.

(61) Brumer, P.; Shapiro, M. *Acc. Chem. Res.* **1989**, *22* (12), 407–413.

(62) Sakurai, J. J. *Modern Quantum Mechanics*; Addison-Wesley Publishing Company: Reading, MA, 1994.

(63) Wheeler, J. A., Zurek, W. H. Princeton University Press: Princeton, NJ, 1983.

(64) Roa, L.; Delgado, A.; Ladron de Guevara, M. L.; Klimov, A. B. *Phys. Rev. A* **2006**, *73*, 012322.

(65) Roa, L.; Olivares-Renteria, G. A. *Phys. Rev. A* **2006**, *73*, 062327.

(66) Roa, L.; Olivares-Renteria, G. A.; de Guevara, M. L. L.; Delgado, A. *Phys. Rev. A* **2007**, *75*, 014303.

(67) Pechen, A.; Il'in, N.; Shuang, F.; Rabitz, H. *Phys. Rev. A* **2006**, *74*, 052102.

(68) Shuang, F.; Zhou, M.; Pechen, A.; Wu, R.; Shir, O. M.; Rabitz, H. *Phys. Rev. A* **2008**, *78*, 063422.

(69) Prezhdo, O. V. *Phys. Rev. Lett.* **2000**, *85*, 4413.

(70) Jacobs, K.; Steck, D. A. *Contemp. Phys.* **2006**, *47* (5), 279–303.

(71) Diosi, L.; Halliwell, J. J. *Phys. Rev. Lett.* **1998**, *81* (14), 2846.

(72) Halliwell, J. J. *Int. J. Theor. Phys.* **1999**, *38* (11), 2969.

(73) Brun, T. A. *Am. J. Phys.* **2002**, *70* (7), 719.

(74) Mensky, M. *Phys. Lett. A* **1994**, *196*, 159–167.

(75) Mensky, M. B. *Phys. Lett. A* **2003**, *307*, 85–92.

(76) Tannor, D. J. *Introduction to Quantum Mechanics: A Time-dependent Perspective*; University Science Books: New York, 2007.

(77) Feynman, R. P.; Vernon, F. L.; Hellwarth, R. W. *J. Appl. Phys.* **1957**, *28*, 49.

(78) Sorensen, D. C. *SIAM J. Matr. Anal. Apps.* **1992**, *13*, 357–385.

(79) Golub, G. H.; van Loan, C. F. *Matrix Computations*; The Johns Hopkins University Press: Baltimore, MD, 1996.

(80) Knapp, M. J.; Rickert, K.; Klinman, J. P. *J. Am. Chem. Soc.* **2002**, *124*, 3865.

(81) Hay, S.; Scrutton, N. S. *Biochemistry* **2008**, *47*, 9880–9887.

(82) Meyer, M. P.; Tomchick, D. R.; Klinman, J. P. *Proc. Natl. Acad. Sci., U. S. A.* **2008**, *105*, 1146.

(83) Schenk, G.; Neidig, M.; Zhou, J.; Holman, T.; Solomon, E. *Biochemistry* **2003**, *42* (24), 7294–7302.

(84) Modi, K.; Shaji, A. *Phys. Lett. A* **2007**, *368*, 215–221.

(85) Fischer, M. C.; Gutierrez-Medina, B.; Raizen, M. G. *Phys. Rev. Lett.* **2001**, *87*, 040402.

CT900630N

JCTC Journal of Chemical Theory and Computation

A Coarse Grained Model for Atomic-Detailed DNA Simulations with Explicit Electrostatics

Pablo D. Dans, Ari Zeida, Matías R. Machado, and Sergio Pantano*

Institut Pasteur de Montevideo, Mataojo 2020, CP 11400 Montevideo, Uruguay

Received December 4, 2009

Abstract: Coarse-grain (CG) techniques allow considerable extension of the accessible size and time scales in simulations of biological systems. Although many CG representations are available for the most common biomacromolecules, very few have been reported for nucleic acids. Here, we present a CG model for molecular dynamics simulations of DNA on the multi-microsecond time scale. Our model maps the complexity of each nucleotide onto six effective superatoms keeping the “chemical sense” of specific Watson−Crick recognition. Molecular interactions are evaluated using a classical Hamiltonian with explicit electrostatics calculated under the framework of the generalized Born approach. This CG representation is able to accurately reproduce experimental structures, breathing dynamics, and conformational transitions from the A to the B form in double helical fragments. The model achieves a good qualitative reproduction of temperature-driven melting and its dependence on size, ionic strength, and sequence specificity. Reconstruction of atomistic models from CG trajectories give remarkable agreement with structural, dynamic, and energetic features obtained from fully atomistic simulation, opening the possibility to acquire nearly atomic detail data from CG trajectories.

Introduction

Computer simulations have become a reliable tool for the study of structure and dynamics of soft condensed matter systems, as they expose molecular insights that can be difficult or impossible to obtain with experimental techniques. The continuous motivation to expand the limits imposed by the available computer power has prompted scientists to develop simplified representations that reduce the complexity, size, and conformational degrees of freedom of molecular systems while keeping the physical essence of the interactions that rule their behavior.[1] The remarkable improvement in accuracy and reliability achieved by the so-called coarse-grain (CG) representations, together with the development of new algorithms and computer power, offers currently the possibility to reach biologically relevant time scales and system sizes (see ref 2 for an exhaustive review of the latest developments in CG techniques applied to molecular systems). A wide variety of CG representations are available for the most common biological macromolecules, including highly complex lipid−protein systems (see, for instance refs 3 and 4). Nevertheless, only a few implementations have been reported for nucleic acids. Among these applications, notable success has been achieved in the description of DNA structure, dynamics, and melting.[5−8] At the base level, some interesting DNA models inspired us in developing our CG model. Zhang and Collins described the B-DNA as a sequence of rigid bodies (base-ribose) connected by flexible rods. Depending on the type of nucleic base (A/T or G/C), four to five centroids were used in the contraction scheme. Molecular dynamics simulations of thermal melting transition were performed using DNA fragments of 100 base pairs (bp).[9] Tepper and Voth developed a DNA model with explicit solvent particles using 14 uniformly distributed centroids per base pair, covalently linked to reproduce the spontaneous formation of the double helix.[5] In the model by Knotts et al.,[6] each base was reduced to three interaction sites with *ad hoc* potentials for stacking and base pairing. This model successfully reproduced salt-dependent melting, bubble formation, and rehybridization. Using wavelet projection to obtain the effective CG potential between effective centroids, the overall deformation response of a DNA

* Corresponding author. Tel.: +598-2522 0910. Fax: +598-2522 0910. E-mail: spantano@pasteur.edu.uy.

10.1021/ct900653p
Published on Web 04/14/2010

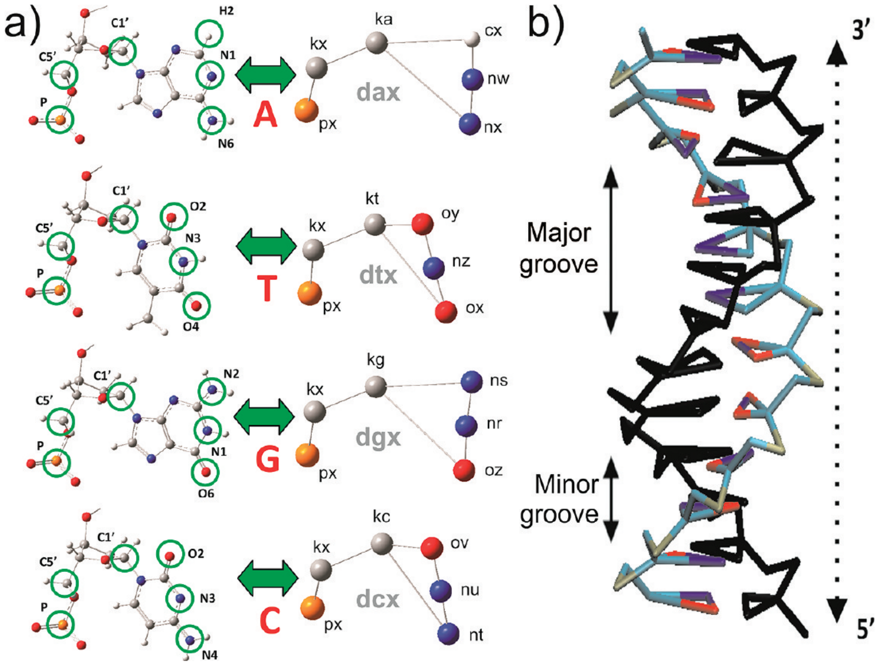


Figure 1. Mapping scheme between atomistic and CG models. (a) Circles highlight the coordinates of the elements from the all-atom representation preserved in the CG model. The residue, superatom, and connectivity are displayed. (b) CG representation of a 12-mer double helix DNA in the canonical B-form that illustrates grooves and 5′−3′ direction (black strand).

molecule was achieved with molecular dynamics (MD) techniques.[7] Representing the DNA as a worm-like polymer and using the "rigid base pair model", homogeneous elastic properties were reproduced by fitting the model against experimental data.[8] In the Mergell et al. model of DNA, each base pair was represented by a rigid ellipsoid linked to the backbone by semirigid harmonic springs.[10] Recently, CG models of DNA were devoted to protein−DNA docking, by optimizing the interaction surface between the macromolecular partners.[11] Similarly, simplified Go-models for RNA have accomplished the description of folding dynamics under varying temperatures and mechanical stretches.[12,13] With a less detailed representation, RNA[14] and also DNA[15] molecules were reduced to only one centroid per nucleotide to study the packing dynamics of a virus genome inside the protein capsid. In this last DNA study, an implicit solvent approach was used to mimic the biological environment.[15] These kinds of models have also been applied with success to the description of large molecular aggregates such as nucleosomes and ribozymes.[16−22]

In this contribution, we present a new CG model for MD simulation of nucleic acids ruled by a Hamiltonian function identical to that used by the most popular MD simulation packages. Electrostatic interactions are treated within the framework of the generalized Born model for implicit solvation.

The model reproduces canonical structures as well as conformational transitions from the A to B form of DNA. We obtain also a good reproduction of the temperature, size, and sequence-specific and ionic strength driven melting. The breathing dynamics of poly(AT) domains were compared with experiments raising comparable life times for end-fraying and also internal hydrogen bonds disruption at the base pair level. Reconstruction of all-atom trajectories from CG MD runs shows a high-quality reproduction of geometrical features with maximum deviations on the order of 2−3 Å with respect to the experimental structures and/or all-atom simulations.

Methods

Coarse Grain Mapping. Our CG model reduces the complexity of a nucleotide to six effective interaction sites (hereafter called superatoms) for each type of canonical nucleotide in DNA (A, T, C and G). This defines four different coarse-grained bases (dax, dtx, dcx, and dgx), which map to the all-atom nucleotides as illustrated in Figure 1a retaining the "chemical sense" of the interactions. Each of the six superatoms was placed on the Cartesian coordinates of one element in the all-atom representation and condensed the molecular information from its atomic neighborhood. The number of superatoms chosen retains the Watson−Crick interaction sites and preserves the asymmetry in the backbone, the identity of the minor and major grooves, as well as the 5′−3′ polarity of the DNA strands (see Figure 1b). Under this scheme, the total mass of the individual atoms of the real nucleotides, including hydrogen, is condensed onto the superatoms, as shown in Table 1.

Phosphate groups are represented by the px superatoms placed on the position of the corresponding phosphorus. The position of the C5′ atom was used to place the superatom kx, which serves to establish the 5′−3′ direction of each DNA strand and allows for the formation of the major and minor grooves (see Figure 1b). The kn superatom (where kn = ka, kt, kc, or kg) lays at the position of the C1′ atom. The superatoms that participate in the Watson−Crick interactions are placed in the same position as the corresponding atoms preserving the molecular specificity between

Table 1. Masses, Charges, and Lennard-Jones Parameters Assigned to the Superatoms

				Lennard-Jones	
superatoms[a]	mass	atoms represented[b]	charges (e)	ε (kcal/mol)	σ (Å)
px	78.97	P+O1P+O2P+O5′	−1.00	0.2000	2.6000
kx	73.07	C5′+C4′+C3′+O3′+O4′	0.00	0.1094	2.4080
ka	41.05	C1′+C2′+N9	0.00	0.1094	1.9080
nx	40.03		0.35	0.1900	1.8240
nw	40.03	(C8+N7+C5+C4+C6+N6+N1+N3+C2)[c]	−0.35	0.1900	1.8240
cx	40.03		0.00	0.1094	1.9080
kt	41.05	C1′+C2′+N1	0.00	0.1094	1.9080
ox	37.03		−0.35	0.2400	1.6612
nz	37.03	(C6+C5+O4+C4+N3+O2+C2+C)[c]	0.70	0.1900	1.8240
oy	37.03		−0.35	0.2400	1.6612
kg	41.05	C1′+C2′+N9	0.00	0.1094	1.9080
oz	45.71		−0.70	0.3100	1.6612
nr	45.71	(C8+N7+C5+C4+N3+C2+N1+C6+O6+N2)[c]	0.35	0.2600	1.8240
ns	45.71		0.35	0.2600	1.8240
kc	41.05	C1′+C2′+N1	0.00	0.1094	1.9080
nt	32.03		0.70	0.2600	1.8240
nu	32.03	(C6+C5+C4+N4+N3+O2+C2)[c]	−0.35	0.2600	1.8240
ov	32.03		−0.35	0.3100	1.6612

[a] The types of the superatoms match those included in the coordinate and topology files that are available from the authors upon request. [b] Hydrogen atoms are omitted for brevity. Their masses are added to the corresponding heavy atoms. [c] The sum of the masses is equally distributed among the three superatoms.

both DNA strands. In this sense, all-atom Watson−Crick hydrogen bonds are shrunk to two-point electrostatic interactions in the CG model.

This scheme leads to an easy mapping/back-mapping from all-atom to CG representation and vice versa. Using internal coordinates and canonical distances, angles, and dihedrals from the B-form of Arnott et al.,[23] we can recover the complete all-atom picture. Dynamic events in the ps−ns time scale can be followed within a multi-microsecond trajectory calculated at the CG level. To this aim, we developed an algorithm that uses as input the instantaneous position of three superatoms to infer the Cartesian coordinates of the atoms in the neighborhood in each MD frame. A Fortran 90 implementation of the homemade algorithm is provided in Table S1 as Supporting Information. The reconstruction to the all-atom picture is made in three steps proceeding from the base to the phosphate moiety (see Table S2 in the Supporting Information for a pseudo-code explaining the algorithm). Since we have less information about the sugar conformation and the dihedrals involved in the phosphodiester bond, a loss of accuracy of the back-mapped coordinates in the backbone region can be expected (see Figure S1 in the Supporting Information). To correct the positioning of the sugar moiety and the distances of the phosphodiester bonds, 150 steps of geometric optimization were performed on each frame after the complete CG to all-atom reconstruction (see Figure S2 in the Supporting Information).

Parameterization. With the aim of maximizing the transferability between different MD packages, our model employs a widely used Hamiltonian function:

$$U = \sum_{\text{bonds}} k_{\text{b}}(r_{ij} - r_{\text{eq}})^2 + \sum_{\text{angles}} k_{\theta}(\theta - \theta_{\text{eq}})^2 + \sum_{\text{dihedrals}} \frac{V_k}{2}[1 + \cos(n_k\varphi - \gamma_k^{\text{eq}})] + \sum_{l}^{N}\sum_{l>m}^{N}\left\{4\varepsilon\left[\left(\frac{\sigma}{r_{lm}}\right)^{12} - \left(\frac{\sigma}{r_{lm}}\right)^{6}\right] + \frac{q_l q_m}{\in r_{lm}}\right\} \tag{1}$$

where k_{b} is the bond stretching constant, $r_{ij} = r_i - r_j$, and r_{eq} is the equilibrium bond distance between two linked elements. k_θ is the bond angle constant. θ is the instantaneous angular value defined by three successive elements, and θ_{eq} is the equilibrium bond angle. V_k is the height of the torsional barrier; n_k is its periodicity. φ is the torsion angle defined by four consecutively bonded elements, and γ_k^{eq} is the phase angle. In the fourth term, the sum runs over all the particles of the system (N). This term corresponds to the Lennard-Jones and Coulombic potentials, in which ε is the maximum depth of the function and σ is the zero energy point or van der Waals diameter. While the values of ε were used as free parameters, those of σ for the backbone superatoms were set to roughly match the excluded volume of the groups of atoms represented (see Table 1). Superatoms participating in the base preserve the σ values coming from the corresponding heavy atoms to avoid artifacts that could disrupt the intra-base-pair step (rise). Lastly, $q_{l,m}$ is the charge of each superatom, and $\in$ is the vacuum permittivity.

Hydration and ionic strength effects were taken into account using the generalized Born (GB) model[24] for implicit solvation as implemented in AMBER.[25] The Born effective radii were fixed to 1.5 Å for all superatoms.

In the present model, the equilibrium bond distances and bond angles were taken from the canonical B-form of Arnott et al.[23] The bond stretching and bond angle constants were fixed to 400 kcal/mol·Å^2 and 75 kcal/mol·rad^2 for all bonds and angles, respectively (eq 1). The torsional barrier for the three dihedral angles of the backbone was fixed to 10 kcal/mol (see Φ, Ξ, and Ψ in Figure 2 and Table 2). The periodicity of dihedral angles was set to nearly reproduce the canonical conformations of the B-form of Arnott et al.[23] To complete the model, two more torsionals, Γ_{dnx} and Ω_{dnx}, that act on the same bond as Ω were added (where dnx stands for each of the four bases: dax, dtx, dgx, and dcx). The parameters for the Γ_{dnx} and Ω_{dnx} dihedral angles, which can be visualized in Figure 2, are specific for each nucleic base. All the torsional parameters used in our model are displayed in Table 2.

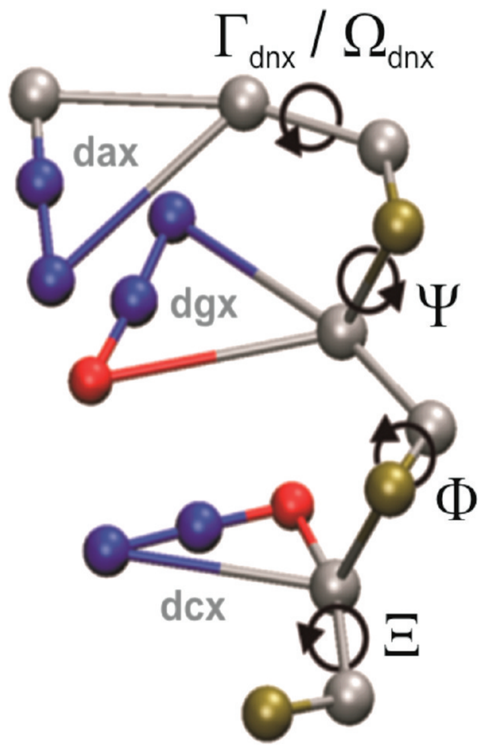


Figure 2. Dihedral angles used in the CG model. Three dihedrals account for the backbone movements for which the parameters are the same regardless of the nucleobase (Φ = kn-*px*-*kx*-kn, Ξ = px-*kx*-*kn*-px, and Ψ = kx-*kn*-*px*-kx where kn = ka, kt, kc, or kg). The dihedral angles Ξ, Ω_{dnx}, and Γ_{dnx} act on the same bond but are defined using different superatoms (dnx = dax, dtx, dgx, dcx). See Table 2 for dihedral angles definition.

Benchmark System: The Drew−Dickerson Dodecamer. To validate the structural, dynamical, and energetic behavior of our CG scheme, the results presented in the first part of this contribution correspond to the Drew−Dickerson dodecamer of DNA (also called the *Eco*RI dodecamer),[26−28] which was used as a benchmark system. This dodecamer of sequence 5′-d(CGCGAATTCGCG)-3′ has been largely studied by means of experimental and theoretical works, giving rise to a solid bibliographic base to compare our results.[29−33] As the starting structure for the CG simulation (labeled DDcgB), the Drew−Dickderson dodecamer was built[25] in the canonical B-form of Arnott el al.[23] During simulation, nonbonded interactions were calculated up to a cutoff of 18 Å within the GB approximation, and the salt concentration was set to 0.15 M. Temperature was controlled using a Langevin thermostat[34,35] with a friction constant of 50 ps^{-1}, which approximates the physical collision frequency for liquid water.[36] The random seed generator of the stochastic force was randomly changed every restart of the simulation (every 1 μs) to avoid quasi-periodic oscillations. The temperature was raised linearly from 0 to 298 K in 5 ns. After that point, production runs of 5 μs were performed, and snapshots were recorded for analysis every 50 ps using a time step of 5 fs to integrate the classical equation of motion. To avoid the fraying of the helix ends frequently observed in long MD simulations,[37] loose harmonic restraints of 3.0 kcal/mol·Å^2 were added to preserve the Watson−Crick hydrogen bonds of the capping base pairs.

To compare our results with state-of-the-art molecular dynamic simulations, the same sequence was built in the Arnott B-form,[23] solvated with explicit water molecules, and surrounded by K^+ and Cl^- ions to mimic the physiological conditions (this system was labeled DDaaB). The all-atom molecular dynamic simulation of the unconstrained Drew−Dickerson dodecamer was performed using the parm99[38] force-field with the correction proposed by Orozco and coworkers for nucleic acids (parmbsc0).[39] Ions were treated with the same force-field. The final system contained 36 K^+, 14 Cl^-, and 3926 TIP3P water molecules[40] in a truncated octahedral box. Initially, the water molecules and ions were relaxed by 1000 steps of energy minimization imposing harmonic restraints of 25 kcal/mol·Å^2 to DNA. Subsequently, four energy minimization runs were performed (with the same number of steps) where the restraints on DNA were gradually reduced from 20 to 5 kcal/mol·Å^2. All optimizations and equilibration MD simulations were performed using constant volume. Long-range interactions were treated using the PME approach[41] with a 12 Å direct space cutoff. The last optimized structure was taken as the starting point for the MD simulations. The entire system was then heated from 0 to 300 K during a 200 ps MD run with harmonic restraints of 5.0 kcal/mol·Å^2 imposed to DNA at a constant volume. Final temperature and a constant pressure of 1 atm were then reached by coupling the system to the Berendsen thermostat and barostat, respectivelly.[42] Fifty nanoseconds of production MD simulation were performed in the isobaric−isothermal ensemble. An integration time step of 2 fs was used, and all

Table 2. Torsional Parameters Used in eq 1 for the CG-DNA Model[a]

	torsional parameters											
dihedral	V_1[b]	V_2	V_3	V_4	n_1	n_2	n_3	n_4	γ_1^{eq}	γ_2^{eq}	γ_3^{eq}	γ_4^{eq}
kn[c]-px-kx-kn (Φ)[c]	10.0				8				161.0			
px-kx-kn-px (Ξ)	10.0				8				–153.2			
kx-kn-px-kx (Ψ)	10.0				4				–29.3			
px-kx-ka-nx (Ω_{dax})	10.0	6.0	7.0	10.0	1	7	2	1	118.0	47.0	20.0	–220.0
px-kx-ka-cx (Γ_{dax})	6.0	4.0	2.0		1	3	4		65.0	145.0	130.0	
px-kx-kt-ox (Ω_{dtx})	10.0	5.0	7.0	10.0	1	8	2	1	117.0	47.0	20.0	–140.0
px-kx-kt-oy (Γ_{dtx})	6.0	4.0	2.0		1	3	4		65.0	145.0	130.0	
px-kx-kg-oz (Ω_{dgx})	10.0	6.5	7.0	10.0	1	6	2	1	110.0	90.0	20.0	–220.0
px-kx-kg-oz (Γ_{dgx})	6.0	4.0	2.0		1	3	4		65.0	145.0	130.0	
px-kx-kc-nt (Ω_{dcx})	10.0	5.0	7.0	10.0	1	8	2	1	117.0	47.0	20.0	–140.0
px-kx-kc-ov (Γ_{dcx})	6.0	4.0	2.0		1	3	4		65.0	135.0	130.0	

[a] See Figure 2 for a comprehensive identification of the Φ, Ξ, Ψ, Ω_{dnx}, and Γ_{dnx} angles. [b] See third term in eq 1. [c] Where kn = ka, kt, kc, or kg.

bond lengths involving hydrogen atoms were restrained using the SHAKE algorithm.[43]

Using the *ptraj* utility of AMBER,[25] root mean square deviations (RMSD) were calculated on all the superatoms/atoms of each residue. The mobility of the bases relative to the backbone was evaluated by comparing atomic B-factors against experimental data. We calculated the quotient between the B-factors of the phosphate atom/superatom and the central heavy atom/superatom engaged in the Watson−Crick interaction (N1 for purines and N3 for pyrimidines). The CG trajectories were back-mapped to all-atom representation and, together with the state-of-the-art MD simulations, analyzed with the program Curves 5.1[44] to monitor the effects of thermal fluctuations upon the major determinants of the B-DNA molecular structure. Root mean square fluctuations (RMSF) and time evolution were calculated for selected helical parameters. The *anal* module of AMBER[25] was used to calculate the interaction energies between bases, strands, GC pairs, and AT pairs in terms of electrostatic and van der Waals contributions. When analyzing back-mapped trajectories, in all the cases, only a discontinuous 50-ns-long trajectory containing the final 10 ns of each microsecond was taken into account for shortness. For comparison purposes, calculated properties were also obtained for crystallographic and averaged NMR derived data (PDB structures 1BNA[45] and 2DAU,[46] respectively).

All MD simulations were carried out using the *sander* module of AMBER 10.[25] Molecular drawings were performed with VMD 1.8.6.[47]

DNA Melting. The CG model was tested to reproduce thermal melting for several systems analyzing the effect of variable length, GC content, and ionic strength of the medium. The sequences chosen were taken from the recently determined experimental work by Owczarzy and co-workers:[48]

(i) 5′-d(ATCGTCTGGA)-3′ (seq10)

(ii) 5′-d(TACTAACATTAACTA)-3′ (seq15a)

(iii) 5′-d(GCAGTGGATGTGAGA)-3′ (seq15b)

(iv) 5′-d(GCGTCGGTCCGGGCT)-3′ (seq15c)

(v) 5′-d(AGCTGCAGTGGATGTGAGAA)-3′ (seq20)

Separated runs were carried out for ionic strengths of 0.07, 0.12, 0.22, and 1.0 M. The melting protocol was the same for each sequence studied and consisted of 3.0 μs of MD simulation in which the temperature was raised 100 °C in five steps of 20 °C. Each step consisted of 0.1 μs of heating followed by 0.5 μs simulated at constant temperature. No restraints were added to the capping base pairs.

To define a melting criterion, hydrogen bonds between base pairs were considered to exist if the distance between the corresponding "acceptor" and "donor" superatom was less than 4.0 Å. The characteristic melting temperature is reached when 50% of the base pairs are in an open state. To generate the melting curves, the percentage of the opened base pairs within the sequence was calculated for each frame of the simulation. Adjacent averaging every 500 frames was performed to clean out the noise. Averaged points were sorted from lowest to highest temperatures, and a sigmoid fit with the Gompertz 4 parameters equation was applied:

$$y_0 + ae^{-e^{-(T-T_0)/b}} \tag{2}$$

This procedure yields one single continuous function of temperature. In eq 2, T_0 is the abscissa of the inflection point, which corresponds to the calculated melting temperature. The regression coefficients for all the sigmoid fits were always >0.8. Results were integrally obtained from the total CG trajectories. Notice that the back-mapping procedure was not applied.

The A to B Transition. The Drew−Dickerson sequence was also built in the A-form of Arnott et al.[23] to test the capability of the model to reproduce a conformational transition from the A to the B form (DDcgA). Five microseconds of coarse grained MD simulations were run under the same conditions used in the DDcgB system. RMSDs with respect to the experimental and canonical B-form structures, pitch, and minor and major groove width were calculated to evaluate the structural transition.

DNA Breathing Dynamics. Finally, we studied the breathing movement of the Drew−Dickerson dodecamer and a 29-bp-long double-stranded DNA: 5′-d(GGCGCCCAATATAAAATATTAAAATGCGC)-3′. The sequence contains a GC clamp domain (G1 to C7) and a long AT track that corresponds to a breathing domain (A8 to A24). The simulation conditions were fixed to roughly match the experimental work by Altan-Bonnet and co-workers.[49] The most relevant difference resided in the fact that the sequence used by Altan-Bonnet et al. contained a thymine tetraloop to avoid the separation of both strands. However, since the structure of this loop is unknown, we decided to replace it by loose harmonic restraints of 3.0 kcal/mol·Å^2 to preserve the Watson−Crick hydrogen bonds of the last base pair (5′-C$_{29}$-3′ in strand1 and 5′-G$_1$-3′ in strand2).

The criterion to define the base opening/closing was identical to that established for melting. MD simulations of 4 μs at 37 °C with an ionic strength of 0.1 M were performed.

Results and Discussion

A major goal for molecular simulations is not only the reproduction of stable trajectories of molecular systems oscillating around equilibrium conformations but also to achieve the capacity to explore the accessible conformational space and evolve toward more stable conformations. In the following paragraphs, we provide some examples of the performance of our model to reproduce the structure, energetics and dynamics of stable trajectories around equilibrium configurations, melting of DNA, conformational transitions, and breathing dynamics.

Benchmark System: CG Model vs All-Atom. All simulations started with the canonical B-form and were stable along all the simulation time. A first measure of the quality of the CG model can be obtained from a direct comparison between the whole trajectories of CG and all-atom representations. To this aim, we calculated the RMSD using all the superatoms in the CG model and the corresponding atoms in the all-atom trajectory (according to the mapping presented in Figure 1). We found that the intrinsic fluctuations during CG and all-atom schemes were very similar. Furthermore, the structural models obtained from both simulations with respect to the experimental structures are practically identical

Table 3. Structural Comparison between CG and All-Atom Simulations[a]

	mean during MD trajectory	starting conformer (B form)	X-ray (1BNA[45])	NMR (2DAU[46])
DDcgB	1.0 ± 0.3	1.8 ± 0.3	2.3 ± 0.3	3.1 ± 0.3
DDaaB	1.6 ± 0.4	2.8 ± 0.4	2.6 ± 0.4	2.7 ± 0.4

[a] RMSD are calculated over 5 μs and 50 ns for the CG and all-atom trajectories, respectively. Values are reported in Angstroms.

(Table 3). Only subtle differences appear when comparing both trajectories against the reference structures.

To analyze the internal flexibility of the dodecamer, B-factors were calculated for selected groups of atoms/superatoms and were compared with the values coming from the X-ray experiments (PDB structure 1BNA). Absolute B-factors calculated from the all-atom trajectory differ significantly from those determined using the CG approach and the X-ray experiments. Only global qualitative trends for the structure as a whole could be obtained. However, the B-factors of the phosphorus elements relative to those of atoms belonging to the base moiety are good descriptors of the relative mobility of different segments of the nucleobases. A comparison between these values indicates that the all-atom simulation (DDaaB) always has the highest mobility, while the coarse-grained version (DDcgB) always presents the lowest (Figure 3). As shown, the relative values were always greater than 1.0 for all the systems, pointing out, as expected, the higher mobility of the backbone with respect to the base. In general, we observe that the relative mobility is lower in the CG model. This can be related to the reduced number of degrees of freedom or to a nonoptimal mass distribution.

Benchmark System: Back-Mapped CG Model vs All-Atom. Despite these encouraging results, it becomes difficult to establish a direct comparison between both simulations. Therefore, we sought to extract atomistic information from our CG model. To this end, we back-mapped the last 10 ns of each microsecond from our CG trajectory (DDcgB). This generated an atomistic noncontiguous 50-ns-long trajectory that is directly comparable with that of the all-atom simulation (DDaaB).

Table 4. Structural Comparisons for the Drew−Dickerson Structure d(CGCGAATTCGCG)$_2$[a,b]

	DDcgB	DDaaB	Arnott-A	Arnott-B	1BNA	2DAU
DDcgB			6.5	1.8	2.3	3.1
DDaaB			5.6	3.0	2.8	2.8
Arnott-A	1.7	2.0		6.3	6.0	4.8
Arnott-B	0.9	1.5	1.5		1.4	3.4
1BNA	1.3	1.2	1.9	0.9		3.3
2DAU	1.5	1.4	1.9	1.5	1.6	

[a] Heavy-atom RMSD between the specified structures. [b] The upper-right portion represents RMSD fit measured in Å calculated over all the atoms. The lower-left portion represents RMSD fit calculated for the four base pairs underlined in the heading, i.e., residues 3−6 and 19−22.

Structural and Dynamical Comparison. Table 4 presents a comparative view of both simulations against the canonical A and B conformations and two experimental structures. The averaged RMSD for the DDaaB simulation was 2.8 Å apart from both the crystallographic (1BNA) and NMR (2DAU) structures and 3.0 Å with respect to the canonical B-form. Analogously, the family of structures obtained with the CG model remained 2.3 Å, 3.1 Å, and 1.8 Å apart from the X-ray structure, NMR structure, and canonical B-form, respectively (upper-right portion of Table 4).

If we consider the averaged RMSD calculated for the selected inner four base pairs (residues 3−6 and 19−22), the values are almost the same between DDcgB and DDaaB with respect to both experimental structures (lower-left portion in Table 4). We can conclude that the differences between all-atom and back-mapped CG simulations are rather subtle, and that both simulations sample very similar or equivalent conformational spaces.

A more stringent evaluation of the quality of the B-form reached by the CG model can be obtained from a comparison of the fluctuations of some selected helical parameters (Figure 4). RMSFs were calculated for the Slide, Rise, Roll and Twist, which are the most distinctive base pairs parameters between the A and B canonical forms (Figure 4a). The large fluctuations observed in the helix ends of DDaaB were not present in DDcgB due to the loose harmonic restraints imposed to preserve the Watson−Crick hydrogen bonds of the capping base pairs in the implicit solvent simulation.

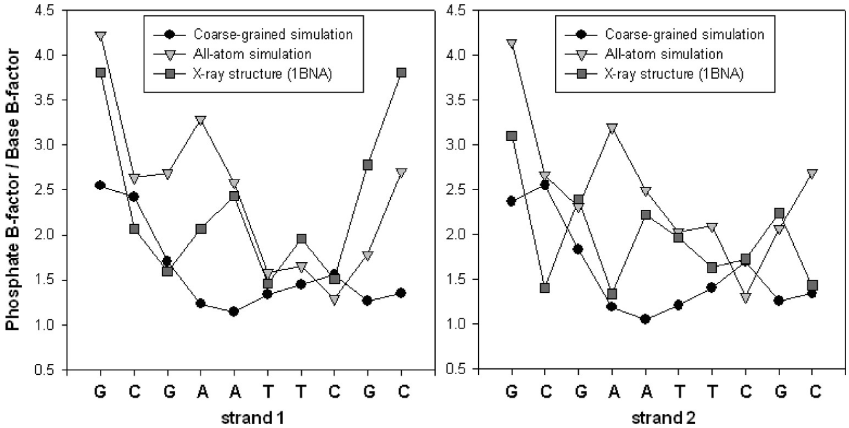


Figure 3. Higher mobility of phosphate groups. B-factors for the phosphorus atoms/superatoms relative to the central elements in the Watson−Crick interaction region along both strands. The coarse-grained (DDcgB) and the all-atom simulation (DDaaB) were compared to the experimental B-factors obtained from the X-ray structure with the PDB code 1BNA.

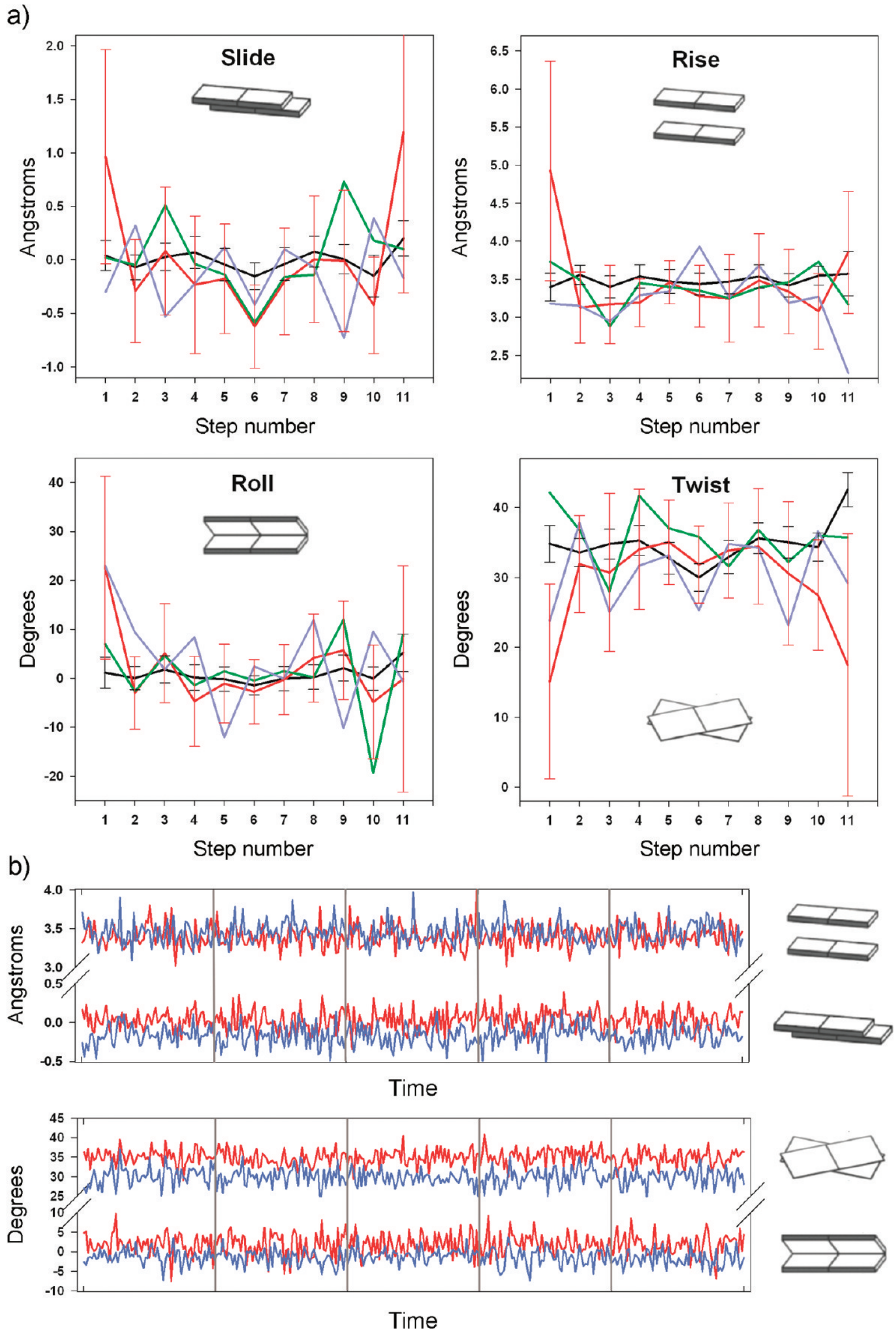


Figure 4. Selected helical parameters. (a) RMSF of the Slide, Rise, Roll, and Twist. The red line corresponds to DDaaB and the black line to DDcgB. Experimental structures 1BNA and 2DAU are represented by the green and the blue lines, respectively. Average values and standard deviations are plotted in Angstroms for the Slide and Rise and in degrees for the Roll and Twist parameters. The values are presented along the helix from the 5′ to 3′ direction (*x* axis). (b) The same helical parameters for two selected intra-base steps (C3/G4 in blue and A6/T7 in red) were plotted along 50 noncontiguous nanoseconds of the back-mapped DDcgB simulation.

Although the fluctuations about the mean values were in general somewhat larger in DDaaB versus DDcgB, the averages exhibited similar trends, especially in the Slide and Twist parameters. Compared to the all-atom simulation, the coarse-grained model exhibited a similar sequence-dependent trend in the Slide and Twist parameters for the CG, GA, AA, AT, TT, and TC dinucleotides (DNA steps 3−8 in Figure 4a).

A more dynamical picture of the structural stability can be acquired following the instantaneous values of the helical parameters during the simulation time. The same selected helical parameters are plotted against time for the back-mapped noncontiguous 50 ns trajectory. For the sake of brevity and clarity, only the C3/G4 and A6/T7 dinucleotides are plotted in Figure 4b. A first global inspection of Figure 4b illustrates the stability of the simulation, as no drift could be observed in the values of the parameters against the simulation time. The Rise and Slide fluctuated around the canonical values, and the Roll showed a distinctive behavior between the C3/G4 and A6/T7 dinucleotides comparable with

Table 5. Comparison of Averaged[a] Electrostatic and van der Waals (VdW) Interactions

	electrostatic (kcal/mol)		VdW (kcal/mol)	
	DDcgB	DDaaB	DDcgB	DDaaB
St1[b] vs St2	1449 ± 38	1434 ± 68	–66 ± 5	–72 ± 8
G4-C21 bp	4 ± 3	6 ± 2	–2 ± 1	–1 ± 1
A5-T20 bp	19 ± 2	11 ± 2	–2 ± 1	–1 ± 1

[a] The averages were calculated over 50 contiguous (DDaaB) or noncontiguous (DDcgB) nanoseconds. [b] St1 stands for strand 1 and St2 for strand 2.

that observed in the all-atom MD simulation.[50] The slight separation between the Twist and Roll traces observed in Figure 4b may suggest a sequence-specific behavior. To shed light on this issue, an exhaustive and systematic study of the helical parameters for all the possible unique combinations of dinucleotides and tetranucleotides (for a total of 146 possible combinations) should be carried out and compared against recent results coming from molecular dynamic simulations.[50,51] Such study is clearly beyond the scope of the present contribution.

Energetic Comparison. In order to further validate the back-mapping procedure and obtain further support on the equivalence between the conformational spaces sampled by the CG and atomistic models, we compared the nonbonded interaction terms of the energy. Calculations were done averaging the results in vacuum using in both cases the same force field (parm99) applied to the all-atom MD and back-mapped trajectories. Comparisons for the van der Waals (VdW) and electrostatic components of the interaction energy between (i) the two strands, (ii) the bases of a GC pair, and (iii) the bases of an AT pair are shown in Table 5.

In light of the correspondent values within the standard deviations, the electrostatic and VdW interactions between DNA strands were virtually the same for both simulations. The good correspondence between both nonbonded interaction terms points out that the conformational space sampled by the CG model was energetically compatible with the state-of-the-art molecular dynamics. Note that the electrostatic contributions in Table 5 are always positive numbers since we computed the Coulombic interaction between two negatively charged strands. When comparing selected GC or AT base pairs, some subtle differences in the averaged electrostatics arise between both approaches. In our back-mapped CG model, the GC base pairs are slightly more stable, whereas the AT base pairs showed an opposite trend. Aimed at acquiring a more global picture, we looked at the electrostatic interactions per residue. For this task, we computed a 12 × 12 electrostatic interactions matrix. The results are presented as an interaction map in Figure 5. A very good correlation between both maps can be observed, providing further support for the compatibility between both approaches.

DNA Melting. Experimentally, the melting temperature (T_0) can be defined for an ensemble of double-stranded DNA molecules as the temperature at which half of the population is in the double-helical state and half in "random-coil" states. This type of definition, which is a good approximation for short DNA sequences, matches with the assumption that melting occurs in a two-state transition. The melting temperature is highly dependent on the length of the double-stranded DNA. Furthermore, because GC base-pairing is generally stronger than AT base-pairing, the amount of guanine and cytosine (called the "GC content") can be estimated by measuring the temperature at which DNA melts. T_0 also depends on the salt concentration or ionic strength of the surrounding medium, as a higher electrostatic screening reduces the mutual repulsion between the negatively charged backbones of each strand in the macromolecule. In other words, T_0 can be used as an indirect measurement of the thermodynamic stability of a double-stranded DNA filament. In terms of the modeling, a good reproduction of the melting process may be indicative of a well-balanced energetic representation of the molecule under study.

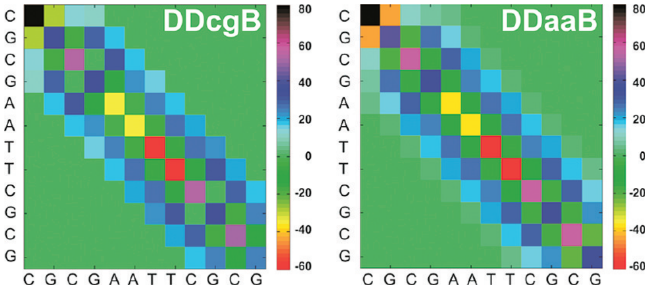


Figure 5. Color map of the averaged electrostatic interaction between the 12 nucleotides within the same strand. Comparison between the back-mapped coarse grained (DDcgB) and the all-atom (DDaaB) simulations. The color scale ranges from −60 to +80 kcal/mol, which are the lower and upper boundary values in the all-atom simulation. It must be noticed that these values were obtained from an effective force field and must not be taken as real energies. The average was calculated over 50 contiguous (DDaaB) or noncontiguous (DDcgB) nanoseconds.

To analyze the energetic features of the CG model, we followed the melting process of five sequences of different lengths, varying also the GC content and the ionic strength according to the Debye−Hückel screening parameter κ.[52] Our results were compared with recent experimental determinations for the same DNA sequences under nearly the same conditions.[48] No back-mapping was performed, as the fraction of native contacts can be measured directly from the CG trajectories.

We studied the length and GC-content dependence of the melting behavior for double-stranded DNA in implicit solvation. Melting temperatures were obtained from single simulations of double-stranded DNA where the temperature was raised in discrete steps of 20° to determine the melting point.

At first glance, good qualitative agreement can be found. As expected, increasing the base pairs number produced a higher T_0 (Figure 6a). Similarly, a higher GC content shifts the T_0 to higher temperatures (Figure 6b). However, in light of standard deviations in the temperature measurement (Table 6), the results could be considered rather qualitative.

There was no variation in T_0 for seq15b at 0.07, 0.12, and 0.22 salt concentrations, for which the calculated melting point was always 63 °C (see Table 6). The only significantly

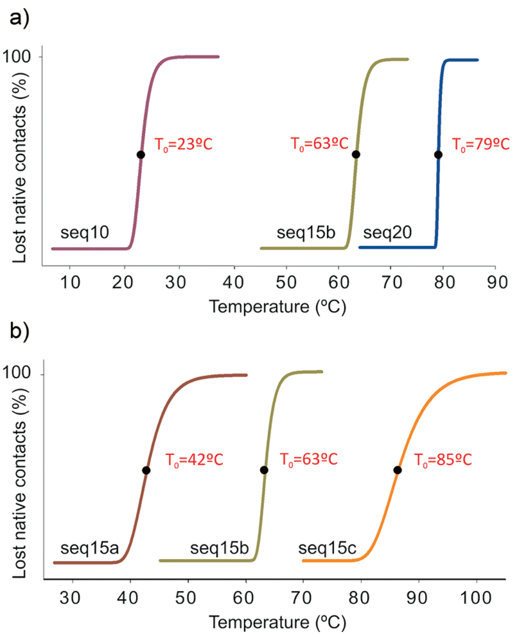


Figure 6. Fittted melting curves. (a) Sequences containing 10 (seq10), 15 (seq15b), and 20 (seq20) bp and 50−53% GC content. (b) Sequences with 15 bp for which the GC content is 20% (seq15a), 53% (seq15b), and 80% (seq15c), respectively. The inflection points (see eq 2) that determine the melting temperatures are indicated with black dots. Notice that the melting curves were obtained after a fitting procedure (see Methods). The numeric values along with the corresponding standard deviations are displayed in Table 6.

different T_0 was obtained at a 1.0 M salt concentration. This is probably due to the way in which the salt effects are incorporated into the GB model. In practice, the linearized Debye−Hückel approximation gives salt effects that are somewhat larger than those predicted by more accurate methods.[52] Saturation of salt effects takes place near 1.0 M, and the best fit with more accurate Poisson−Boltzmann estimations occurs for values from 0.1 to 0.4 M.[52] Previous MD simulations of nucleic acid structures carried out with either a 0.1 or 0.2 M salt concentration showed almost identical results.[53] Recent work describing the melting reaction in DNA hexamers using the same force field (parm99 with the Perez and co-workers modification[39]) and more accurate all-atom simulations for sampling of the free energy landscape also gave only qualitative results.[54]

The aim of this last set of simulations discussed was to test the qualitative dependence of the melting point upon variations of different factors. A precise determination of the melting temperature would need a better sampling such as, for instance, that performed by Knotts and co-workers.[6] They used replica exchange methods to achieve a more quantitative determination. We decided to not perform this kind of calculation, as there is a rather large arbitrariness in the molecular level definition of the melting point. For instance, a small variation (even of tenths of an angstrom) in the cutoff criteria for a native contact between two interacting bases can significantly shift the position of the melting points.

A clear advantage of using MD simulations is that the dynamic behavior of the melting process can be followed on the molecular scale. Thus, sequence- and location-dependent initiation and propagation of the steps that leads to DNA denaturation can be analyzed in detail. In all the sequences studied here, the melting of the helix started from the termini and proceeded toward the center (as an example, the movie for seq15b at 0.12 M is provided in the Supporting Information). This suggests that the loss of internal Watson−Crick interactions has a high-energy cost if the terminal base pairs are still formed as observed in other all-atom simulation work,[54] making internal fraying less frequent.

The A to B Transition. A celebrated result of effective force fields was the capability to reproduce complex conformational changes such as the A to B transition in duplex DNA.[55,56] Therefore, we faced the challenge of reproducing with our CG model the transition from the canonical A to B form, which is the physiologically more stable conformation of double-stranded DNA.

We prepared the same Drew−Dickerson dodecamer studied in the previous section but in the canonical A-form. To follow the A→B transition along the simulation, we calculated the RMSD of all the superatoms with respect to the corresponding atoms in the canonical B-form (see mapping scheme in Figure 1) and the two experimental structures. The results for 5 μs of simulation are shown in Figure 7. The conformational transition took place progressively in a relatively long time window, arriving at final state after nearly 1.2 μs (Figure 7a).

The final RMSD value reached after the transition was 3.3 Å with respect to the canonical B-form, e.g., a value comparable with the deviations obtained from atomistic simulations of duplex B-DNA using the generalized Born approximation.[37]

To reach the final B-form structure (between 1.2 and 5 μs), the conformational transition occurred in three steps:

Table 6. Reference Names and DNA Sequences Used in the Melting Experiments for which the GC Content and Salt Concentrations Are Indicated

reference name	DNA sequence (5′–3′)	GC content (%)	salt concentration (M)	T_0 expt[a] (°C)	T_0 calcd (°C)	st. dev.
seq10	ATCGTCTGGA	50	0.12	37.4	23	25
seq15a	TACTAACATTAACTA	20	0.12	40.4	42	20
seq15b	GCAGTGGATGTGAGA	53	0.07	51.2	63	22
			0.12	54.8	63	22
			0.22	58.0	63	22
			1.00	63.3	100	26
seq15c	GCGTCGGTCCGGGCT	80	0.12	67.7	85	25
seq20	AGCTGCAGTGGATGTGAGAA	50	0.12	63.5	79	19

[a] Taken from ref 48.

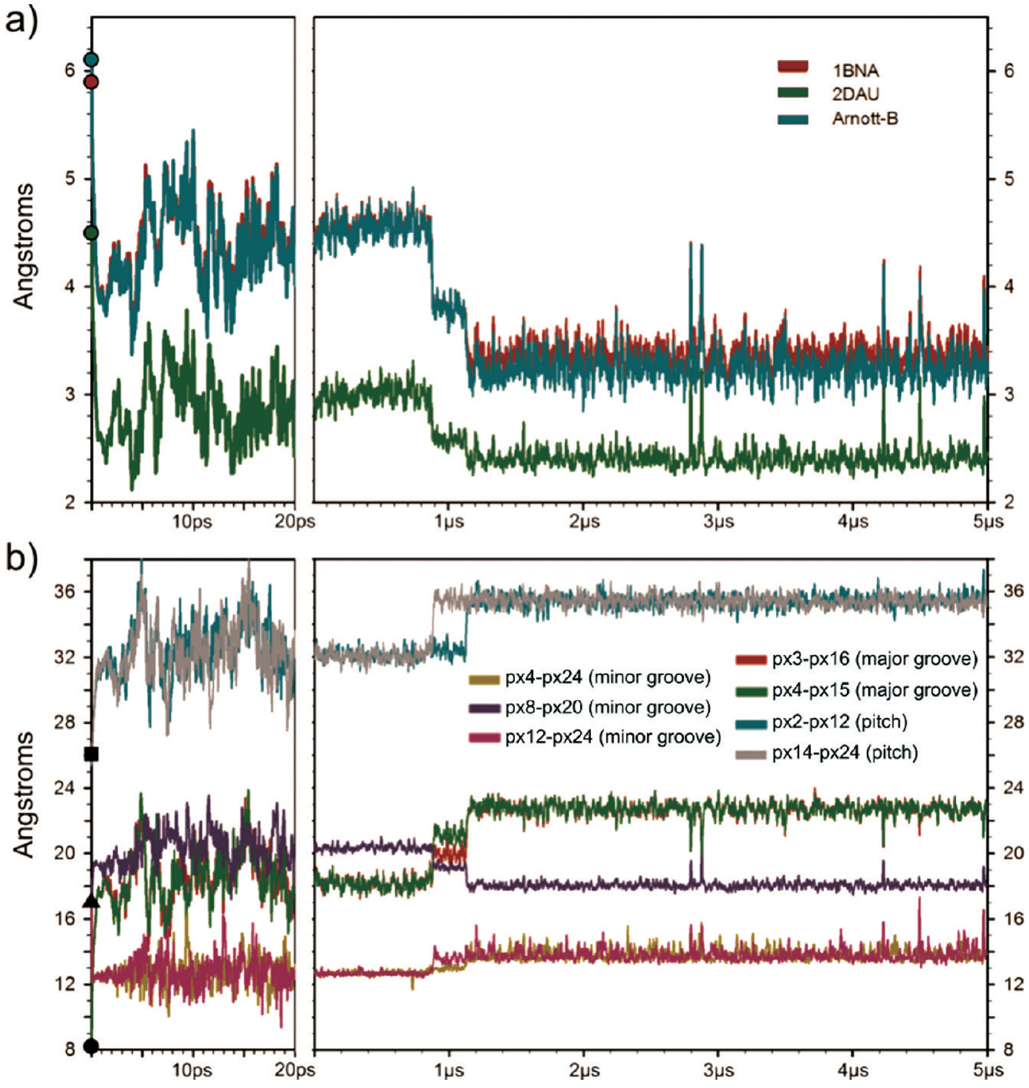


Figure 7. Time evolution of the A to B conformational transition. (a) RMSD using as a reference the canonical B-form (Arnott-B, blue line) and the X-ray and NMR structures (1BNA, dark red line, and 2DAU, green line, respectively). Colored dots indicate the RMSD of the initial conformer with respect to the reference structures. (b) Time evolution of selected distances (pitch, minor and major grooves) during simulation (color codes are indicated in the picture). Black squares, triangles, and circles indicate the starting values for pitch and minor and major grooves, respectively. In both cases, the data shown in the left panels correspond to instantaneous values, while data presented in the right panels correspond to a running average every 200 frames.

(i) In the first few picoseconds (left panel in Figure 7a), the initial structure (canonical A-form) underwent an abrupt conformational change that mainly affected the width of the major groove and, in a second degree, the overall pitch (see Figure 7b). On average, the major groove went from 8 to 18 Å and the pitch from 26 to 32 Å. These changes gave rise to a first cluster of structures 2.6 Å apart from the canonical A-form that remained stable during the first ~900 ns (step 1 in Figure 8). Using the generalized Born model, Tsui and Case[53] showed the convergence from an A-form DNA to a cluster of structures near the B-form within 20 ps of simulation. The quick transition was characterized by the rapid increase of the major groove and the end-to-end length (pitch). The same behavior was observed in the first 20 ps of the CG simulation (Figure 7b). Obviating that the DNA sequence is not strictly the same, visual inspection of the final structure obtained by Tsui and Case[53] after the transition looks very similar to the first cluster of structures obtained in the first picosecond of our CG model (compare the second structure in Figure 8 with Figure 9 in ref 53).

(ii) The following ~300 ns were characterized by a second cluster of structures 3.3 Å apart from the initial structure (first shoulder in Figure 7a). As shown in Figure 7b, the major groove continued to increase from 18 to 21 Å. This movement was followed by a decrease in the wideness of the minor groove measured in the central part of the sequence (from residues 8 and 20, dark blue line). In this case, the pitch underwent an asymmetric transformation to first rearrange the 3′−5′ strand; subsequently the 5′−3′ strand changed its value from 32 to 35 Å (a value very near the 34 Å of the canonical B-form).

(iii) Finally, between 1.2 μs and the end of the simulation, a last cluster of conformers 3.0 Å apart from the reference structure could be found. To reach this last state, the pitch in the 5′−3′ strand went to a final value of 35 Å. The major groove experienced a subsequent increase accompanied by

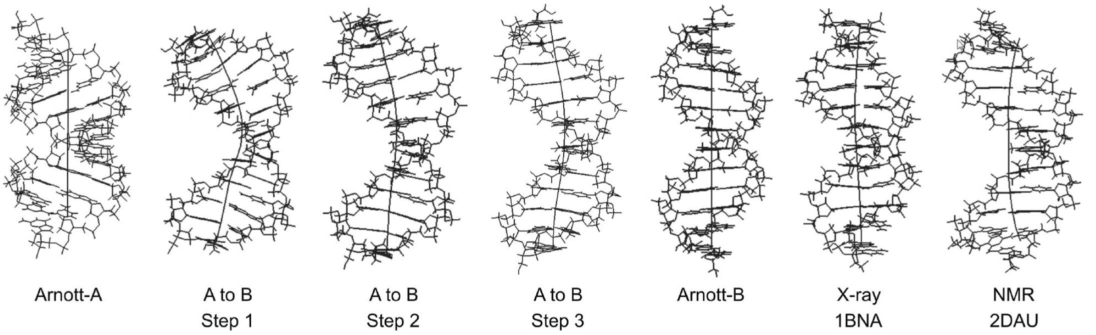


Figure 8. Comparison between back-mapped snapshots and atomistic structures. The conformers labeled steps 1−3 correspond to back-mapped representative snapshots from the conformational A to B transition: steps 1 (0−900 ns), 2 (900−1200 ns), and 3 (1.2−5.0 μs). The DNA axis was calculated with the Curves program.[44]

a ~1 Å narrowing in the minor groove. Note that, along the 5 μs of simulation, the minor groove measured in the extremity of the sequence (between residues 4−24 and 12−16) only underwent slight changes.

In short, the A→B transition can be characterized by global changes in the major structural determinants of double-helical DNA (pitch and groove measurement) in a way that reminds the motion of a "crankshaft". Worth notice is the presence of some peaks in the RMSD after 2 μs of simulation. These correspond to little shifts between the two strands in the AT track that produce transient changes in the minor and major grooves. This behavior was only observed in the central tract and can be associated with breathing movements in the double helix (see next section).

As shown in Figure 8, the conformational changes seem to begin in the central part of the double helix and propagate to the ends, in the same way reported by Cheatham and Kollman in the first simulation on the A to B transition of DNA using all-atom simulations in explicit solvent.[53,55]

The comparison of the A to B transition with the work of Tsui and Case[53] appears to be relevant in the context of the actual time scale sampled by our CG scheme. This is always a complicated issue when dealing with CG simulations, as it is expected that the reduction of degrees of freedom translates to a flattening of the conformational space. The putative correspondence between our work and that of Tsui and Case seems to suggest some equivalence between both simulation schemes. However, the correspondence in the conformational transition may be an artifact of the model that is parametrized to reproduce the B-DNA. To further explore this issue, we sought to test our model against experimental data for which characteristic times ranging from picoseconds to hundreds of microseconds have been reported.

DNA Breathing Dynamics. The microsecond time scale for the full A to B transition begs the question of the correspondence between the real and simulated times. Some insights about this issue can be obtained from a comparison with published simulations on the microsecond time scale. Along the CG simulations of the Drew−Dickerson dodecamer, some transient base pair opening events occurred during the trajectory, especially at the AT pairs. The average lifetime of an open base pair is typically on the order of few picoseconds, but some opening events last for hundreds of picoseconds. These results are in very good agreement with the work of Perez and co-workers,[57] who performed the atomistic simulation of the Drew−Dickerson over 1.2 μs.

Aimed at directly comparing our model with well established experimental results and acquiring a more global perspective, we sought to perform the simulation of a 29-bp-long double-stranded DNA trying to mimic the laboratory conditions.[49] Base pair opening/closing dynamics have been reported for this kind of system on time scales ranging from picoseconds to nanoseconds[58] to hundreds of microseconds.[49] This would allow us to set the time frame of our simulations within a time scale window of near 8 orders of magnitude, covering (i) end-fraying, (ii) breathing, i.e, opening/closing of internal base pairs, and (iii) bubble formation, i.e., temporary opening of internal base pairs implying a partial loss of the double-helical structure.

Following the criterion to define an open state (see the Methods), we calculated the instantaneous state of each base pair (open/close) for each frame of the simulation and the time and sequence extension of those events. As was expected, significantly fewer open states were found in the GC clamp region compared to the AT domain (Figure 9a). Fraying events typically involved few base pairs (typically one or two, Figure 9b) that relax reaching the closed state in dozens to hundreds of picoseconds. This effect is compatible with X-ray,[59] NMR,[60,61] and computer[31,61] studies indicating that fraying is largely confined to the last two base pairs. The CG model also agrees with time-resolved Stokes shifts spectroscopy measurements that restrict the base-opening time to the range of dozens of picoseconds to a few nanoseconds.[62] Nevertheless, during the 4 μs of simulation, we found two events where the end-fraying spread even up to the sixth base pair (Figure 9b,c).

In the AT domain, a nearly continuous breathing dynamic was found along the simulation (Figure 9a), registering several opening/closing events. These events remained in the open state on the nanosecond time scale (see Figure 9b right). The global deformation and the time scale are well comparable with the NMR imino-proton exchange measurements.[58] In this technique, only slight opening of the base pairs, as those observed in the CG model, would be sufficient for the reaction to occur.

Notably, simultaneous opening/closing events with extensions from 2 to 10 consecutive base pairs were frequently observed (Figure 9b). Although with a much shorter time

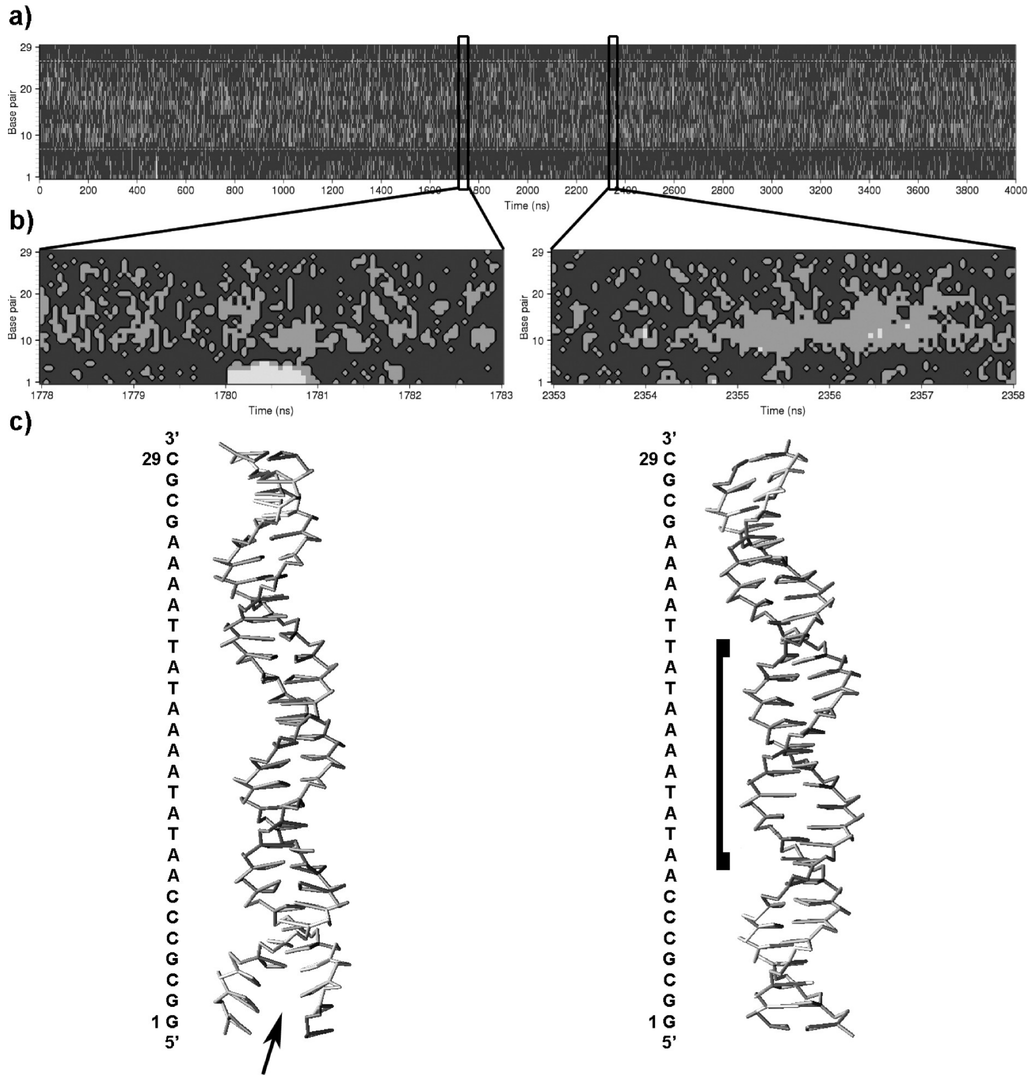


Figure 9. Breathing dynamics of the 29-bp-long double-stranded DNA. Base pairs (*y* axis) are plotted versus time (*x* axis) in nanoseconds. (a) Overview of the breathing along the trajectory. Dark gray color represents closed state base pairs (inter base distance lower than 4 Å). Open states were divided into two ranges: from 4 to 6 Å (light gray) and more than 6 Å (white). White dashed lines delimit the AT breathing domain.[49] (b) Five nanosecond closeups of the trajectory. (c) Representative structures of the end-fraying at the GC clamp (left) and AT breathing domain (right). Fraying and breathing are evidenced with an arrow and square bracket, respectively.

range, these results agree with multiexponential kinetics inferred from fluorescence relaxation times in an analogous molecular system for which opening/closing times of 20−100 μs were reported.[49] It is worth note that these data were obtained from fluorescence quenching experiments, which require a significant distortion in the double-helical structure (bubble formation) in order to be detectable. Such large deformations were never observed along our simulations.

The correspondence with previous theoretical work[57] and NMR studies[58] suggests that the time scale sampled by our model may roughly match the real one. Should this be true, a simulation time on the order of milliseconds would be needed to properly sample the ~100 μs process of bubble formation reported for 29-bp-long double-stranded DNA.[49] Alternatively, the absence of large deformations in our CG simulations could be related to the relative stiffness in the torsional parameters used. A larger number of simulations on different systems and comparison against experimental data are needed to further clarify this point.

Conclusions

We presented herein a nontopological CG model for MD simulations of DNA with explicit electrostatics that offers the possibility to fully recover the atomistic information. Back-mapped CG trajectories gave geometries with maximum deviations of a few angstroms from experimental values, which may be compatible with all-atom simulations offering a considerable speedup. Coarse-grained simulations were carried out in a single node with eight Intel Xeon 2.66 GHz cores at a rate of ~100 μs/superatom/day. At this rate, we performed 1 μs of the coarse-grained simulation using the Drew−Dickerson system in ~1.5 days. Around 850 days would be needed to run 1 μs of the all-atom simulation described herein. Globally, a speedup by a factor of nearly

600 is granted using the CG model. An advantage of the present contribution is that many of the published CG simulation schemes are implemented in *ad hoc* codes or require tailor-made modifications of standard simulation packages, which are often difficult to access and/or operate for the general public. A notable exception of this is the MARTINI force field.[63] The evaluation of the interactions using a classical Hamiltonian allows for a straightforward porting to any other publicly available MD simulation package (topologies and parameters files in AMBER format are available from the authors upon request).

Although the sampling time remains a not completely solved issue, this kind of implementation may open new alternatives to the study of dynamic properties of nucleic acids at longer time scales and for larger systems.

Finally, we would like to stress the fact that the results showed here cover only applications where DNA exists near its B-form. Clearly, Hoogsteen and sugar-edge pairs are out of reach for the present model. This begs the question of whether noncanonical structural motifs can be also well described (structure of telomeres, circular DNA, etc.). This is particularly relevant for the case of RNA where a multiplicity of structural motifs is present (bulges, wobbles, hairpins, and internal loops, etc.). Work is currently ongoing in our group to expand the description to these more challenging cases.

Acknowledgment. This work was supported by ANII (Agencia Nacional de Investigación e Innovación), Programa de Apoyo Sectorial a la Estrategia Nacional de Innovación—INNOVA URUGUAY (Agreement n8 DCI - ALA/2007/19.040 between Uruguay and the European Commission), and Grant FCE_60-2007. M.R.M. is a beneficiary of the National Fellowship System of ANII.

Supporting Information Available: Fortran 90 implementation of the homemade algorithm needed for the reconstruction of the CG trajectories is provided. A pseudocode version explaining the homemade algorithm and two figures illustrating its accuracy (before and after the energy minimization) are also provided along with a movie of the melting process for seq15b at a 0.12 M salt concentration. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Klein, M. L.; Shinoda, W. Large-scale molecular dynamics simulations of self-assembling systems. *Science* **2008**, *321* (5890), 798–800.

(2) Voth, G. A. *Coarse-Graining of Condensed Phase and Biomolecular Systems*, 1st ed.; Taylor & Francis Group: New York, 2009; pp 1−455.

(3) Treptow, W.; Marrink, S. J.; Tarek, M. Gating motions in voltage-gated potassium channels revealed by coarse-grained molecular dynamics simulations. *J. Phys. Chem. B* **2008**, *112* (11), 3277–3282.

(4) Ollila, O. H.; Risselada, H. J.; Louhivuori, M.; Lindahl, E.; Vattulainen, I.; Marrink, S. J. 3D pressure field in lipid membranes and membrane-protein complexes. *Phys. Rev. Lett.* **2009**, *102* (7), 078101.

(5) Tepper, H. L.; Voth, G. A. A coarse-grained model for double-helix molecules in solution: spontaneous helix formation and equilibrium properties. *J. Chem. Phys.* **2005**, *122* (12), 124906.

(6) Knotts, T. A.; Rathore, N.; Schwartz, D. C.; de Pablo, J. J. A coarse grain model for DNA. *J. Chem. Phys.* **2007**, *126* (8), 084901.

(7) Chen, J.-S.; Teng, H.; Nakano, A. Wavelet-based multi-scale coarse graining approach for DNA molecules. *Finite Elem. Anal. Des.* **2007**, *43*, 346–360.

(8) Becker, N. B.; Everaers, R. From rigid base pairs to semi-flexible polymers: coarse-graining DNA. *Phys. Rev. E.: Stat. Nonlin. Soft. Matter Phys.* **2007**, *76* (2 Pt 1), 021923.

(9) Zhang, F.; Collins, M. A. Model simulations of DNA dynamics. *Phys. Rev. E* **1995**, *52* (4), 4217–4224.

(10) Mergell, B.; Ejtehadi, M. R.; Everaers, R. Modeling DNA structure, elasticity, and deformations at the base-pair level. *Phys. Rev. E* **2003**, *68*, 021911.

(11) Poulain, P.; Saladin, A.; Hartmann, B.; Prévost, C. Insights on Protein-DNA Recognition by Coarse Grain Modelling. *J. Comput. Chem.* **2008**, *29*, 2582–2592.

(12) Hyeon, C.; Thirumalai, D. Mechanical unfolding of RNA hairpins. *Proc. Natl. Acad. Sci. U. S. A* **2005**, *102* (19), 6789–6794.

(13) Hyeon, C.; Thirumalai, D. Forced-unfolding and force-quench refolding of RNA hairpins. *Biophys. J.* **2006**, *90* (10), 3410–3427.

(14) Zhang, D.; Konecny, R.; Baker, N. A.; McCammon, J. A. Electrostatic interaction between RNA and protein capsid in cowpea chlorotic mottle virus simulated by a coarse-grain RNA model and a Monte Carlo approach. *Biopolymers* **2004**, *75* (4), 325–337.

(15) Forrey, C.; Muthukumar, M. Langevin Dynamics Simulations of Genome Packing in Bacteriophage. *Biophys. J.* **2006**, *91*, 25–41.

(16) Voltz, K.; Trylska, J.; Tozzini, V.; Kurkal-Siebert, V.; Langowski, J.; Smith, J. Coarse-grained force field for the nucleosome from self-consistent multiscaling. *J. Comput. Chem.* **2008**, *29* (9), 1429–1439.

(17) Hyeon, C.; Dima, R. I.; Thirumalai, D. Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. *Structure.* **2006**, *14* (11), 1633–1645.

(18) Tan, R. K. Z.; Petrov, A. S.; Harvey, S. C. YUP: A Molecular Simulation Program for Coarse-Grained and Multiscaled Models. *J. Chem. Theory Comput.* **2006**, *2*, 529–540.

(19) Korolev, N.; Lyubartsev, A. P.; Nordenskiold, L. Computer modeling demonstrates that electrostatic attraction of nucleosomal DNA is mediated by histone tails. *Biophys. J.* **2006**, *90* (12), 4305–4316.

(20) Wocjan, T.; Klenin, K.; Langowski, J. Brownian Dynamics Simulation of DNA Unrolling from the Nucleosome. *J. Phys. Chem. B* **2009**, *113* (9), 2639−2646.

(21) Langowski, J. Polymer chain models of DNA and chromatin. *Eur. Phys. J. E. Soft. Matter* **2006**, *19* (3), 241–249.

(22) Langowski, J.; Heermann, D. W. Computational modeling of the chromatin fiber. *Semin. Cell Dev. Biol.* **2007**, *18* (5), 659–667.

(23) Arnott, S.; Campbell-Smith, P. J.; Chandrasekaran, R. Handbook of Biochemistry and Molecular Biology, 3rd Nucleic Acids ed.; CRC Press: Cleveland, OH, 1976; Vol. II, pp 411−422.

(24) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.* **1996**, *100*, 19824–19839.

(25) *AMBER 10*; University of California: San Francisco, CA, 2008.

(26) Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E. Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. U. S. A* **1981**, *78* (4), 2179–2183.

(27) Dickerson, R. E.; Drew, H. R. Structure of a B-DNA dodecamer. II. Influence of base sequence on helix structure. *J. Mol. Biol.* **1981**, *149* (4), 761–786.

(28) Drew, H. R.; Dickerson, R. E. Structure of a B-DNA dodecamer. III. Geometry of hydration. *J. Mol. Biol.* **1981**, *151* (3), 535–556.

(29) McConnell, K. J.; Beveridge, D. L. DNA structure: what's in charge. *J. Mol. Biol.* **2000**, *304* (5), 803–820.

(30) Phan, A. T.; Leroy, J. L.; Gueron, M. Determination of the residence time of water molecules hydrating B′-DNA and B-DNA, by one-dimensional zero-enhancement nuclear Overhauser effect spectroscopy. *J. Mol. Biol.* **1999**, *286* (2), 505–519.

(31) Young, M. A.; Ravishanker, D.; Beveridge, D. L. A 5-ns Molecular Dynamics Trajectory for B-DNA: Analysis of Structure, Motions, and Solvation. *Biophys. J.* **1997**, *73*, 2313–2336.

(32) Denisov, V. P.; Carlstrom, G.; Venu, K.; Halle, B. Kinetics of DNA hydration. *J. Mol. Biol.* **1997**, *268* (1), 118–136.

(33) Duan, Y.; Wilkosz, P.; Crowley, M.; Rosenberg, J. M. Molecular dynamics simulation study of DNA dodecamer d(CGCGAATTCGCG) in solution: conformation and hydration. *J. Mol. Biol.* **1997**, *272* (4), 553–572.

(34) Pastor, R. W.; Brooks, B. R.; Szabo, A. An analysis of the accuracy of Langevin and molecular dynamics algorithms. *Mol. Phys.* **1988**, *65*, 1409–1419.

(35) Wu, X.; Brooks, B. R. Self-guided Langevin dynamics simulation method. *Chem. Phys. Lett.* **2003**, *381*, 512–518.

(36) Izaguirre, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D. Langevin stabilization of molecular dynamics. *J. Chem. Phys.* **2001**, *114*, 2090–2098.

(37) Cheatham, T. E., III; Case, D. A. *Computational Studies of RNA and DNA*; Springer: Dordrecht, The Netherlands, 2006; pp 45−71.

(38) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(39) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., III; Laughton, C. A.; Orozco, M. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **2007**, *92* (11), 3817–3829.

(40) Jorgensen, W. L. Transferable Intermolecular Potential Functions for Water, Alcohols, and Ethers. Application to Liquid Water. *J. Am. Chem. Soc.* **1981**, *103*, 335–340.

(41) Darden, T. A.; York, D.; Pedersen, L. Particle mesh Ewald: An $N\cdot\log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(42) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3691.

(43) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

(44) Lavery, R.; Sklenar, H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.* **1988**, *6* (1), 63–91.

(45) Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E. Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **1981**, *78* (4), 2179–2183.

(46) Denisov, A. Y.; Zamaratski, E. V.; Maltseva, T. V.; Sandstrom, A.; Bekiroglu, S.; Altmann, K. H.; Egli, M.; Chattopadhyaya, J. The solution conformation of a carbocyclic analog of the Dickerson-Drew dodecamer: comparison with its own X-ray structure and that of the NMR structure of the native counterpart. *J. Biomol. Struct. Dyn.* **1998**, *16* (3), 547–568.

(47) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.

(48) Owczarzy, R.; You, Y.; Moreira, B. G.; Manthey, J. A.; Huang, L.; Behlke, M. A.; Walder, J. A. Effects of sodium ions on DNA duplex oligomers: improved predictions of melting temperatures. *Biochemistry* **2004**, *43* (12), 3537–3554.

(49) Altan-Bonnet, G.; Libchaber, A.; Krichevsky, O. Bubble dynamics in double-stranded DNA. *Phys. Rev. Lett.* **2003**, *90* (13), 138101.

(50) Lavery, R.; Zakrzewska, K.; Beveridge, D.; Bishop, T. C.; Case, D. A.; Cheatham, T., III; Dixit, S.; Jayaram, B.; Lankas, F.; Laughton, C.; Maddocks, J. H.; Michon, A.; Osman, R.; Orozco, M.; Perez, A.; Singh, T.; Spackova, N.; Sponer, J. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.* **2010**, *38* (1), 299–313.

(51) Dixit, S. B.; Beveridge, D. L.; Case, D. A.; Cheatham, T. E., III; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Sklenar, H.; Thayer, K. M.; Varnai, P. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.* **2005**, *89* (6), 3721–3740.

(52) Srinivasan, J.; Trevathan, M. W.; Beroza, P.; Case, D. A. Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. *Theor. Chem. Acc.* **1999**, (101), 426–434.

(53) Tsui, V.; Case, D. A. Molecular Dynamics Simulations of Nucleic Acids with a Generalized Born Solvation Model. *J. Am. Chem. Soc.* **2000**, *122*, 2489–2498.

(54) Piana, S. Atomistic simulation of the DNA helix-coil transition. *J. Phys. Chem. A* **2007**, *111* (49), 12349–12354.

(55) Cheatham, T. E., III; Kollman, P. A. Observation of the A-DNA to B-DNA transition during unrestrained molecular dynamics in aqueous solution. *J. Mol. Biol.* **1996**, *259* (3), 434–444.

(56) Soliva, R.; Luque, F. J.; Alhambra, C.; Orozco, M. Role of sugar re-puckering in the transition of A and B forms of DNA

in solution. A molecular dynamics study. *J. Biomol. Struct. Dyn.* **1999**, *17* (1), 89–99.

(57) Pérez, A.; Luque, F. J.; Orozco, M. Dynamics of B-DNA on the Microsecond Time Scale. *J. Am. Chem. Soc.* **2007**, *129*, 14739–14745.

(58) Gueron, M.; Leroy, J. L. Studies of base pair kinetics by NMR measurement of proton exchange. *Methods Enzymol.* **1995**, *261*, 383–413.

(59) Holbrook, S. R.; Kim, S. H. Local mobility of nucleic acids as determined from crystallographic data. I. RNA and B form DNA. *J. Mol. Biol.* **1984**, *173* (3), 361–388.

(60) Fujimoto, B. S.; Willie, S. T.; Reid, B. R.; Schurr, J. M. Position-dependent internal motions and effective correlation times for magnetization transfer in DNA. *J. Magn Reson. B* **1995**, *106* (1), 64–67.

(61) Kojima, C.; Ono, A.; Kainosho, M.; James, T. L. DNA duplex dynamics: NMR relaxation studies of a decamer with uniformly 13C-labeled purine nucleotides. *J. Magn. Reson.* **1998**, *135* (2), 310–333.

(62) Andreatta, D.; Sen, S.; Pérez Lustres, J. L.; Kovalenko, A. E. N. P. M. C. J.; Coleman, R. S. B M. A. Ultrafast Dynamics in DNA: "Fraying" at the End of the Helix. *J. Am. Chem. Soc.* **2006**, *128*, 6885–6892.

(63) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **2007**, *111* (27), 7812–7824.

CT900653P

JCTC Journal of Chemical Theory and Computation

Ligand Affinities Estimated by Quantum Chemical Calculations

Pär Söderhjelm,[†] Jacob Kongsted,[‡] and Ulf Ryde*,[†]

Department of Theoretical Chemistry, Lund University, Chemical Centre, P.O. Box 124, 221 00 Lund, Sweden, and Department of Physics and Chemistry, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark

Received December 30, 2009

Abstract: We present quantum chemical estimates of ligand-binding affinities performed, for the first time, at a level of theory for which there is a hope that dispersion and polarization effects are properly accounted for (MP2/cc-pVTZ) and at the same time effects of solvation, entropy, and sampling are included. We have studied the binding of seven biotin analogues to the avidin tetramer. The calculations have been performed by the recently developed PMISP approach (polarizable multipole interactions with supermolecular pairs), which treats electrostatic interactions by multipoles up to quadrupoles, induction by anisotropic polarizabilities, and nonclassical interactions (dispersion, exchange repulsion, etc.) by explicit quantum chemical calculations, using a fragmentation approach, except for long-range interactions that are treated by standard molecular-mechanics Lennard-Jones terms. In order to include effects of sampling, 10 snapshots from a molecular dynamics simulation are studied for each biotin analogue. Solvation energies are estimated by the polarized continuum model (PCM), coupled to the multipole-polarizability model. Entropy effects are estimated from vibrational frequencies, calculated at the molecular mechanics level. We encounter several problems, not previously discussed, illustrating that we are first to apply such a method. For example, the PCM model is, in the present implementation, questionable for large molecules, owing to the use of a surface definition that gives numerous small cavities in a protein.

Introduction

A major goal within theoretical chemistry is to accurately predict the free energy for the binding of a ligand to a macromolecule. If such binding affinities could be accurately predicted, large parts of the drug development could be performed by computer simulations rather than by costly experiments, because essentially all drugs evoke their action by binding to a target macromolecule. Likewise, many interesting questions in biochemistry can be formulated as the differential binding affinities of a substrate, product, or transition state to a protein or enzyme.

Consequently, numerous theoretical methods have been developed to estimate ligand affinities.[1] The most accurate ones are based on free-energy perturbation (FEP) and related approaches.[2] Unfortunately, they are extremely time-consuming, and the results typically converge only when the difference in binding affinity of similar ligands is considered, i.e., for relative binding affinities. Therefore, many more approximate methods have been suggested. Some of them are still based on extensive sampling of the phase space, e.g., linear-response approximation (LRA), the semimacroscopic protein-dipole Langevin-dipole approach (PDLD/S-LRA), the linear interaction energy (LIE), and molecular mechanics Poisson−Boltzmann surface area (MM/PBSA) approaches.[3−7] Other methods use a single molecular conformation and estimate the binding affinities by methods based on either physics or statistics.[1]

Most of the physical methods are based on calculations with a molecular mechanics (MM) force field. These force fields enable fast energy evaluations that allow extensive sampling. Moreover, they have the advantage of being

[*] Corresponding author e-mail:Ulf.Ryde@teokem.lu.se.

[†] Lund University.

[‡] University of Southern Denmark.

10.1021/ct9006986 © 2010 American Chemical Society
Published on Web 04/15/2010

tunable for specific systems and allowing contributions from the surrounding solvent to be included in an effective way through parametrization. Nevertheless, although the accuracy of ligand-affinity calculations is often limited by the extent of phase-space sampling, it can also be limited by the accuracy of the underlying force field. The deviation from experimental results seen even in well-converged free-energy calculations has often been attributed to imperfect force fields.[8] In fact, the results obtained with various MM force fields may differ strongly; for example, for biotin binding to avidin, there was a 91 kJ/mol difference in the interaction energy calculated with the Amber 1994 and 2002 force fields.[9] Likewise, FEP estimates of the binding affinities of five ligands to serine proteases differed by up to 36 kJ/mol between the MMFF and QMPFF force fields.[10] From a more conceptual point of view, it is also valuable to use an energy function that is less empirical, so that the results depend less on error cancellation.

Therefore, there has recently been great interest in developing ligand-binding methods that are based on quantum mechanics (QM), rather than on a MM force field.[11] Such methods are typically based on either semiempirical calculations[12,13] or on higher-level methods, using fractionation approaches, e.g., the fragment molecular orbital method (FMO)[14,15] or the molecular fractionation with conjugate caps (MFCC) and related methods.[16−22] However, it is well-known that calculations of dispersion effects generally require a very high level of theory.[23] Likewise, accurate predictions of polarization and dispersion effects require the use of a large and flexible one-electron basis set.[23,24] Only one of these previous studies[22] has been performed at a level (MP2/6-311(+)G(2d,p)) for which there is hope that dispersion and polarization effects are treated in a balanced and satisfactory way.

Recently, we have developed an approach that is intended to provide accurate interaction energies between a ligand and a macromolecule at a proper level of theory.[9] It is called PMISP (polarizable multipole interactions with supermolecular pairs). It treats electrostatic interactions by multipoles up to quadrupoles, induction by anisotropic polarizabilities, and nonclassical interactions (dispersion, exchange repulsion, etc.) by explicit quantum mechanical calculations, using a fragmentation approach similar to MFCC. It has given an accuracy of 2−5 kJ/mol for neutral and ∼10 kJ/mol for charged ligands compared to a full QM treatment.[9] This error could be reduced to 5 and 3 kJ/mol if the Hartree−Fock (HF) calculation for the full system is possible. For calculations with a whole protein, much computer time can be saved if long-range interactions are treated by a QM/MM approach (PMISP/MM).[25] If the boundary between the PMISP and MM systems is chosen far enough from the ligand, this approximation does not add any additional uncertainty. By this approach, we have illustrated the importance of using a proper level of theory. For example, the difference in interaction energy between two biotin analogues binding to avidin varied by 108 kJ/mol depending on the basis set employed (6-31G* or aug-cc-pVTZ at the MP2 level).[25]

However, in order to provide reliable ligand-binding energies, more terms than the pure interaction energy need to be considered. In particular, the effects of the surrounding solvent, entropy, and sampling need to be taken into account.[1,7] Only a few of the previous attempts to calculate ligand-binding energies with pure QM methods[12−22] have taken into account effects of solvation[12,13,15] (by a self-consistent reaction field Poisson−Boltzmann model and a surface area model) and entropy[12,13] (by counting the number of rotable bonds that are fixed during binding), and none of them consider sampling.

In this paper, we present what seems to be the first realistic QM estimation of ligand-binding affinities at a proper level of theory and at the same time taking into account the combined effects of solvation, entropy, and sampling. We employ the PMISP/MM method at the MP2/cc-pVTZ level within the framework of the MM/PBSA approach. We study the affinities of seven biotin analogues to the full avidin tetramer. This system is well characterized by X-ray crystallography,[26−29] and experimental binding free energies for a number of ligands (biotin analogues) are available.[30−32] Moreover, it has been investigated using several different theoretical methods.[33−39]

Methods

The PMISP/MM Method. The PMISP and PMISP/MM approaches have previously been thoroughly described.[9,25] Therefore, we here only provide a short summary of the methods. We consider the binding of a ligand (L) to a protein (P):

$$\mathrm{P} + \mathrm{L} \rightarrow \mathrm{PL} \tag{1}$$

In the PMISP method,[9] the interaction energy is estimated by

$$E_{\mathrm{PMISP}}(\mathrm{PL}) = E_{\mathrm{es}}(\mathrm{PL}) + E_{\mathrm{ind}}(\mathrm{PL}) + E_{\mathrm{nc}}(\mathrm{PL}) \tag{2}$$

where E_{es} and E_{ind} are the electrostatic and induction interaction energies, respectively. All energies in eqs 2−4 are interaction energies between L and P, not the absolute energies of the PL complex. The term E_{es} is calculated from a multicenter−multipole expansion up to quadrupoles, centered at all atoms and bond midpoints in the protein and the ligand. Likewise, E_{ind} is calculated from anisotropic dipole polarizabilities in the same centers in a self-consistent manner. Both these terms are obtained with the LoProp approach.[40] E_{nc} is the nonclassical term, containing mainly dispersion and exchange repulsion but also short-range corrections to the classical terms, e.g., charge penetration. It is estimated by

$$E_{\mathrm{nc}}(\mathrm{PL}) = \sum_{i=1}^{n} c_i(E_{\mathrm{QM}}(\mathrm{P}_i\mathrm{L}) - E_{\mathrm{es}}(\mathrm{P}_i\mathrm{L}) - E_{\mathrm{ind}}(\mathrm{P}_i\mathrm{L})) \tag{3}$$

where the protein has been divided into a number of fragments (P_i), using the molecular fractionation with conjugate caps (MFCC) method.[41] In this paper, each amino acid constitutes one fragment, and they are capped with CH_3CO- and $-NHCH_3$ groups. The caps from neighboring fragments are joined to form a $CH_3CONHCH_3$ conjugated cap (concap) for each peptide bond, and the energies of these concaps are subtracted ($c_i = -1$ in eq 3) from the energies of the capped amino acid fragments ($c_i = 1$). This has been shown to be an excellent approximation, giving errors of only ~1 kJ/mol.[9] $E_{QM}(P_iL)$ is the counterpoise-corrected quantum mechanical (QM) interaction energy of the P_i−L pair. A similar formula is used to derive properties (multipoles and polarizabilities) for the whole protein from fragment-wise calculations.[9] E_{QM} was calculated at the MP2/cc-pVTZ level, which has been shown to provide dispersion energies similar to coupled-cluster methods with larger basis sets, owing to error cancellation.[25,42] The multipoles and polarizabilities were calculated at the B3LYP/6-31G* level, which has been shown to be a good approximation for the much more expensive MP2/cc-pVTZ properties, provided that the same properties are used in both eqs 2 and 3.[25]

For a large protein, only a few fragments P_i are in close contact with the ligand, so the direct use of eq 2 would be very inefficient. Therefore, we can save much time without compromising the accuracy by using a QM/MM approach, PMISP/MM:[25] For a model containing residues close to the ligand (M), the full PMISP approach is used, whereas for more distant residues, E_{nc} is approximated by the Lennard-Jones term from a classical force field, E_{LJ}:

$$E_{PMISP/MM}(PL) = E_{es}(PL) + E_{ind}(PL) + E_{nc}(ML) + E_{LJ}(PL) - E_{LJ}(ML) \quad (4)$$

Thus, we use the same accurate multipole-polarizability model for the whole protein. In this work, the E_{LJ} term is taken from the Amber 1994 force field (the same terms are also used in the newer Amber 2003 and the polarizable 2002 force fields).[43−45] Naturally, the accuracy of this approximation will improve as the size of the M region is increased.[25] In this work, we have used all atoms that are within 4 Å of the ligand in at least one snapshot and added enough atoms to obtain chemically reasonable groups, such as aromatic rings or amide groups. For groups that form exceptionally strong interactions with the ligand (distances shorter than 1.7 Å), the model was extended with an extra CH_2 group, to avoid the largest errors observed previously[25] (e.g., Ser-73 was modeled by ethanol, rather than methanol). Thus, M consisted of 165−271 atoms, depending on the ligand (but the same M region was used for all snapshots with the same ligand).

It was previously shown that the PMISP error is rather insensitive to the quantum-chemical method and basis set employed and thus that one can exploit error cancellation.[9] Therefore, we also performed PMISP and full supermolecular calculations for the M region of each snapshot at a lower level of theory, HF/6-31G*, and subtracted the resulting deviation from the $E_{nc}(ML)$ term. The average correction was 6 kJ/mol, i.e., similar to the errors observed before for the same systems but with snapshots taken from a simulation with another force field.[9] By this procedure, the estimated error compared to full MP2/cc-pVTZ calculations is reduced to 3 kJ/mol for the M region,[9] whereas the protein environment adds an uncertainty of 5−8 kJ/mol.[25] All PMISP calculations were performed with the Molcas 7.2 software,[46] applying the Cholesky decomposition approximation to the two-electron integrals[47,48] in combination with the local-exchange algorithm.[49] We confirmed that the decomposition threshold used (10^{-4}) gave less than 1 kJ/mol error in the interaction energies.

PMISP/MM differs in several respects from standard QM/MM methods.[50] First, it uses a polarizable MM force field, which has been used in some previous studies[50,51] but is not routinely used. Second, a more advanced MM potential is used for the electrostatic interactions, including multipoles up to quadrupoles. Third, and most importantly, both the polarizabilities and all the multipoles are determined for each conformation of the protein by residue-wise QM calculations of the whole protein, ensuring that the conformational dependence of the polarizabilities and multipoles is explicitly accounted for. This conformational dependence has been shown to be significant, leading to errors of 3−43 kJ/mol for the electrostatic interaction energy between ligands and a protein or water solution.[52−55] Fourth, a large QM system is employed, 165−271 atoms, which ensures that the most important short-range interactions between the ligand and the protein are explicitly treated by QM, e.g., exchange, dispersion, charge transfer, charge penetration, as well as cross-terms and coupling to electrostatics and polarization. Fifth, a higher level of QM theory is employed than normally is used, MP2/cc-pVTZ. On the other hand, no geometry optimization at the PMISP/MM level is performed, and a fragmentation scheme is used to make the QM calculations feasible.

Solvation Calculations with the PCM Method. To accurately estimate ligand-binding affinities, an accurate estimate of the change in solvation energy upon ligand binding is needed. The standard continuum solvation methods for MM/PBSA in the AMBER software,[56] the Poisson−Boltzmann or generalized Born methods, cannot handle a multipole expansion or polarizabilities. Therefore, we instead decided to use the PCM method, which has recently been extended to be used with the effective fragment potential method (which also uses a polarizable force field with a multipole expansion).[57] We used the integral-equation formulation of PCM, IEFPCM,[58] which exhibits a better numerical stability than other formulations of PCM, and it is the default PCM method in the Gaussian software.[59] Owing to the large size of the molecular systems, the PCM-induced charges were obtained using a direct inversion of the iterative subspace procedure,[60] as implemented in the GAMESS software.[61] Thus, no explicit matrix inversion is needed. The PCM calculations were performed at the MM level, using the same multipoles and polarizabilities as in the PMISP calculations.

Like all continuum-solvation approaches, PCM employs dielectric cavities defined by a set of atomic radii. For accurate predictions of solvation energies, it is mandatory to use optimized cavity parameters. Several such sets of parameters are available for PCM at various levels of theory, e.g., Hartree−Fock[62] and density functional theory (UAHF and UAKS, i.e., united-atom topological model for Hartree−Fock and Kohn−Sham theory). Since we base our predictions on B3LYP and MP2 calculations, we decided to use the latter radii, which were optimized using the PBE0 functional.

Table 1. Calibration of the PCM Method for PMISP[a]

	scaling factor of radii for polar term												
	1.20	1.19	1.18	1.17	1.16	1.15	1.14	1.13	1.12	1.11	1.10	ΔG_{np}	exp.
H_2O	7.9	7.1	6.1	5.2	4.1	3.1	1.9	0.7	−0.5	−1.9	−3.3	6.3	−26.4[b]
CH_3OH	9.8	9.1	8.3	7.5	6.6	5.7	4.7	3.6	2.5	1.3	0.0	7.4	−21.4[b]
ethanol	10.7	9.9	9.1	8.3	7.3	6.4	5.4	4.2	3.0	1.7	0.4	8.2	−21.0[b]
p-$CH_3C_6H_4OH$	9.9	8.8	7.7	6.5	5.2	3.9	2.4	0.9	−0.8	−2.6	−4.6	10.0	−25.7[b]
NH_3	8.4	7.9	7.3	6.7	6.1	5.4	4.7	4.0	3.2	2.3	1.5	6.5	−17.9[b]
CH_3NH_2	11.7	11.1	10.5	9.8	9.1	8.3	7.5	6.6	5.7	4.7	3.6	7.5	−19.1[b]
CH_3CONH_2	13.7	12.6	11.5	10.3	9.0	7.6	6.2	4.7	3.2	1.6	−0.2	8.3	−40.6[b]
propionamide	15.2	14.2	13.1	11.9	10.7	9.4	8.0	6.5	5.0	3.4	1.7	8.9	−39.2[b]
CH_4	−1.4	−1.4	−1.4	−1.4	−1.4	−1.4	−1.5	−1.5	−1.5	−1.6	−1.6	7.1	8.4[b]
propane	0.1	0.1	0.1	0.0	0.0	0.0	0.0	−0.1	−0.1	−0.2	−0.2	8.5	8.2[b]
n-butane	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.1	0.1	0.0	9.3	8.7[b]
isobutane	−0.9	−0.9	−1.0	−1.0	−1.0	−1.1	−1.1	−1.1	−1.2	−1.2	−1.3	9.1	9.7[b]
toluene	5.9	5.5	5.2	4.8	4.4	4.0	3.5	3.0	2.5	1.9	1.3	9.6	−3.7[b]
CH_3SH	5.1	4.8	4.5	4.1	3.8	3.4	3.0	2.6	2.1	1.6	1.1	7.9	−5.2[b]
$CH_3SC_2H_5$	7.3	7.0	6.7	6.3	6.0	5.6	5.2	4.7	4.2	3.7	3.2	9.3	−6.2[c]
3-methylindol	8.2	7.1	6.0	4.8	3.6	2.3	0.9	−0.6	−2.2	−3.9	−5.7	10.7	−24.6[c]
4-methylimidazole	12.4	10.9	9.4	7.8	6.2	4.4	2.5	0.5	−1.6	−3.8	−6.3	9.1	−43.0[c]
N-propyl guanidine	17.3	15.5	13.7	11.7	9.6	7.4	5.0	2.5	−0.3	−3.3	−6.4	10.2	−45.7[d]
$CH_3NH_3^+$	36.8	34.2	31.6	28.9	26.1	23.3	20.4	17.4	14.4	11.3	8.1	7.8	−319.7[e]
imidazoleH$^+$	−9.2	−12.0	−14.8	−17.8	−20.9	−24.0	−27.3	−30.8	−34.3	−38.1	−42.0	8.5	−248.9[f]
$HCOO^-$	15.3	12.6	9.9	7.2	4.3	1.4	−1.6	−4.7	−7.9	−11.2	−14.6	7.2	−318.8[e]
CH_3COO^-	21.6	19.0	16.3	13.5	10.7	7.8	4.9	2.0	−1.0	−4.1	−7.3	8.1	−324.7[e]
MAD, all	10.4	9.6	8.8	8.0	7.1	6.2	5.4	4.7	4.4	4.8	5.2		
MAD, neutral	8.1	7.5	6.8	6.0	5.2	4.4	3.5	2.7	2.2	2.3	2.3		

[a] The total solvation free energies of 22 organic molecules and ions were calculated with the PCM+SASA method, using different values for the scaling factor of the radii for the electrostatic term (1.10−1.20). The SASA nonpolar energy, calculated with Parse radii,[67] was added to these values, and the results were compared to experiments. In the table, the differences compared to experiments are given, as well as the nonpolar energy term (ΔG_{np}) and the experimental data (exp.)[64−68] (all in kJ/mol). [b] Data from ref 64. [c] Data from ref 66. [d] Data from ref 67. [e] Data from ref 65. [f] 18.8 kJ/mol was added to the value in ref 68 to use the same value of the absolute solvation energy of a proton as in ref 65.

These radii, although not yet properly published, are available in the Gaussian 03 suite of programs.[63]

Since we use the UAKS parameters at the MM level, recalibration of these parameters is strictly needed. However, we limited the recalibration to a scaling of the radii for the electrostatic component in the PCM solvation energy calculation. For the original UAKS radii, this scaling parameter is 1.2. The calibration was based on a test-set of 22 small organic molecules, listed in Table 1. These molecules were selected to represent models of the peptide backbone and all amino acid side chains. For these, we constructed distributed multipoles up to quadrupoles and anisotropic polarizabilities in the same way as for PMISP.[9] The multipoles and polarizabilities were calculated using the B3LYP/6-31G* method (6-31+G* for the two anions), and the solvation energies were then evaluated using the PCM approach for various values of the scaling parameter. The nonpolar solvation terms (cavitation, dispersion, and exchange repulsion[62]) are independent of this scaling factor and were therefore calculated only once. As will be discussed below, we encountered serious problems with the nonpolar terms in the PCM model. Therefore, the final calibration of the PCM method (Table 1) employed instead the nonpolar energy from the standard MM/PBSA method. A fitting to experimental data[64−68] gave a scaling factor of 1.12 (with the nonpolar terms from PCM, the optimum scaling factor was 1.15). This decrease in the scaling factor is expected, because there is no charge penetration at the MM level. The scaled model gave MADs of 2 and 4 kJ/mol, for the neutral molecules and all molecules, respectively. This is only slightly worse than for the UAHF parameters (1 and 3 kJ/mol), similar to the UAKS parameters (1 and 5 kJ/mol), and appreciably better than seven different Poisson−Boltzmann and 11 generalized Born methods (3−9 and 7−18 kJ/mol).[69] For the seven biotin analogues, this recalibrated PCM method gives a MAD of 10 kJ/mol, compared to a weighted average of 24 different continuum solvation methods,[69] which again is slightly worse than for the original UAHF and UAKS methods (4 and 7 kJ/mol).

MM/PBSA. The calculations in this paper are based on the MM/PBSA approach.[7] We selected the MM/PBSA approach because it is widely used and has been shown to give reasonable results for many systems.[7,35,38,39,70−72] It also contains no adjustable parameters and has a modular approach with separate energy terms, which facilitates the incorporation of QM data. However, we do not claim that this approach is more accurate or effective than other approaches.

In this method, the binding affinity (the free energy of the reaction in eq 1, ΔG_{bind}) is estimated from the free energies of the three reactants,

$$\Delta G_{\mathrm{bind}} = G(\mathrm{PL}) - G(\mathrm{P}) - G(\mathrm{L}) \qquad (5)$$

where all species are assumed to be in water solution. The free energy of each of the reactants is estimated as a sum of four terms:

$$G = \langle E_{\mathrm{MM}}\rangle + \langle G_{\mathrm{solv}}\rangle + \langle G_{\mathrm{np}}\rangle - T\langle S_{\mathrm{MM}}\rangle \qquad (6)$$

where G_{solv} is the polar solvation energy of the molecule, estimated by the solution of the Poisson−Boltzmann (PB) equation,[73] G_{np} is the nonpolar solvation energy, estimated from the solvent-accessible surface area (SASA) of the molecule,[74] T is the temperature, S_{MM} is the entropy of the molecule, estimated from a normal-mode analysis of harmonic frequencies calculated at the molecular mechanics (MM) level, and E_{MM} is the MM energy of the molecule, i.e., the sum of the internal energy of the molecule (i.e., bonded terms, E_{bond}), the electrostatics (E_{es}), induction energy (E_{ind}, only if a polarizable force field is used), and van der Waals interactions (E_{vdW}):

$$E_{\mathrm{MM}} = E_{\mathrm{bond}} + E_{\mathrm{es}} + E_{\mathrm{ind}} + E_{\mathrm{vdW}} \tag{7}$$

All of the terms in eq 6 are averages of energies obtained from a number of snapshots taken from MD simulations. To reduce the time-consumption and increase the precision, the same geometry is normally used for all three reactants (complex, ligand, and receptor); i.e., only the PL complex is explicitly simulated by MD.[75] Thereby, E_{bond} cancels in the calculation of ΔG_{bind}.

In this investigation, we test if the binding-affinity predictions can be improved by replacing some of these terms with estimates using other methods. Thus, we replace the E_{MM} term by the PMISP/MM interaction energy between the ligand and the protein (eq 6). Second, we replace the $G_{\mathrm{solv}} + G_{\mathrm{np}}$ estimates of the solvation energies by the corresponding terms within the PCM model. The E_{ind} and G_{solv} terms are computed in a self-consistent way, as described in ref 57 and implemented in GAMESS.[61] Thus, the apparent surface charges and the induced dipoles are simultaneously iterated to self-consistency. Then, E_{ind} is defined as the energy of the induced dipoles in the electric field of the multipoles and G_{solv} as the energy of the apparent surface charges in the electric potential of the multipoles, with both terms divided by two to account for the self-energy of polarization. This decomposition is only used in the qualitative discussion; only the sum of these terms is well-defined and influences the result. Other approaches to replace the E_{MM} term with a standard QM/MM term have been tested, both for calculations of ligand-binding affinities and for other energies.[76−78] A possible problem, common to these methods and to PMISP/MM/PCM, is that the geometries are not generated by the same energy function as used to evaluate the binding affinities. However, it should be noted that a similar problem exists for the original MM/PBSA method when the geometries are generated using explicit solvent, but the energies are calculated with implicit solvent. Moreover, this problem is reduced by using multiple snapshots instead of a single minimized structure.

Thus, only the S_{MM} term is kept from the original MM/PBSA method, but it is calculated according to our recently developed method to improve the precision of this estimate.[79] In the original approach,[7] the protein was truncated 8 Å from the ligand, and it was then freely optimized, using a distance-dependent dielectric constant $\varepsilon = 4r$. We have shown that this gives a large statistical uncertainty in the entropy estimate, which can be reduced by a factor of 2−4 if a buffer region of 4 Å is used outside the cutoff radius. This buffer region is kept fixed in the geometry optimization and is not included in the estimate of the entropy, but it ensures that the optimized system stays close to the structure in the complex. This also makes the use of the questionable distance-dependent dielectric constant superfluous.

The results of the PMISP/MM/PCM/$T\Delta S$ approach are compared to the results of standard MM/PBSA calculations using the polarizable Amber 2002 force field.[45] These were performed in the same way as in our previous investigation of various force fields for the biotin−avidin system[38] (the 02ohp/02 calculations in that work, although the present calculations are based on the ligand in the fourth subunit in the tetramer, rather than the first one in the previous investigation). This means that G_{solv} is estimated by adding an extra charge close to each atom site to simulate the induced dipoles in the PB calculations. Calculations of both G_{solv} and G_{np} used Parse radii.[67] The calculations were performed with the Amber software,[56] but with the improved entropy estimate[79] (this term is identical to the one used in the PMISP/MM/PCM/$T\Delta S$ approach). Unfortunately, the Amber nmode program does not work properly for a polarizable force field, so the entropy calculations were performed without the polarizabilities.

AutoDock Calculations. Standard docking calculations were performed with AutoDock 4.[80] Three sets of calculations were performed. In the first set, we simply rescored the same snapshots used in the MM/PBSA calculations with the AutoDock scoring function[81] and averaged the results. In the second set, we docked the ligand into the equilibrated protein structure for each ligand, as represented by the first snapshot from the MD simulations. Finally, in the third set, we docked the ligand into the crystal structure. In addition, we tested the influence of the partial charges by performing all calculations with either the default (Gasteiger) charges or the Amber charges used in the MD simulations. Default settings (e.g., atom types) were used. The protein was considered rigid in all docking calculations, whereas the ligands were fully flexible.

Studied Systems. We studied the binding of the seven biotin analogues (BTN1−BTN7) in Figure 1 to avidin. The setup of the molecular dynamics simulations has been described before.[38] We used 10 snapshots (sampled every 20 ps) for each analogue taken from this investigation, performed by the polarizable Amber 2002 force field[45] (the 02ohp simulation in ref 38). Error estimates of the correlation coefficients and the mean absolute deviations were obtained by 10 000 random simulations, as has been described before.[39] It is likely that the reported standard deviations are underestimates because the structures come from a single simulation, rather than several independent simulations.[39]

Timings. A single-point calculation with the PMISP/MM/PCM/$T\Delta S$ method took 38−45 CPU days. Most of this time was spent on the property calculations (∼30 CPU days), which could be significantly sped up by using software optimized for density functional theory. The supermolecular calculations took 5−12 CPU days depending on the ligand size and the PCM calculations took 3 CPU days, whereas the computational time for all other steps was negligible. All calculations were run trivially in parallel. The corresponding time for a single-point MM/PBSA evaluation is less than 1 CPU hour. However, in that case, the computa-

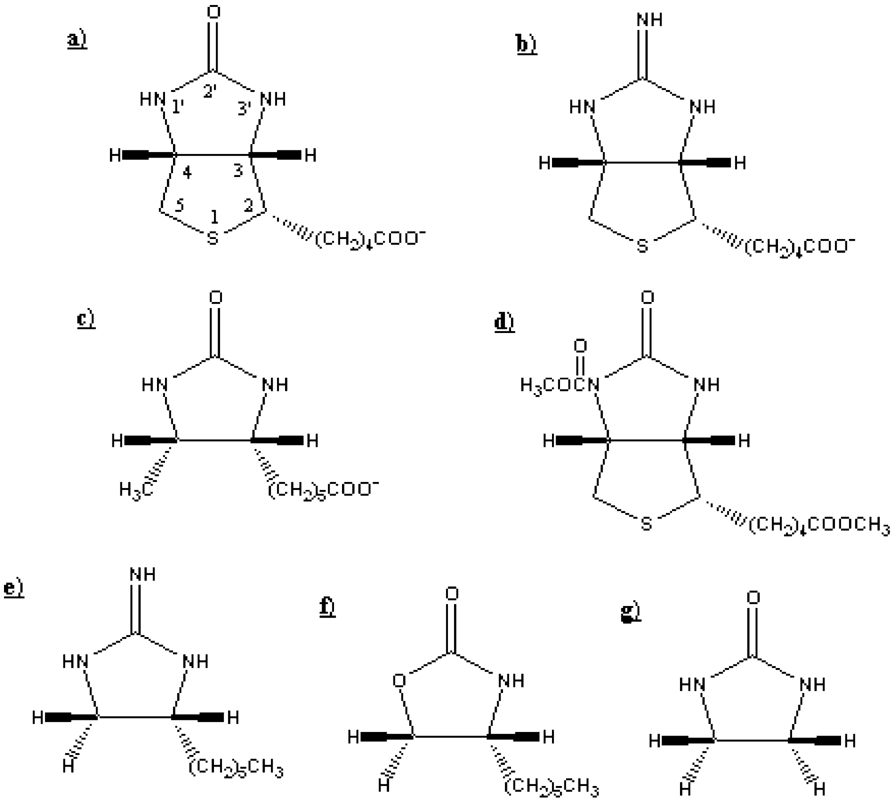


Figure 1. The seven biotin analogues used in this study. (a) BTN1 (biotin), (b−g) BTN2−BTN7.

Table 2. Average Nonpolar Solvation Energies (kJ/mol) in the SASA and PCM Calculations[a]

	SASA				PCM						
	PL	P	L	PL−P−L	PL	P	L	PL−P−L	cav	disp	rep
BTN1	464.5	466.9	14.1	−16.5	22282.5	22156.3	12.3	113.9	−18.5	189.5	−57.2
BTN2	466.1	468.8	14.1	−16.8	22274.8	22144.1	14.4	116.3	−19.8	197.6	−61.4
BTN3	471.5	474.1	14.1	−16.7	22234.1	22116.7	6.7	110.7	−21.6	180.8	−48.5
BTN4	468.7	472.9	16.3	−20.5	22447.5	22267.7	24.4	155.4	−22.4	236.2	−58.5
BTN5	465.5	468.1	13.4	−16.0	22250.3	22137.8	8.0	104.5	−18.2	163.5	−40.9
BTN6	466.7	469.6	13.1	−16.0	22228.3	22119.1	8.9	100.5	−19.1	153.4	−33.9
BTN7	478.4	479.8	9.3	−10.6	21941.9	21892.2	−2.5	52.3	−7.5	89.1	−29.3

[a] The energy contributions for the complex (PL), protein (P), and ligand (L) given, as well as the net contribution to the binding (PL−P−L), for PCM further divided into cavitation (cav), dispersion (disp), and repulsion (rep) contributions.

tional time is dominated by the generation of snapshots, which took 7 CPU days per ligand.[38]

Result and Discussion

Nonpolar Solvation Energy. First, we calculated the solvation energies using the full PCM model implemented in the GAMESS program.[61] However, this gave differential solvation energies (i.e., $G_{\text{solv}}(\text{PL}) + G_{\text{np}}(\text{PL}) - G_{\text{solv}}(\text{P}) - G_{\text{np}}(\text{P}) - G_{\text{solv}}(\text{L}) - G_{\text{np}}(\text{L})$) that were 60−180 kJ/mol more positive than the corresponding results with a PB+SASA model. Further inspection showed that these differences arise almost entirely from the nonpolar part of the solvation energy: In the PB+SASA method, this term is taken from the difference in the solvent-exposed surface area between the complex and the isolated protein and ligand. From the results presented in Table 2, it can be seen that the SASA nonpolar energies are quite small and similar for the complex and the protein, ~470 kJ/mol (corresponding to a SASA of 20 600 Å^2, because $\Delta G_{\text{np}} = \text{SASA} \times 0.0227 - 3.85$ kJ/mol, when SASA is given in Å^2 [7,35,38]). The difference is 1−4 kJ/mol, with the protein having the largest value, indicating that the ligand is mainly buried in the protein. Therefore, the net nonpolar SASA effect comes mostly from the ligand. As an effect, ΔG_{np} in MM/PBSA is small and negative for all complexes, 11−21 kJ/mol, and directly related to the size of the ligand.

However, in the PCM method, the nonpolar solvation energy is calculated from three separate terms: the energy cost of making a cavity in the solvent (the cavitation energy), a favorable term from the dispersion interactions between the solute and the solvent, and the corresponding unfavorable term from exchange repulsion.[62] The former term is calculated from an expression that contains terms involving the radius of each atom to the power of 0, 1, 2, and 3,[82−85] i.e., including a term that is proportional to the volume, whereas the latter two terms are calculated by a surface-based integration method.[86] From Table 2, it can be seen that the PCM energies of the protein and the complex are almost 50 times larger than the SASA energies, ~22 200 kJ/mol. The PCM energies are dominated by the cavitation energy, which is ~28 000 kJ/mol, compared to the dispersion energy of ~−7500 kJ/mol and the exchange repulsion energy of ~2000 kJ/mol. However, when computing the difference upon binding, the cavitation energy is mainly canceled (the net effect is negative and 8−22 kJ/mol; cf. Table 2). This indicates that the volume term of the cavitation energy is dominating the individual energies, because the volume hardly changes during ligand binding. On the other hand, the surface area is reduced, and this causes a positive (unfavorable) contribution from the dispersion term of 89−236 kJ/mol, only partly canceled by the exchange repulsion (negative and 29−61 kJ/mol) and by the small cavitation energy. Therefore, the net ΔG_{np} is 52−155 kJ/mol; i.e., it has the opposite sign and is larger in magnitude compared to the SASA nonpolar solvation energy. It is notable that the two methods are reasonably in accord for the ligand: The SASA energy estimate is 9−16 kJ/mol (corresponding to SASAs of 240−550 Å^2), whereas the PCM nonpolar energies are −3 to +24 kJ/mol with a correlation coefficient $r^2 = 0.85$.

This illustrates a major problem in estimating binding affinities using approaches that involve a continuum estimate of the solvation energy. Apparently, there is no consensus as to how the nonpolar energy should be estimated, and the PCM and SASA approaches give strongly differing results. It has previously been argued that it does not matter whether the area or volume is used to estimate the nonpolar solvation energy.[87] However, the present results show that this is not the case for ligand-binding affinities: When a ligand binds to a complex, the volume of the protein increases approximately by the volume of the ligand (so that the total volume during the binding reaction hardly changes). However, the SASA typically decreases during the binding, because the ligand becomes partly hidden by the protein and an empty cavity in the protein becomes filled by the ligand. In PCM, this is further complicated by the use of several energy terms with different functional forms. In fact, it appears that the cavity term (after cancellation of the volume contributions) contains the same information as the SASA estimate (the difference is always within 5 kJ/mol). The additional terms used in PCM (dispersion and repulsion) do not appear to give any advantage for protein−ligand energies; on the contrary, the results deteriorate. It should be noted that these terms are well motivated from a physical point of view and, for example, that the three-dimensional reference interaction site model (3D-RISM[88,89]) gives nonpolar energies of the same size and sign as PCM.[90] Therefore, this seems to be a parametrization problem, possibly related to the general difficulty of using terms with different signs in a fitting expression. However, we cannot exclude the possibility that the better result for the SASA estimate is fortuitous. We currently investigate this issue for other systems.

Another difference between the two solvation methods is that the PB method is based on the SASA, whereas the cavitation terms in PCM are based on the van der Waals surface of the solute (and the other terms on the solvent-excluded surface area, SESA).[91] The van der Waals surface is simply the surface of the union of spheres on all atoms with the corresponding van der Waals radius, whereas the SASA is the surface defined by the center of a spherical solvent probe that is rolled on the van der Waals surface. Therefore, the radius of a solvent molecule (~1.4 Å for water) is added to the van der Waals radii of each atom, and crevices between the spheres that are not accessible to a solvent probe are considered as a part of the solute. The SESA is similar to the van der Waals surface, but it excludes those crevices. For small molecules, for which the PCM method was calibrated,[62] these three surfaces are rather similar. However, for a large molecule, like a protein, they are totally different, because there are numerous small cavities inside the protein that are not large enough to house a solvent probe. The solvent-accessible surface of the protein will essentially be only the outer surface of the protein, whereas the van der Waals surface will be much larger. For example, for the avidin tetramer, the van der Waals area is 58 000 Å^2, and all atoms contribute to it, whereas the SASA is only 21 000 Å^2, and only 40% of the atoms contribute to it. The SESA is ~20 000 Å^3. It should be noted that this effect becomes apparent already for much smaller molecules. For example, in our study of solvation energies of drug-like molecules, the PCM estimates differed by 100−150 kJ/mol from that of all other methods for the two largest molecules (with 68 and 113 atoms).[69]

It seems quite questionable to use the van der Waals surface to calculate the solvation energy of a protein. Therefore, we tend to prefer the SASA model, which also gives 50 times smaller energies and thus probably more precise differences. We have therefore based the recalibrated PCM model on the nonpolar SASA energies. We do not argue that this is an optimum approach—on the contrary, it would be better to develop a new PCM method that works properly also for a protein, based on the SASA or SESA. Unfortunately, this is a major task, involving both method development and a complete reparametrization of the method so that it works well both for small molecules and for proteins. Moreover, it has to be settled whether the nonpolar term should be based on the volume or the surface area. This is out of the scope of the present investigation.

Binding Affinity Estimates. Table 3 shows the various terms in the full PMISP/MM/PCM/$T\Delta S$ method (with the nonpolar solvation energies from the PCM method; column ΔG_1). It can be seen that the method gives poor absolute affinities, ranging from +30 to +103 kJ/mol, compared to the experimental data, −19 to −85 kJ/mol (Figure 2),[30−32] with a mean absolute deviation from the experimental values (MAD) of 108 kJ/mol. If we allow for a systematic error in the method (i.e., if we translate all points with the mean signed error), we still get a mean absolute deviation (TR MAD) of 19 ± 3 kJ/mol, with the largest error for BTN1. This result is disappointing. It is worse than similar MM/PBSA calculations using various MM force fields for the same system, which gave MADs of 9−19 kJ/mol, and TR MADs of 5−19 kJ/mol.[38] In particular, the standard MM/PBSA calculations for exactly the same snapshots, using the Amber 2002 force field, give a MAD and TR MAD of 13

Table 3. Results of the PMISP/MM/PCM/$T\Delta S$ Method (kJ/mol)[a]

exp	ΔE_{es}	ΔE_{ind}	ΔE_{nc}	$\Delta G_{solv,PCM}$	$\Delta G_{np,PCM}$	$-T\Delta S$	$\Delta G_{np,SASA}$	ΔE_{vdW}	ΔE_{coop}	ΔG_1	ΔG_2	ΔG_3
BTN1	−1061.2	−253.5	−75.0	1236.5	113.9	96.8	−16.9	−143.4	78.9	57.6	−73.1	−141.6
BTN2	−1109.5	−322.7	−67.5	1311.3	116.3	102.4	−17.2	−149.1	94.3	30.4	−103.2	−184.8
BTN3	−1055.0	−282.3	−61.9	1235.0	110.7	93.4	−16.7	−138.6	95.7	39.9	−87.5	−164.1
BTN4	−123.1	−55.9	−151.1	181.1	155.4	96.5	−21.3	−211.0	−0.5	103.0	−73.6	−133.5
BTN5	−112.6	−50.9	−92.4	140.7	104.5	77.4	−16.2	−134.5	−10.3	66.7	−54.0	−96.1
BTN6	−88.5	−45.3	−91.0	132.9	100.4	69.8	−15.8	−131.9	−5.3	78.4	−37.8	−78.8
BTN7	−112.5	−46.3	−22.6	127.4	52.3	66.4	−10.7	−53.6	−14.1	64.6	1.7	−29.3
MAD										107.9	25.5	73.4
TR MAD										19.2	18.6	30.6
R^2										0.27	0.52	0.59

[a] Three different estimates of the total binding energy are given: $\Delta G_1 = \Delta E_{es} + \Delta E_{ind} + \Delta E_{nc} + \Delta G_{solv,PCM} + \Delta G_{np,PCM} - T\Delta S$ is the full PMISP/MM/PCM/TΔS, whereas in $\Delta G_2 = \Delta E_{es} + \Delta E_{ind} + \Delta E_{nc} + \Delta G_{solv,PCM} + \Delta G_{np,SASA} - T\Delta S$, the nonpolar PCM term has been replaced by the nonpolar SASA term, and in $\Delta G_3 = \Delta E_{es} + \Delta E_{ind} + \Delta E_{vdW} + \Delta G_{solv,PCM} + \Delta G_{np,SASA} - T\Delta S$, the ΔE_{nc} term has also been replaced by the Amber van der Waals energy. The mean absolute deviation (MAD), the correlation coefficient (R^2), as well as the MAD after subtraction of the mean signed deviation (TR MAD) are also given for each energy estimate. ΔE_{coop} is the cooperativity of the binding in a vacuum, defined as the difference between the induction energy of the whole protein−ligand complex and the sum of pairwise induction energies for the fragment−ligand dimers.

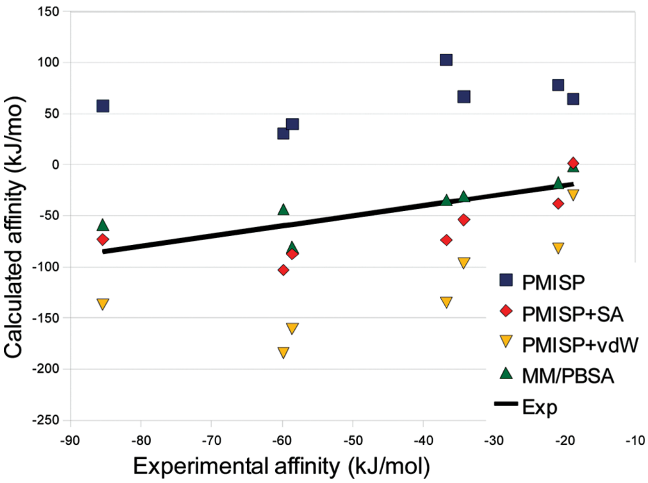


Figure 2. The results of the PMISP/MM/PCM/$T\Delta S$, PMISP/MM/PCM/SASA/$T\Delta S$, PMISP/MM/PCM/SASA/$T\Delta S$/E_{vdW}, and MM/PBSA (with the Amber-02 force field) methods for the binding of seven biotin analogues to avidin.

and 11 ± 3 kJ/mol (Table 4), respectively. In fact, the result is not significantly better than assigning the same affinity to all seven biotin analogues, which gives a TR MAD of 20 kJ/mol. The correlation coefficient is also poor, $R^2 = 0.27 \pm 0.10$, compared to 0.65 ± 0.09 for MM/PBSA and 0.43−0.98 in our previous investigation.[38]

The replacement of the PCM nonpolar term by the SASA term (as discussed above) gave the same TR MAD, although the binding affinities are shifted to a range that is closer to the experimental one, +2 to −103 kJ/mol (column ΔG_2 in Table 3), and the correlation coefficient is improved ($R^2 = 0.52 \pm 0.09$). If the nonpolar solvation energy is omitted, the result is improved in absolute terms, so that both the MAD and TR MAD are 17 ± 2 kJ/mol, but R^2 remains similar, 0.55 ± 0.10.

The standard deviations of the various PMISP/MM/PCM/$T\Delta S_{MM}$ energy terms are listed in Table 5. It can be seen that they are 10−30 kJ/mol for the final binding energy. Thus, the standard errors of the mean values are 3−9 kJ/mol, showing that the statistical precision cannot explain the poor results. The standard deviation is dominated by the electrostatics, induction, polar solvation, and nonclassical terms, which typically give slightly larger standard deviations than the total energy, because some of the variation between these terms is canceled. The standard deviation of the entropy term is also quite large, 9−21 kJ/mol, but it never limits the precision of the method. The standard deviation of the nonpolar solvation energy is always less than 1 kJ/mol. The corresponding standard deviations for the MM/PBSA method are also listed in Table 5. The standard deviations of the electrostatics and entropy terms are similar to that for PMISP, but those of the solvation and the nonclassical terms are somewhat smaller.

To see how these results compare with other simpler methods, we tried to correlate the binding affinities to the molecular weight of the ligands or to the Amber van der Waals term alone. However, this gave poor correlations to the experimental data, $R^2 = 0.20$ and 0.11, respectively.

Next, we performed a docking study of the same ligands using AutoDock.[80] The results are shown in Table 6. First, we rescored the MD snapshots used in the MM/PBSA calculations with the AutoDock scoring function.[81] This gave good agreement with the experimental values for the neutral ligands, whereas the binding free energies of the charged ligands were too positive. Nevertheless, the MAD with Gasteiger charges was 16 kJ/mol, and the TR MAD was 13 kJ/mol. The Amber charges gave consistently less negative binding affinities and thus a larger MAD (30 kJ/mol), but the relative energies were not significantly affected (TR MAD 14 kJ/mol). Interestingly, the standard deviations over the snapshots, listed in Table 5, were significantly smaller than for the PMISP or MM/PBSA calculations, ranging from 1 to 4 kJ/mol for the various ligands.

Next, we docked the ligand into the equilibrated protein structure. This gave similar results (TR MADs of 13−14 kJ/mol), because the best docked binding pose agreed with the one used in the MD simulations (average root-mean-squared deviation in geometries of 0.8 Å) in all cases except one (BTN5 with Amber charges). On average, the binding free energy was 2 kJ/mol more favorable when the ligand was allowed to relax.

Finally, we docked the ligand into the crystal structure. Again, the best docked binding mode agreed with the one

Table 4. Results for the MM/PBSA Calculations Using the Polarizable Amber 2002 Force Field (kJ/mol)[a]

	ΔE_{es}	ΔE_{ind}	ΔE_{vdW}	$\Delta G_{solv,PB}$	$\Delta G_{np,SASA}$	$-T\Delta S$	ΔG_{bind}	exp.
BTN1	−1173.6	−1.6	−143.4	1180.0	−16.9	96.8	−58.6	−85.4
BTN2	−1213.8	14.6	−149.2	1220.0	−17.2	102.4	−43.2	−59.8
BTN3	−1182.2	−0.6	−138.6	1164.1	−16.7	93.4	−80.6	−58.6
BTN4	−127.0	−14.6	−211.0	243.1	−21.3	96.6	−34.2	−36.8
BTN5	−100.7	−16.3	−134.5	159.4	−16.1	77.4	−30.9	−34.3
BTN6	−80.4	−11.2	−131.9	152.1	−15.8	69.8	−17.4	−20.9
BTN7	−111.4	−9.6	−53.6	117.6	−10.6	66.4	−1.2	−18.8

[a] The MAD and TR MAD are 13.2 and 11.5 kJ/mol, respectively, and R^2 is 0.65.

Table 5. Standard Deviations of the Various Terms for PMISP/MM/PCM/$T\Delta S_{MM,}$ MM/PBSA, and AutoDock with Gasteiger (G) and Amber (A) Charges (kJ/mol)[a]

	MM/PBSA								PMISP/MM/PCM/$T\Delta S$							G	A
	E_{es}	E_{ind}	E_{vdW}	G_{PB}	G_{SASA}	$T\Delta S$	G_{bind}	E_{eis}	E_{es}	E_{ind}	E_{nc}	G_{PCM}	$G_{np,PCM}$	G_{bind}	E_{eis}	G_{bind}	G_{bind}
BTN1	20.9	5.4	15.0	18.8	0.2	14.2	18.7	19.6	17.6	17.6	25.8	22.0	0.2	23.4	18.4	1.4	1.4
BTN2	37.6	4.6	14.4	21.0	0.2	21.0	25.4	25.5	32.0	32.2	23.3	45.4	0.2	26.3	33.9	2.7	3.0
BTN3	26.0	6.8	11.6	17.9	0.2	11.7	22.1	29.3	25.1	24.2	31.0	24.8	0.2	14.7	29.6	3.4	3.2
BTN4	15.5	3.8	10.8	16.5	0.3	11.8	20.7	19.8	18.8	9.8	14.1	11.9	0.3	16.9	19.9	3.6	3.4
BTN5	18.9	3.9	8.0	16.7	0.1	12.1	21.8	23.6	13.3	13.3	13.9	15.8	0.1	9.8	14.6	1.2	1.4
BTN6	15.1	3.1	15.3	8.7	0.2	15.3	21.1	17.4	13.1	11.0	19.9	15.4	0.2	14.9	14.0	2.0	1.8
BTN7	10.6	3.1	7.8	10.9	0.1	9.3	18.5	16.6	11.6	6.9	13.2	8.4	0.1	9.9	12.5	1.1	1.2

[a] E_{eis} is the sum of the E_{es}, E_{ind}, and G_{solv} terms.

Table 6. Binding Free Energies (kJ/mol) Obtained from the AutoDock Calculations Based on Rescoring of the MD Snapshots, Docking into Snapshot 1, or Docking into the Crystal Structure, Using Gasteiger or Amber Charges

	rescoring		docking (snapshot)		docking (crystal)	
	Gasteiger	Amber	Gasteiger	Amber	Gasteiger	Amber
BTN1	−37	−24	−40	−26	−38	−27
BTN2	−37	−23	−41	−24	−36	−25
BTN3	−36	−19	−37	−21	−35	−20
BTN4	−35	−15	−37	−17	−33[a]	−11[a]
BTN5	−21	−7	−22	−7[a]	−21[b]	−19[b]
BTN6	−22	−10	−26	−13	−27[a]	−9[b]
BTN7	−14	−8	−15	−8	−13[b]	−9[a]
MAD	16	30	15	28	18	28
TR MAD	13	14	13	14	13	14
R^2	0.68	0.83	0.66	0.79	0.65	0.84
slope	0.32	0.27	0.33	0.28	0.31	0.29

[a] Simulated pose not found. [b] Simulated pose ranked 2nd or 3rd.

used in the MD simulations for all the charged ligands. For the other ligands, the simulated pose was sometimes ranked second and third (with differences of ∼1 kJ/mol), whereas in some cases, the simulated pose was not among the docked binding poses (see Table 6). This probably reflects the fact that the crystal structure was obtained using biotin (BTN1), and thus a rigid protein represents a crude approximation when the ligands differ significantly in size. Reassuringly, the predicted binding free energies from the crystal docking were always similar or slightly larger (up to 6 kJ/mol) than that found in the docking to the MD snapshots (with the same exception as before, BTN5 with Amber charges). This indicates that the simulated pose is the correct one for all ligands. The TR MADs from the crystal docking were again 13−14 kJ/mol.

The correlation coefficients for the AutoDock results (R^2) ranged from 0.65 to 0.83. However, the slopes of the correlation lines are small (0.24−0.33), indicating that the energy scale of the AutoDock scoring function is less quantitative than the methods based on PMISP (which give slopes of 0.5−2.2). Overall, the AutoDock results are similar or sometimes slightly better than those based on PMISP. An important reason for this is the smaller standard deviation of the AutoDock results. A more physical energy function will give energy terms that are larger in magnitude, but to a large extent canceling. Thus, it needs to have a much higher accuracy than a less detailed model to give a better final result. In fact, many energy functions can be improved by simply scaling down all terms. For example, if ΔG_2 in Table 3 (PMISP/MM with SASA) is scaled down by a factor 2, the MAD and TR MAD become 14 and 11 kJ/mol (whereas R^2 does not change from 0.52), respectively, but at the same time, the slope of the correlation line is reduced from 1 to 0.5.

We will try to rationalize the failure of PMISP by analyzing the various terms in the method in comparison to MM/PBSA. The entropy term is identical between the two methods, and the nonpolar solvation term is also identical in the ΔG_2 estimate, so these cannot explain the failure.

The polar solvation energies show differences of −62 to +91 kJ/mol (PMISP/MM/PCM is mostly more negative for the neutral ligands and always more positive for the charged ligands). However, the correlation is excellent for the neutral ligands, $R^2 = 0.97$, and rather good for the charged ones, $R^2 = 0.93$.

The electrostatic and induction energies of the PMISP and MM/PBSA methods are not comparable, because intramolecular induction is not treated in the same way.[9] Therefore, we can only compare the sum of these two terms. It turns out that this sum is always more negative with PMISP than with MM/PBSA, by 37−47 kJ/mol for the neutral ligands, and by 140−233 kJ/mol for the charged ligands. However, again, the two terms are almost perfectly correlated, with an R^2 of 0.98 and 0.96 for the charged and neutral ligands, respectively.

If the solvation energy is added to this sum, the difference is partially canceled, but the PMISP/MM/PCM results are still 28−142 kJ/mol more negative than the MM/PBSA results. Interestingly, the good correlation is completely lost, especially for the charged ligands ($R^2 = 0.11$, versus 0.84 for the neutral ligands).

Finally, the nonclassical (van der Waals) energies also differ by a sizable but rather constant amount, 31−82 kJ/mol, which is slightly larger for the charged ligands than for the neutral ones. The PMISP estimates are always more positive. There is a perfect correlation ($R^2 = 1.00$) between Amber and PMISP for the neutral ligands, but it is much worse for the charged ones ($R^2 = 0.14$). If we replace the nonclassical PMISP term with the Amber van der Waals term, the results become worse, with a TR MAD of 31 kJ/mol (but R^2 increases to 0.59; cf. Figure 2).

Recently, it has been shown that there are strong cooperative effects in the binding of biotin to avidin (∼45 kJ/mol at the MP2/6-31+G** level).[92] Such effects are included in the present calculations through the polarizabilities. However, from the results in Table 3 (column ΔE_{coop}), it can be seen that we actually find strong anticooperative effects for the three charged ligands (by 79−96 kJ/mol for the full protein and 24−31 kJ/mol for the region *M*), whereas we find cooperative effects for the four neutral ligands (0−14 kJ/mol). The reason for this discrepancy is most likely that the previous calculations omitted the carboxylate tail of biotin and therefore its negative charge.

Conclusions

In this paper, we present the first attempt to calculate ligand-binding affinities using high-level QM methods with large basis sets (MP2/cc-pVTZ, i.e., enough to get reasonably accurate dispersion energies), combined with estimates of solvation energies, entropy, as well as sampling effects. To this aim, we have used the recently developed PMISP method, which has been calibrated and tested for the biotin−avidin complexes and has been shown to give protein−ligand interaction energies with an accuracy of 3−5 kJ/mol compared to full QM calculations with the same method.[9] We have also shown that the surrounding protein can be modeled by a PMISP/MM approach, and we have tested different sizes of the PMISP model.[25] In this paper, we have combined this method with the PCM solvation model, which has been used before to calculate ligand-binding affinities.[60] These methods are combined to evaluate binding affinities through the widely used MM/PBSA approach,[7] using a normal-mode estimate of the entropy change during ligand binding and sampling geometries from an MD simulation of the solvated complex.

Unfortunately, the results with this PMISP/MM/PCM/$T\Delta S$ approach are rather poor in both absolute and relative terms, with a TR MAD of 19 kJ/mol, i.e., worse than a standard MM/PBSA method for the same problem (11 kJ/mol) or docking results with AutoDock (13−14 kJ/mol). The reason for the poor absolute binding affinities is probably the contribution from the PCM nonpolar solvation energies, which is 60−180 kJ/mol more positive than that obtained with the simple SASA-based method in standard MM/PBSA. On the other hand, the relative energies are not much improved when the PCM nonpolar solvation energies are replaced by SASA-based estimates (TR MAD = 19 kJ/mol). In fact, the best results are obtained without any nonpolar terms at all (TR MAD = 17 kJ/mol). A possible problem with the present approach is that we calculate the PMISP energies on snapshots from a MD simulation performed with another energy function. It is conceivable that the mismatch between the two potential energy surfaces may give rise to sizable errors.[38]

The use of high-level QM interaction energies allows us to address the important question of whether the accuracy of the MM/PBSA method is limited by the accuracy of the force field. Clearly, our results indicate that this is not the case. However, we cannot say whether the limitation resides in the solvation model or in the statistical-mechanical approximations inherent in the method. For this, it would be necessary to have a solvation model that is specifically parametrized for the electrostatics corresponding to high-level QM calculations and that is consistent for both large and small systems, so that the solvation contribution to binding energies can be accurately calculated. In this study, we have pointed out several qualitative differences between the nonpolar part of the PCM model and the corresponding SASA estimate, thus providing a starting point for understanding how such an ideal method should behave. In particular, it must be settled what type of expression (volume and area terms in PCM, only area terms in SASA) is most transferable between various solute sizes and what type of solute surface (van der Waals surface in PCM, solvent-accessible surface area in SASA) is most easily parametrized. Intuitively, it seems questionable to use a surface (e.g., the van der Waals surface) that contains contributions for atoms deeply buried in the protein and gives rise to many cavities.

Acknowledgment. This investigation has been supported by grants from the Swedish research council and AstraZeneca, as well as by computer resources of Lunarc at Lund University. J.K. thanks the Villum Kann Rasmussen foundation and the Danish research council for financial support.

References

(1) Gohlke, H.; Klebe, G. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644–2676.

(2) Beveridge, D. L.; Dicapua, F. M. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431–492.

(3) Lee, F. S.; Chu, Z. T.; Bolger, M. B.; Warshel, A. *Protein Eng.* **1992**, *5*, 215–228.

(4) Sham, Y. Y.; Chu, Z. T.; Tao, H.; Warshel, A. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 393–407.

(5) Warshel, A.; Sharma, P. K.; Kato, M.; Parson, W. W. *Biochim. Biophys. Acta* **2006**, *1764*, 1647–1676.

(6) Hansson, T.; Marelius, J.; Åqvist, J. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 27–35.

(7) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. *Acc. Chem. Res.* **2000**, *33*, 889–897.

(8) Pearlman, D. A.; Charifson, P. S. *J. Med. Chem.* **2001**, *44*, 3417–3423.

(9) Söderhjelm, P.; Ryde, U. *J. Phys. Chem. A* **2009**, *113*, 617–627.

(10) Khoruzhii, O.; Donchev, A. G.; Galkin, N.; Illarionov, A.; Olevanov, M.; Ozirin, V.; Queen, C.; Tarasov, V. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 10378–10383.

(11) Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; Wollacott, A. M.; Westerhoff, L. M.; Merz, K. M. *Drug Discovery Today* **2007**, *12*, 725–731.

(12) Raha, K.; Merz, K. M. *J. Am. Chem. Soc.* **2004**, *126*, 1020–1021.

(13) Raha, K.; Merz, K. M. *J. Med. Chem.* **2005**, *48*, 4558–4575.

(14) Fukuzawa, K.; Mochizuki, Y.; Tanaka, S.; Kitaura, K.; Nakano, T. *J. Phys. Chem. B* **2006**, *110*, 16102–16110.

(15) Nakanishi, I.; Fedorov, D. G.; Kitaura, K. *Proteins: Struct., Funct., Bioinf.* **2007**, *68*, 145–158.

(16) Zhang, D. W.; Xiang, Y.; Zhang, J. Z. H. *J. Phys. Chem. B* **2003**, *107*, 12039–12041.

(17) Zhang, D. W.; Xiang, Y.; Gao, A. M.; Zhang, J. Z. H. *J. Chem. Phys.* **2004**, *120*, 1145–1148.

(18) Zhang, D. W.; Zhang, J. Z. H. *Int. J. Quantum Chem.* **2005**, *103*, 246–257.

(19) Mei, Y.; Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. *Proteins: Struct., Funct., Bioinf.* **2005**, *59*, 489–495.

(20) He, X.; Mei, Y.; Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 423–432.

(21) Wu, E. L.; Mei, Y.; Han, K.; Zhang, J. Z. H. *Biophys. J.* **2007**, *92*, 4244–4253.

(22) Bettens, R. P. A.; Lee, A. M. *Chem. Phys. Lett.* **2007**, *449*, 341–346.

(23) Jurecka, T.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

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

(25) Söderhjelm, P.; Aquilante, F.; Ryde, U. *J. Phys. Chem. B* **2009**, *113*, 11085–11094.

(26) Weber, P. C.; Ohlendorf, D. H.; Wendolowski, J. J.; Salemme, F. R. *Science* **1989**, *243*, 85–88.

(27) Weber, P. C.; Wendoloski, J. J.; Pantoliano, M. W.; Salemme, F. R. *J. Am. Chem. Soc.* **1992**, *114*, 3197–3200.

(28) Pugliese, L.; Coda, A.; Malcovati, M.; Bolognesi, M. *J. Mol. Biol.* **1993**, *231*, 698–710.

(29) Livnah, O.; Bayer, E. A.; Wilchek, M.; Sussman, J. L. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5076–5080.

(30) Green, N. M. *Biochem. J.* **1966**, *101*, 774–780.

(31) Green, N. M. *Adv. Protein Chem.* **1975**, *29*, 85–133.

(32) Green, N. M. *Methods Enzymol.* **1990**, *184*, 51–67.

(33) Miyamoto, S.; Kollman, P. A. *Proteins: Struct., Funct., Genet.* **1993**, *16*, 226–245.

(34) Wang, J.; Dixon, R.; Kollman, P. A. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 69–81.

(35) Kuhn, B.; Kollman, P. A. *J. Med. Chem.* **2000**, *43*, 3786–3791.

(36) Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. *J. Med. Chem.* **2005**, *48*, 4040–4048.

(37) Brown, S. P; Muchmore, S. W. *J. Chem. Inf. Model.* **2006**, *46*, 999–1005.

(38) Weis, A.; Katebzadeh, K.; Söderhjelm, P.; Nilsson, I.; Ryde, U. *J. Med. Chem.* **2006**, *49*, 6596–6606.

(39) Genheden, S.; Ryde, U. *J. Comput. Chem.* **2010**, *31*, 837–846.

(40) Gagliardi, L.; Lindh, R.; Karlström, G. *J. Chem. Phys.* **2004**, *121*, 4494–4500.

(41) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599–3605.

(42) Riley, K. E.; Hobza, P. *J. Phys. Chem. A* **2007**, *111*, 8257–8263.

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

(44) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(45) Cieplak, P.; Caldwell, J.; Kollman, P. A. *J. Comput. Chem.* **2001**, *22*, 1048–1057.

(46) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222–239.

(47) Beebe, N. H. F.; Linderberg, J. *Int. J. Quantum Chem.* **1977**, *12*, 683–705.

(48) Koch, H.; Sánchez de Merás, A.; Pedersen, T. B. *J. Chem. Phys.* **2003**, *118*, 9481–9484.

(49) Aquilante, F.; Pedersen, T. B.; Lindh, R. *J. Chem. Phys.* **2007**, *126*, 194106.

(50) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.

(51) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.

(52) Stouch, T. R.; Williams, D. E. *J. Comput. Chem.* **1992**, *13*, 622–632.

(53) Reynolds, C. A.; Essex, J. W.; Richards, W. G. *J. Am. Chem. Soc.* **1992**, *114*, 9075–9079.

(54) Sigfridsson, E.; Ryde, U.; Bush, B. L. *J. Comput. Chem.* **2002**, *23*, 351–364.

(55) Söderhjelm, P.; Ryde, U. *J. Comput. Chem.* **2009**, *30*, 750–760.

(56) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker,

R. C.; Zhang, W.; Merz, K. M. ; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G. ; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*; University of California: San Francisco, CA, 2008.

(57) Bandyopadhyay, P.; Gordon, M. S.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **2002**, *116*, 5023–5032.

(58) Cances, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032–3041.

(59) Frisch, A. E.; Frisch, M. J.; Trucks, G. W. *Gaussian 03 User'S Reference*; Gaussian, Inc.: Wallingford, CT, 2003; p 205.

(60) Li, H.; Pomelli, C. S.; Jensen, J. H. *Theor. Chem. Acc.* **2003**, *109*, 71–84.

(61) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

(62) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210–3221.

(63) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C. Pople, J. A. *Gaussian 03*, Revision D.01; Gaussian, Inc.: Wallingford CT, 2004.

(64) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2011–2033.

(65) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *110*, 16066–16081.

(66) Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. *Biochemistry* **1981**, *20*, 849–855.

(67) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978–1988.

(68) Florián, J; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 5583–5595.

(69) Kongsted, J.; Söderhjelm, P.; Ryde, U. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 395–409.

(70) Fratev, F.; Jonsdottir, S. O; Mihaylova, E.; Pajeva, I. *Mol. Pharmaceutics* **2009**, *6*, 144–157.

(71) Grazioso, G.; Cavalli, A.; de Amici, M.; Recanatini, M.; de Micheli, C. *J. Comput. Chem.* **2008**, *29*, 2593–2603.

(72) Fogolari, F.; Moroni, E.; Wojciechowski, M.; Baginski, M.; Ragona, L.; Molinari, H. *Proteins* **2005**, *59*, 91–103.

(73) Gilson, M. K; Honig, B. *Proteins: Struct., Funct., Genet.* **1998**, *4*, 7–18.

(74) Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754–2759.

(75) Swanson, J. M. J.; Henchman, R. H.; McCammon, J. A. *Biophys. J.* **2004**, *86*, 67–74.

(76) Gräter, F.; Schwarzl, S. M.; Dejaegere, A.; Fischer, S.; Smith, J. C. *J. Phys. Chem. B* **2005**, *109*, 10474–10483.

(77) Wang, M. L.; Wong, C. F. *J. Chem. Phys.* **2007**, *126*, 026101.

(78) Kaukonen, M.; Söderhjelm, P.; Heimdal, J.; Ryde, U. *J. Phys. Chem. B* **2008**, *112*, 12537–12548.

(79) Kongsted, J.; Ryde, U. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 63–71.

(80) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639–1662.

(81) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. *J. Comput. Chem.* **2007**, *28*, 1145–1152.

(82) Cossi, M.; Tomasi, J.; Cammi, R. *Int. J. Quant. Chem. Quant. Chem. Symp.* **1995**, *29*, 695–702.

(83) Pullman, B. *Intermolecular Interactions, from diatomics to biomolecules*; John Wiley & Sons: Chichester, U.K., 1978; p 69.

(84) Pierotti, R. A. *Chem. Rev.* **1976**, *76*, 717–726.

(85) Caillet, J.; Claverie, P. *Acta Crystallogr., Sect. B* **1978**, *34*, 3266–3273.

(86) Floris, F.; Tomasi, J. *J. Comput. Chem.* **1989**, *10*, 616–627.

(87) Tan, C.; Tan, Y.-H.; Luo, R. *J. Phys. Chem. B* **2007**, *111*, 12263–12274.

(88) Chandler, D.; Andersen, H. C. *J. Chem. Phys.* **1972**, *57*, 1930–1937.

(89) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *112*, 10391–10417.

(90) Genheden, S.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Söderhjelm, P.; Ryde, U. An MM/3D-RISM approach for ligand-binding affinities. *J. Phys. Chem. B*, Submitted.

(91) Cossi, M; Barone, V; Cammi, R; Tomasi, J. *Chem. Phys. Lett.* **1996**, *255*, 327–335.

(92) DeChancie, J.; Houk, K. N. *J. Am. Chem. Soc.* **1999**, *129*, 5419–5429.

CT9006986

JCTC Journal of Chemical Theory and Computation

Active Site, Catalytic Cycle, and Iodination Reactions of Vanadium Iodoperoxidase: A Computational Study

Luis F. Pacios*,† and Oscar Gálvez‡

Departamento de Biotecnología, Unidad de Química y Bioquímica, E.TSI Montes, Universidad Politécnica de Madrid, 28040 Madrid, Spain, and Departamento de Física Molecular, Instituto de Estructura de la Materia, C.S.I.C., Serrano 121, 28006 Madrid, Spain

Received January 21, 2010

Abstract: A combined computational study using molecular surfaces and Poisson−Boltzmann electrostatic potentials for proteins and quantum calculations on complexes representing the vanadate cofactor throughout the catalytic cycle is employed to study the activity of vanadium iodoperoxidase (VIPO) from alga *Laminaria digitata*. A model structure of VIPO is compared with available crystal structures of chloroperoxidases (VClPOs) and bromoperoxidases (VBrPOs) focusing on properties of the active site that concern halogen specificity. It is found that VIPO displays distinctive features regarding electrostatic potentials at the site cavity and the local topography of the cavity entrance. Quantum calculations on cofactor stages throughout the catalytic cycle reveal that, while steps involving binding of hydrogen peroxide and halide oxidization agree with available data on VBrPO, final formation and subsequent release of hypohalous acid could follow a different pathway consisting of His476-assisted protonation of bonded hypoiodite and further displacement by a water molecule. *Ab initio* free energies of reaction computed to explore iodination of organic substrates predict strongly exoergonic reactions with HOI, whereas other possible iodination reagents give thermodynamically disfavored reactions.

Introduction

Atmospheric iodine and its potential role in the catalytic destruction of ozone have attracted considerable attention in the two past decades.[1−3] It has been shown that condensable iodine oxide vapors can nucleate efficiently to form aerosols, which may contribute to form cloud condensation nuclei and hence have an impact on the climate.[4] The most abundant oxide, IO, is formed after photolysis of reactive iodine precursors and subsequent reaction of I atoms with atmospheric O_3. Recent measurements have shown high levels of IO in coastal Antarctica,[5] which, together with the relevance of iodine for atmospheric chemistry, opens the question of the release mechanisms required to account for such large amounts of iodine. High concentrations of IO have been explained by a mechanism for iodine release triggered by the biological processing of iodide (I^-) and the production of hypoiodous acid (HOI) from algae.[3] It is generally assumed that iodine in the atmosphere has a natural origin since no anthropogenic sources are known.[6] While low levels are found in soils, continental waters, and terrestrial plants, oceans are a major source. The iodine biogeochemical cycle involves large exchanges in the marine boundary layer (MBL) in which iodine is transferred from oceans to the atmosphere.[7] This process occurs by direct emission of I_2 and volatile iodinated compounds from open ocean (via phytoplankton) and coastal areas where macroalgae are a major contributor through the production of volatile iodocarbons.[8]

Large uncertainties still remain in assessing global emissions of volatile iodine compounds at the MBL in the iodine biogeochemical cycle, yet it is generally assumed that both seaweeds and marine phytoplankton release iodocarbons.[9]

* Corresponding author. E-mail: luis.fpacios@upm.es.

† Universidad Politécnica de Madrid.

‡ C.S.I.C.

10.1021/ct100041x
Published on Web 04/14/2010

In coastal environments, brown macroalgae and especially kelps are considered as major contributors to the flux of iodine.[10,11] Kelp species such as *Laminaria digitata* are indeed the most effective iodine accumulators among all living systems.[12] In addition to iodinated organic compounds, IO is also detectable in the atmosphere above kelp beds,[7b] and it has been reported that kelps release I_2, a major source of particle formation in coastal areas.[13] However, the reasons why seaweeds produce iodocarbons as well as the ecological significance of these compounds remain largely unknown. It has been recently shown that I^- accumulation provides kelps with an inorganic antioxidant, the first described in living systems, that constitutes a protection against oxidative stress.[11]

In seawater, iodide reacts spontaneously with hydrogen peroxide to produce HOI:

$$I^- + H_2O_2 \rightleftarrows HOI + OH^- \qquad (1)$$

but this reaction is slow. It was already observed in 1929 that this process occurs efficiently in cell walls of *L. digitata*, which led to the proposal of an "iodine oxidase" responsible for the catalyzed oxidation of iodide.[14] More recently, this oxidation has been confirmed to occur in the *Laminaria* apoplast where hypoiodous acid is formed virtually undissociated (pK = 10.64) at the pH of seawater, being itself in equilibrium with molecular iodine:[12]

$$I^- + HOI + H^+ \rightleftarrows H_2O + I_2 \qquad (2)$$

Triiodide is in turn formed as the result of the equilibrium

$$I^- + I_2 \rightleftarrows I_3^- \qquad (3)$$

but I_2/I^- complexation is weak in diluted solutions of iodide.[10] It is currently well established that the general catalytic role of halide oxidation in marine algae is actually played by vanadium-dependent haloperoxidases (VHPOs).[10−13,15−18] A VHPO enzyme specialized in oxidating iodide in cell walls of *L. digitata* should explain the high efficiency for iodine accumulation in this organism. Since both HOI and I_2 are more lipophilic than I^-, their transport should be facilitated across membrane lipid bilayers at both iodine uptake and efflux occurring in response to biotic and abiotic stresses. Iodide detoxifies both aqueous (mainly hydrogen peroxide) and atmospheric (mainly ozone) oxidants,[11] and HOI is the central intermediate for all cases. In the absence of organic substrates, VHPOs catalyze the oxidation of a second equivalent of H_2O_2, resulting in the formation of singlet dioxygen and halide. In the presence of organic substrates, VHPOs catalyze the halogenation of a wide range of organic molecules.[15] The presence of these enzymes in the apoplasm of marine algae could thus explain the production of I_2 and iodocarbons, compounds for which a defense function has been proposed on the basis of their high microbial toxicity.[11,19]

The majority of naturally occurring organohalogens and nearly all brominated and iodinated natural products are produced by marine organisms on a large scale.[20] Although the biogenesis of these compounds (many of them with biological activities of pharmacological interest) has been studied for a long time, more new enzymes for halogenation have been discovered in the past five years than in the four decades before.[18] Haloperoxidases are major halogenating enzymes classified according to the most electronegative halide they oxidize. VHPOs, present in macroalgae, fungi, and bacteria, contain a ligated vanadate ion and use hydrogen peroxide to oxidize a halide (X^-) to its corresponding hypohalous acid (HOX), an intermediate chemically equivalent to an electrophile reagent "X^+".[17−19] This reaction product has been customarily considered as a mixture of different species "X^+" = HOX, X_2, and X_3^- all of them able to halogenate appropriate organic substrates if present.[15−19]

Crystal structures have been reported for H=Cl (VClPO) haloperoxidase from the fungus *Curvularia inaequalis*,[21,22] H=Br (VBrPO) from the brown alga *Ascophyllum nodosum*,[23] and VBrPO from the red alga *Corallina officinalis.*[24] Kinetics and structural studies of peroxide-bound forms and mutants of VClPO as well as mutants of VBrPO have been also developed.[25] All these studies have led to a general consensus regarding the role of the residues that bind the vanadate cofactor within the active site.[17,18] However, for H=I (VIPO) haloperoxidases, no experimental structure is available. A distinct vanadium-dependent iodoperoxidase has been purified from *L. digitata*, and kinetic studies have shown that it oxidizes iodide specifically, although competition experiments indicate that bromide is a competitive inhibitor.[16] This VIPO is the obvious candidate to explain not only iodine uptake and efflux observed in kelps but also the formation of iodocarbons released by these algae. The amino acid sequence of *L. digitata* VIPO shows the conserved residues in the active site of all VHPOs except two significant differences, one with respect to VClPOs and another with respect to VBrPOs, which point to specific features that are discussed below.

Since the resolution of crystal structures of these proteins precludes solving hydrogen atoms from experimental electron densities, details regarding the different protonation states of the vanadate cofactor throughout the catalytical cycle—a central issue in the mechanism—have to be addressed theoretically. Reports on this subject have been been published in recent years for chlorine and bromine[26−31] but not for iodine. The features that confer halide selectivity remain as yet unexplained. The nature of the species "X^+" that ultimately halogenates organic compounds is an unanswered question. Although on a kinetic and structural basis it was considered unlikely that the halide-binding sites were found at the vanadium center,[32,33] recent experimental work[26] suggests otherwise. The specificity for halogenation of organic substrates is also an as yet unsolved question. Recent studies have shown specificity in bromination of cyclic molecules by VBrPOs,[15,17,18] but other results point to a lack of selectivity,[34] which has been interpreted to mean that the "Br^+" intermediate was freely diffusible to facilitate bromination outside the active site.[35]

In this contribution, we focus on the iodoperoxidase VIPO enzyme identified in *L. digitata*.[16] On the basis of a reliable three-dimensional (3D) model of the whole protein and by means of a theoretical combined approach that uses quantum calculations and structural modeling along with Poisson−Boltzmann (PB) electrostatic potentials and protein surface analyses, we address the following issues: (1) The active site of VIPO is compared to the sites of VClPOs and VBrPOs with known structures to identify differences. (2) The topography of entrances to the active site cavity in the protein surfaces of VHPOs are compared, highlighting putative distinctive features in VIPO provided by the PB electrostatic potential. (3) Characterization of the vanadate cofactor and its protonation states throughout the catalytic cycle is explored on the basis of quantum calculations and 3D modeling of the active site. Results point to a specific protonation of the iodine-bound cofactor assisted by a nearby histidine, a step not proposed before. (4) Iodination reactions of several organic compounds selected to account for representative volatile and nonvolatile iodocarbons are thermodynamically studied at high-level *ab initio* correlated calculations. Free energies of reactions with the three possible iodinating species HOI, I_2, and I_3^- are calculated. Results indicate that only hypoiodous acid give clearly exergonic iodination of organic substrates.

Methods

The 624-amino acid sequence of the vanadium peroxidase (VIPO) purified from *L. digitata*[16] deposited in the PeroxiBase database[36] (entry 4072, UnitProtKB: Q4LDE6) was utilized to generate a model structure using the Swiss-model homology-modeling server.[37] The model was constructed from the X-ray structure of *A. nodosum* VBrPO[23] as a template and includes residues 66−623 (coordinates in PDB format and Figure S1 in the Supporting Information). Sequence identity between VIPO and VBrPO was 57.3%, and their structural superposition gave a RMSD for backbone atoms of 0.26 Å. A geometry obtained in quantum calculations for vanadate bonded to an imidazole ring was inserted in the VIPO model structure to fit the His555 side chain. Insertion of heterogroups and structure analyses were performed with Chimera 1.3.[38]

Besides the modeled structure of VIPO, the following crystal structures of VHPOs were used for comparison: resting forms of native (PDB code 1IDQ[21]) and recombinant (PDB code 1VNI[22]) VClPO from *C. inaequalis*, its peroxide form VClPO (PDB code 1IDU[21]), and its resting form of VBrPO from *A. nodosum* (PDB code 1QI9[23]). Solvent-excluded molecular surfaces for the five structures were obtained for a 1.4-Å-radius probe sphere and rendered with PyMOL 1.2.[39] Amino acid contributions to accessible surface areas and solvent exposure percentages were computed with Arvomol[40] and SurfRace.[41] Poisson−Boltzmann (PB) electrostatic potentials were obtained with APBS 1.2.1,[42] assigning AMBER99 charges[43] and atomic radii with PDB2PQR[44] to all the atoms including hydrogens added and optimized with this program. PB potentials were obtained at ~0.5 Å grid spacings around the more than 8000 atoms of which these proteins are composed by solving the nonlinear PB equation[45] in single-point multigrid calculations at meshes of 225 × 193 × 225 points at 298.15 K, 0.150 M ionic strength, and dielectric constants of 4 for proteins and 78.54 for water. Output meshes were processed in scalar OpenDX format with PyMOL 1.2. PB potential values are given in units of kT/e (k, Boltzmann's constant and e, unit charge; 1 kT/e = 2.48 kJ mol^{-1} at 298.15 K). Since no AMBER parameters exist for vanadate, atomic charges were obtained at separate B3LYP/cc-pVTZ calculations to produce charges fit to the quantum electrostatic potential according to the CHelpG scheme.[46] An atomic radius of 1.6612 Å for oxygen (AMBER value for phosphate) and 2.0 Å for vanadium were assumed for vanadate.

Structures of vanadium cofactors at distinct protonation states throughout the catalytic cycle of VIPO were optimized in quantum calculations with the B3LYP hybrid functional and cc-pVTZ basis sets upon using an imidazole ring to represent the His555 side chain. Harmonic vibrational frequencies were then computed at the B3LYP geometries. The cc-pVTZ basis set for vanadium atom was taken from the basis set library by Balabanov and Peterson for transition elements,[47] while an equivalent valence-only basis set was developed for the iodine atom using a shape-consistent averaged relativistic effective potential (AREP)[48] to replace the 46-electron core.[49] This pseudopotential formalism has been shown to account for relativistic effects when used with properly optimized basis sets.[50] The cc-pVTZ basis set for iodine was generated following the method prescribed by Martin and Sundermann to construct correlated consistent basis sets for relativistic effective core potentials[51] using MOLPRO2006.[52] This AREP-optimized basis set (Table S1 in the Supporting Information) was tested in benchmark calculations on molecules containing iodine (for an illustrative example of its reliable performance, see Table 3 below). Full geometry optimizations using analytic gradients without symmetry constraints were performed and frequencies analyzed to characterize minima. Environmental effects were accounted for by computing Polarizable Continuum Model (PCM) energies with three dielectric constants: $\varepsilon = 4$ to simulate protein interiors,[42] $\varepsilon = 40$ to simulate charge−charge interactions in protein active sites as used before in quantum calculations on VBrPO,[29,30] and $\varepsilon = 78.39$ for an aqueous medium.

With the aim of thermodynamically studying iodination reactions of organic substrates, we selected the following iodocarbons: CH_3I, CH_2I_2, and iodopropene isomers ($CHI{=}CH{-}CH_3$, $CH_2{=}CI{-}CH_3$, and $CH_2{=}CH{-}CH_2I$) to represent volatile iodocarbons and CH_2ICHO, CH_2ICOOH, and iodophenol isomers to represent nonvolatile iodocarbons. The corresponding reactions of the possible iodinating species HOI, I_2, and I_3^- with methane, propene, acetaldehyde, acetic acid, and phenol to yield iodocarbons and byproducts H_2O, HI, and I^- were studied by means of *ab initio* calculations. Free energies were computed at 298.15 K as $\Delta_R G^0_{298} = \Delta_R H^0_{298} - T\Delta_R S^0_{298}$ from enthalpies of formation $\Delta_f H^0_{298}$ and absolute entropies S^0_{298} obtained with aug-cc-pVTZ basis sets (an AREP-optimized aug-cc-pVTZ set was constructed for iodine: see Table S1 in the Supporting Information). Geometries and vibrational frequencies to compute zero-point energies (ZPE) and thermal corrections were obtained at MP2 calculations. CCSD(T) energies were also computed at MP2 geometries. ZPE-corrected $\Delta_f H^0_0$ values were first calculated at 0 K from known enthalpies of formation of isolated atoms, as suggested by Curtiss et al.[53] for the set of molecules (*A*) H_2O, CH_4, I_2, HI, HOI, $CH_2{=}CH{-}CH_3$, CH_3CHO, and CH_3COOH. $\Delta_f H^0_0$ values for the set of molecules (*B*) CH_3I and CH_2I_2 were obtained from isogyric reactions $CH_4 + IH \rightarrow (B) + H_2$, while those for (*C*) the rest of the iodocarbons, $\Delta_f H^0_0$, were obtained from isodesmic reactions substrate $+ CH_3I \rightarrow (C) + CH_4$. Reaction 3 was used for I_3^-. $\Delta_f H^0_{298}$ values were then calculated by correction to $\Delta_f H^0_0$ using atomic $H^0_{298} - H^0_0$ thermal corrections from NIST-JANAF data[54] for molecules in set *A* and MP2 values for sets *B* and *C*. MP2 results for S^0_{298} were used in all cases except I^-, for which the experimental value was taken. Thermodynamical computed data and experimental values where available are gathered in Table S2 in the Supporting Information. All the calculations to obtain geometries, energies, frequencies, and PCM results were performed with Gaussian 03.[55]

Results and Discussion

Active Site of Vanadium Haloperoxidases. We start addressing the reliability of the model structure of VIPO used in this work. Homology modeling, the approach followed to obtain the model structure, is currently considered the method of choice to predict the protein structure if the target protein has ~50% sequence identity with a template protein.[37,56] Estimating the accuracy of homology-modeled structures has been an issue thoroughly studied for more than 20 years. During the past decade, blind tests such as those provided by the CASP (Critical Assessment of protein Structure Prediction) experiments have permitted the objective evaluation of the reliability of protein modeling methods by examining the quality of predictions. It has been demonstrated that the core atoms of protein models sharing 50% sequence identity with their templates deviate by RMS ~ 1.0 Å from crystal structures.[56] Given the typical resolution of X-ray protein geometries, this deviation is more than acceptable. What is indeed more important here is that, if one is dealing with a family of proteins in which the function is maintained, as it is the case for VHPOs, binding and active sites are found to show even less deviation, as evolution tends to alter those sites rather conservatively.[37,56]

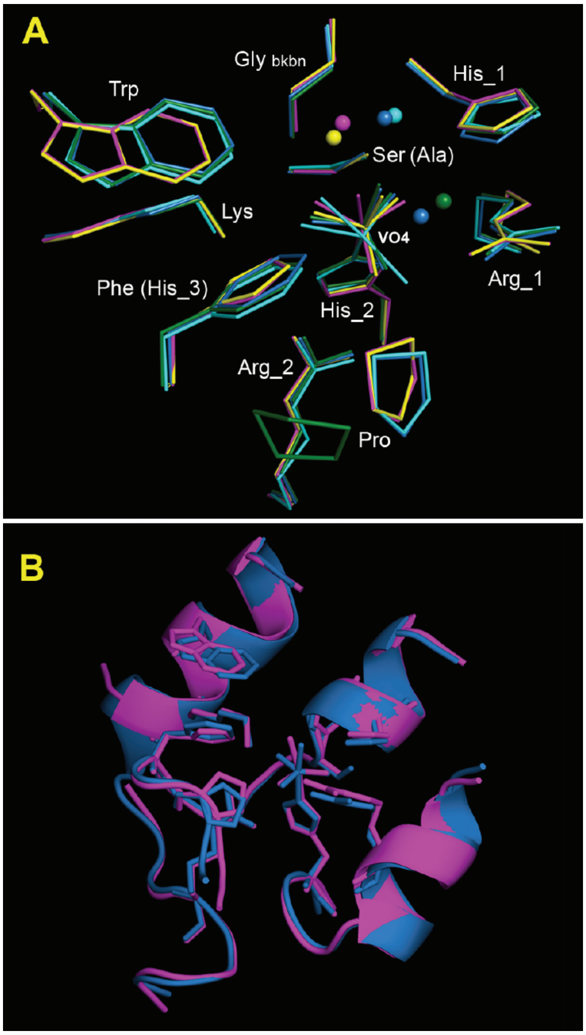
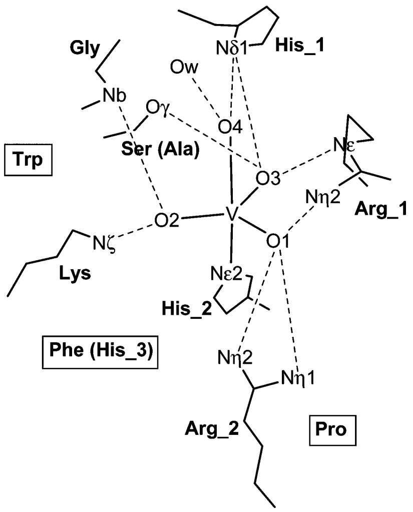


Figure 1. (A) Site defined by residues within a 4 Å radius around vanadate in crystal structures of two resting forms of VClPOs (1IDQ, deep blue; 1VNI, green), peroxo form of VClPO (1IDU, cyan), resting form of VBrPO (1QI9, yellow), and model structure of VIPO (magenta). Spheres represent water molecules. Apart from the glycine backbone, only side chains are shown. Residue numbering is given in Scheme 1. (B) Structural region around the entry channel to the active site in the crystal structure of resting form of VClPO (1IDQ, blue) and model structure of VIPO (magenta). Sticks represent residues at the active site. Arrangements depicted in A and B result from superposing the whole proteins without constraining fit in the regions displayed.

Most of the residues in active sites of VClPOs and VBrPOs are conserved in *L. digitata* VIPO.[16] Superposition of the VIPO model structure with available crystal structures of *C. inaequalis* VClPOs[21,22] and *A. nodosum* VBrPO[23] shows a remarkable fit of residues within 4 Å around vanadate (Figure 1A). Given that VBrPO was used as a template to model VIPO and that these proteins share 57% sequence identity, the near coincidence between their active sites is the expected result. However, neither of the resting VClPO structures was used to model VIPO; hence, the good agreement between their active sites lends additional support to the predicted geometry of VIPO active site. Since the entry channel leading to the active site seems to play a crucial role in fine-tuning the activity of VHPOs (see the next subsection), this structural region must be also reliably predicted. Figure 1B shows the comparison between structural sections around the entry channel for the resting form of *C. inaequalis* native VClPO[21] and VIPO (the comparison with the template VBrPO yields a nearly indistinguishable superposition). It is seen that not only α helices but also coil segments in this region are in a rather satisfactory agreement. The good match of conformation around this entry for the three VHPOs gives thus a sound basis for comparing its local surface topography, as discussed below. It must be stressed that Figure 1 was produced upon superposing the proteins to fit their whole structures, not the displayed regions alone.

On the other side, the close agreement between resting and peroxide forms of VClPO indicates that the catalytic activity is carried out without noticeable structural changes at the active site. Vanadium is coordinated to the protein by a single axial His_2 ligand in a trigonal bipyramidal

Scheme 1. Schematic Drawing of Residues in the Active Site of Vanadium Haloperoxidases VHPOs (Figure 1), Atom Labeling and Residue Numbers in the Corresponding Sequences of H = Cl, Br, and I Enzymes[a]



	VClPO	VBrPO	VIPO
Arg_1	490	480	549
His_1	404	418	483
Gly	403	417	482
Ser (Ala)	S402	S416	A481
Trp	350	338	400
Lys	353	341	403
Phe (His_3)	F397	H411	H476
Arg_2	360	349	411
Pro	395	409	474
His_2	496	486	555

[a] Amino acid atoms at intermolecular distances from vanadate oxygens shorter than 3.5 Å (dashed lines) are labeled. Nb is glycine backbone nitrogen, and Ow is water oxygen. Names inside boxes indicate residues not directly interacting with vanadate.

geometry, while oxygens of vanadate are hydrogen-bonded to Arg_1, His_1, Lys, Arg_2, side chains, and the Gly backbone (Scheme 1). The spatial coincidence of these residues in all VHPOs demonstrates that they define a rigid scaffold setting a hydrogen bond network around vanadate regardless of the halogen specificity. However, some differences are noticed. As it was observed before,[16] VIPO has an alanine at a position occupied by serine in other haloperoxidases. It has been noted that this substitution, which results in the loss of one of the hydrogen bonds of vanadate, could be related with the inability of VIPO to oxidize bromide or chloride.[16] VBrPO and VIPO have histidine substituting phenylalanine in VClPO at a location where no direct interaction with the vanadium cofactor is possible. It was conjectured that, whereas Phe397 in VClPO might take part in halide binding through its aromatic ring,[17,57] His411 in VBrPO could participate in proton transfer during the enzymatic reaction,[15,18] though this role was not specifically elucidated. We propose in this work that His476 (VIPO numbering) should in fact participate in a water-assisted proton transfer to the apical oxygen bonded to iodine before releasing hypoiodous acid (see below). Two other differences that have apparently gone as yet unnoticed concern the locations of Trp and Pro, residues conserved in VHPOs not interacting with vanadate. Trp and Pro show structural displacements between VClPO on one side and VBrPO/VIPO on the other side, particularly Pro in recombinant *C. inaequalis* VClPO.[22] Since halides are known to bind to hydrophobic pockets (apart from basic residues)[16,18] and it seems unlikely that halide specificity, an as yet unsolved question, could arise from the rigid scaffold, it seems reasonable to think of Trp and Pro besides Phe (His_3) as residues involved in halide selectivity. We point in the next subsection to the topography of the active site surface and the local electrostatic potential as properties displaying differentiated features in this regard.

In contrast to other haloperoxidases which function as redox catalysts, the vanadium atom in VHPOs remains in the (v) oxidation state throughout the catalytic cycle; hence its role is to act as a Lewis acid in the activation of the primary oxidant H_2O_2.[21−24] Excluding the largely distorted geometry of the peroxo form, vanadate groups differ significantly even though His_2 imidazole rings overlap when VHPO's structures are superposed (Figure 1). Given that the resolution of the available experimental structures does not allow solving hydrogens and the uncertainty of X-ray derived bond distances precludes unambiguous assignments, the protonation states of vanadate have to be theoretically deduced as was done before for VClPO and VBrPO[27−31] and is reported here for VIPO. Nevertheless, vanadate bond lengths in crystal geometries hint at differences even between both resting forms of VClPO. As shown below, V−O bond lengths of about 1.5−1.6 Å are typical of V═O double bonds, while lengths longer than 1.8 Å indicate V−OH bonds. All the equatorial O−V bonds in resting forms of VClPO and VBrPO have distances between 1.52 and 1.64 Å. However, apical O4−V lengths are 1.88 Å in VClPO (1IDQ) and 1.77 Å in VBrPO, both values consistent with V−OH bonds, but 2.15 Å in VClPO (1VNI) and 2.19 Å in the VIPO model, which suggests instead a V••OH_2 interaction (see below).

Surfaces and Local Electrostatic Potential. X-ray structures of VHPOs show that the vanadium-binding site is positioned at the bottom of a deep funnel-shaped channel.[21−23] Although the VIPO structure is a model and as such merely temptative, the evidence presented in the preceding subsection allows one to reasonably discuss features in the region around the entry channel in VClPO, VBrPO, and VIPO on a similar footing. Comparison among molecular surfaces (Figure 2) reveals a different local topography in the funnel entrance. While VClPO exhibits a narrow entrance which is besides partially buried, VBrPO and VIPO display open, wider entrances, much greater in the latter. These differences become striking when one compares surface contributions from residues in the active site (Figure 3). All VHPOs have Ser(Ala) and Lys completely buried, while the bottom surface is made of Arg_2 and His_2 contributions. Arg_1, His_1, and Gly contribute significantly to the pocket wall surface, but as the cleft entrance is deeper in VClPO, their surfaces are barely accessible from outside (Figure 3A). On the contrary, Arg_1, Gly, and His_1 surfaces form large outer patches around the channel entrance in VIPO (Figure 3B), though only Arg_1 and His_1 have surfaces accessible to

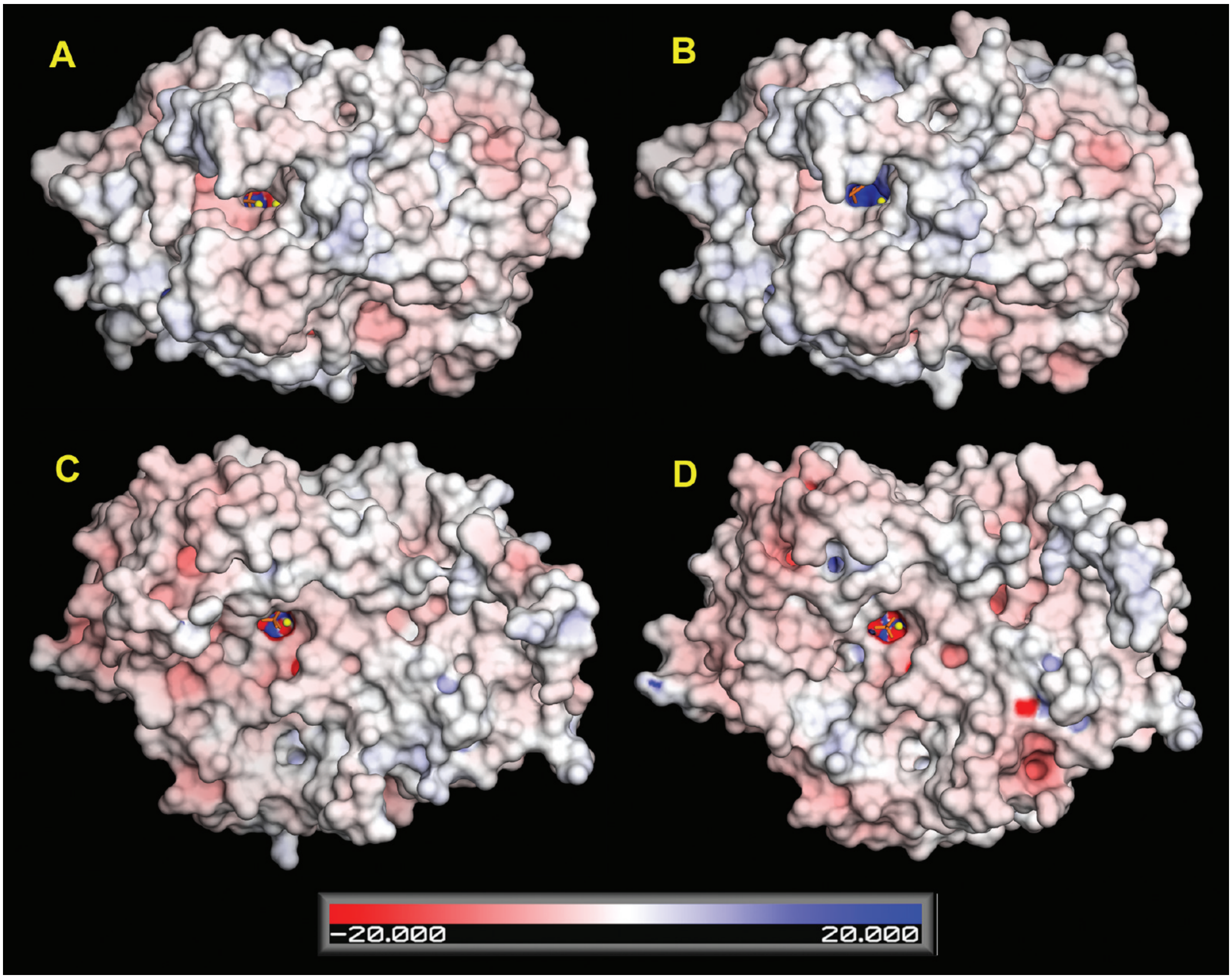


Figure 2. Poisson−Boltzmann electrostatic potential in the range −20 to +20 (units of *kT*/*e*) mapped onto the molecular surface of vanadium haloperoxidades. Vanadate cofactor in the substrate binding pocket is drawn as orange sticks and water molecules as yellow spheres. (A) VClPO (resting form, 1IDQ). (B) VClPO (peroxo form, 1IDU). (C) VBrPO (1QI9). (D) VIPO.

the solvent (Table 1). More remarkable are the differences regarding residues found to be shifted in structural superpositions discussed above. In VClPO, Pro and Phe form hydrophobic patches with similar surface areas at the entrance region inaccessible to solvent, and Trp is completely buried (Figure 3A, Table 1). In VIPO, His_3 and especially Pro and Trp are a great part of the surface around the cavity entrance forming a large hydrophobic area accessible to solvent (Figure 3B, Table 1). Comparing the small deep channel in VClPO with the wide open cleft in VIPO, one could conclude that the access cavity topography is a clearly differentiated feature which might affect halide selectivity.

Another revealing difference appears when one analyzes a property that has been hitherto scarcely explored: the electrostatic potential. The dominance of acidic versus basic amino acids yields very large negative total electric charges in these enzymes: −26*e* in VClPO, −43*e* in VBrPO, and −45*e* in VIPO. Therefore, the electrostatic potential at their surfaces displays a predominantly negative character, more significant in VIPO or VBrPO than in VClPO (Figure 2). However, in choosing a large range of values as that used to render Figure 2, the active site pocket shows a local electrostatic potential markedly stronger than any other region at the protein surface. Gathering four basic residues (Arg_1, His_1, Lys, and Arg_2) and four electronegative vanadate oxygens at the active site produces strong local electrostatic potentials at the pocket (Figure 4). Comparing these local potentials thus gives useful information that complements that provided by local topography. Note that the three resting forms show at the bottom of the pocket two negative regions, but whereas VClPO has a positive area in the surface near Arg_1 and His_1 (Figure 4A), VBrPO displays there a neutral/negative domain (Figure 4C), and VIPO shows only a negative domain (Figure 4D). This surface patch turns out to be located just above Lys and Ser (Ala), both completely buried residues (Table 1). In the crystal structures of VHPOs, it is found that Ser402 is at a short distance from vanadate equatorial oxygens in VClPO (O3···Oγ distance ~2.6 Å; see Scheme 1), and Ser416 is farther apart in VBrPO (O3···Oγ distance ~2.9 Å), whereas VIPO has alanine instead serine at that position. Since the scaffold around the cofactor is nearly identical in all VHPOs and the atomic charges used for vanadate to compute PB potentials in the three resting forms were exactly the same, the Ser/Ala change had to give rise to local differences in the electrostatic potential at the cavity.

Another factor of interest is the rather distinct depth of the substrate cleft. It must be stressed that all pockets displayed in Figure 4 correspond to the same clipping 20 Å depth from the outer surface. As discussed above with regard

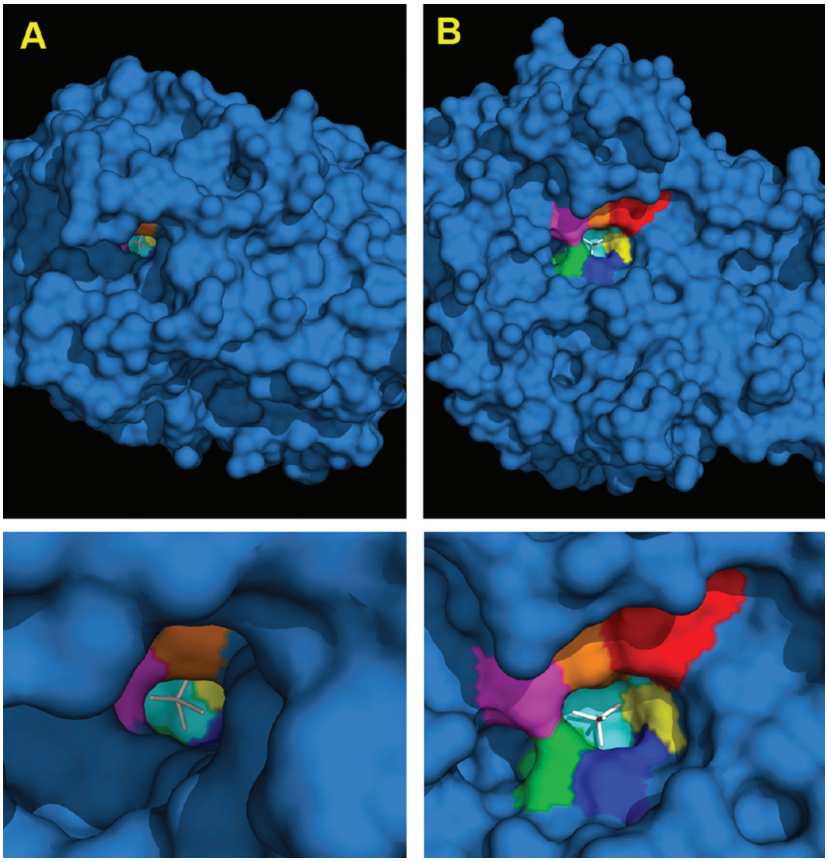


Figure 3. Molecular surface of VClPO (A) and VIPO (B) showing contributions of residues in the active site. Top: general views of the protein surface. Bottom: closeup views of the entrance to the substrate channel. Vanadate cofactor drawn as white sticks. Bottom of the pocket colored cyan and residue contributions colored as follows: Arg_1, green; His_1, marine blue; Gly, yellow; Trp, red; Phe (His_3), orange; and Pro, magenta. Residue numbering in Scheme 1.

Table 1. Surface Areas and Solvent-Exposure Percentages (In Parentheses) of Residues at the Active Site in Vanadium Haloperoxidases[a]

residue	VClPO resting (1IDQ)	VClPO peroxo (1IDU)	VBrPO resting (1QI9)	VIPO resting (Model)
Arg_1	8.7 (0.0)	11.5 (0.0)	32.8 (3.4)	39.2 (5.9)
His_1	29.8 (0.0)	23.4 (0.0)	37.1 (0.4)	36.3 (6.3)
Gly	13.3 (0.0)	15.5 (0.0)	17.8 (0.0)	18.8 (0.0)
Ser (Ala)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Trp	12.1 (0.6)	10.3 (0.6)	37.3 (4.5)	65.2 (10.0)
Lys	2.8 (0.0)	3.5 (0.0)	2.4 (0.0)	2.6 (0.0)
Phe (His_3)	26.3 (1.7)	26.5 (2.7)	23.9 (2.2)	39.6 (4.4)
Arg_2	27.3 (3.1)	29.3 (2.5)	21.7 (2.2)	20.0 (1.5)
Pro	26.4 (2.3)	29.8 (4.2)	46.2 (14.5)	48.0 (15.0)
His_2	12.5 (0.9)	14.7 (0.9)	12.1 (0.0)	12.4 (0.0)

[a] Areas in Å^2. Residue labels in Scheme 1.

to the topography of cavity entrances, most of the residues in VClPO lie under the outer surface (Figure 4A), whereas in VBrPO and especially VIPO, the outermost part of the cavity merges with the external protein surface (Figure 4C and D). Note finally that, as far as we chose the activated peroxo form of vanadate to represent the halide binding stage in the catalytic cycle and this corresponds to a positive charge of the cofactor (see next subsection), the peroxo form of VClPO shows a strongly positive electrostatic potential at the active site which covers nearly all the pocket surface (Figure 4B).

Vanadate Cofactor Throughout the Catalytic Cycle. The protonation state of oxygen atoms of the vanadate cofactor is a key issue in elucidating the catalytic activity of vanadium haloperoxidases. Proposals on their mechanism derived from crystal structures[21−26] have been recently complemented by computational studies,[27−31,58,59] which has led to a better understanding of the catalytic properties of chlorine and bromine peroxidases. By using quantum calculations as explained in the Methods, we address now the theoretical analysis of the catalytic cycle of iodoperoxidase focusing on hitherto less explored details regarding the binding of iodide and the final release of hypoiodous acid.

X-ray structures show for vanadate equatorial O−V bond lengths ranging from 1.60 to 1.64 Å in VClPO and between 1.52 and 1.60 Å in VBrPO. According to the evidence provided by previous quantum calculations[27−31] and in agreement with our results (Figure 5), these bond lengths may suggest that equatorial oxygen atoms are deprotonated in resting states of VHPOs. One should recall that these oxygens are stabilized by hydrogen bonds in the active site that anchor the cofactor at the protein (Scheme 1). As noted above, more noteworthy differences are found for apical oxygens even between both resting forms of VClPO, with O−V bond lengths of 1.88 and 2.15 Å for 1IDQ and 1VNI structures, respectively. According to our calculations and in agreement with other reports,[27−31,58,59] such bond lengths may be ascribed to singly and doubly (as a water molecule)

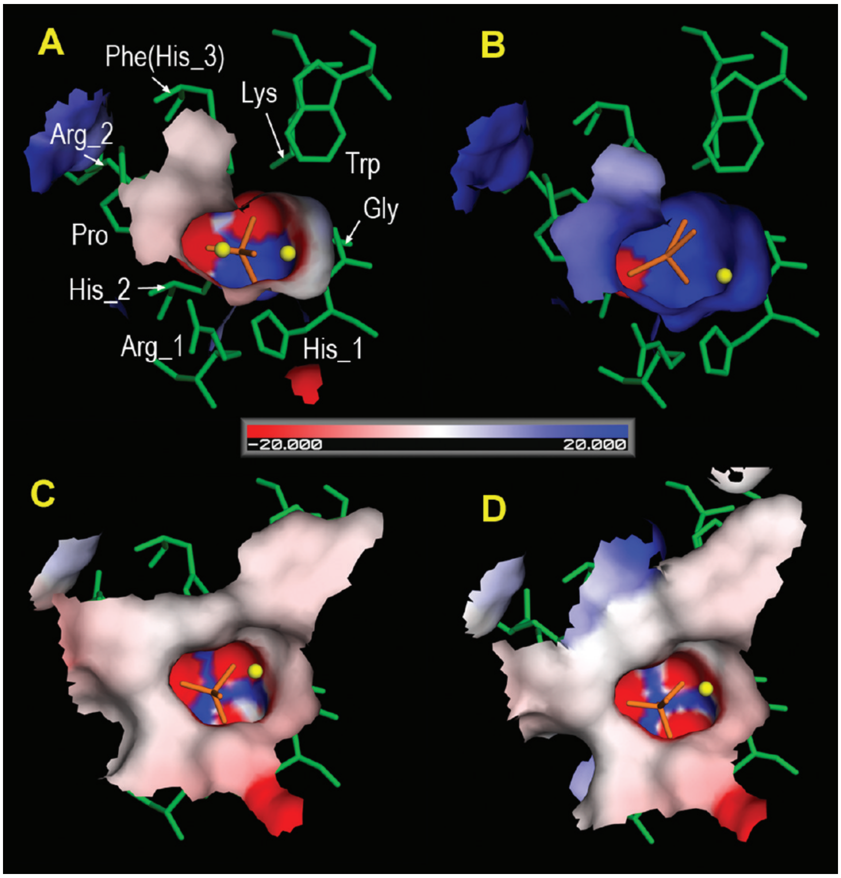


Figure 4. Poisson−Boltzmann electrostatic potential in the range −20 to +20 (units of *kT*/*e*) mapped onto the molecular surface at the active site pocket. Surface clippings at 20 Å depth from the outer protein surface rendered at the same orientation. Residues drawn as green sticks, vanadate group as orange sticks, and water molecules as yellow spheres. Residue numbering in Scheme 1. (A) VClPO (resting form, 1IDQ). (B) VClPO (peroxo form, 1IDU). (C) VBrPO (1QI9). (D) VIPO.

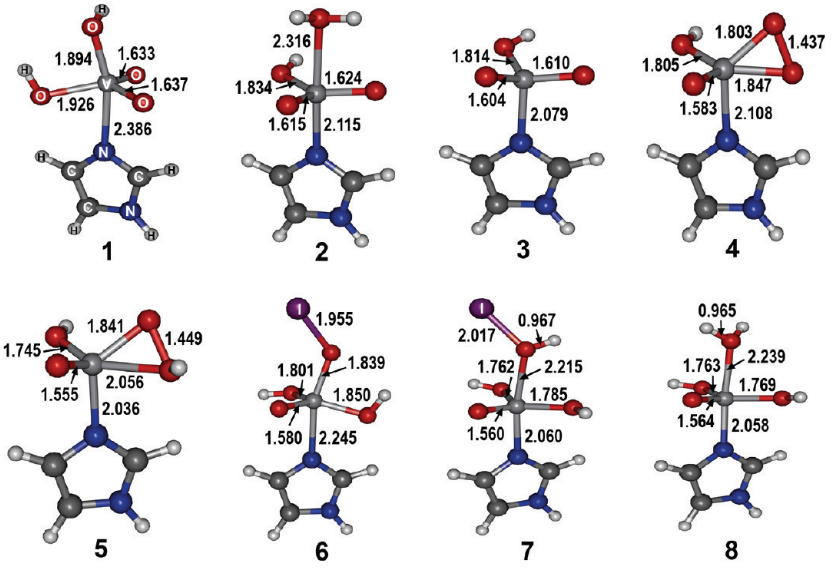
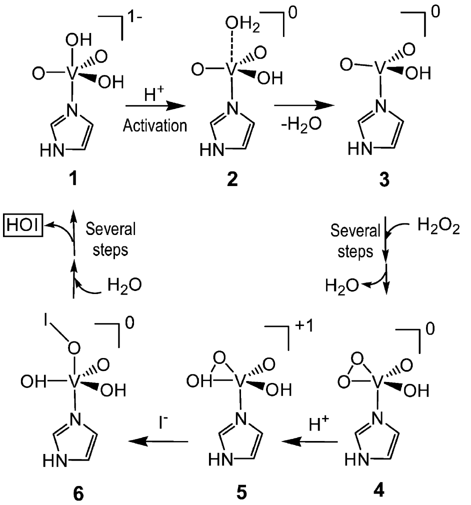


Figure 5. Optimized structures of model complexes involved in the catalytic cycle of VIPO. Distances in Ångstroms.

protonated oxygen atoms, respectively (Figure 5). The apical O−V distance for VBrPO, 1.77 Å, suggests the presence of a hydroxyl group. Apical oxygen is also stabilized by hydrogen bonds with His_1 and water molecules (Figure 1A, Scheme 1).

Recent experimental and theoretical studies have shown that the vanadate anion in resting states of VHPOs can be described as an equilibrium between a structure with hydroxyl groups in equatorial and apical positions, **1** (structure numbers and bond lengths refer hereafter to Scheme 2 and Figure 5), and another structure with an apical water molecule and three equatorial oxo bonds.[27,58−60] A better description of observed ^{51}V NMR chemical shifts and UV−vis spectra of VClPO is indeed provided when the

Scheme 2. Catalytic Cycle for Vanadium Iodoperoxidase



cofactor has apical water.[31] However, it must be noted that our calculations were unable to find optimized geometries of the isolated cofactor with unprotonated equatorial oxygens. This result, observed before in calculations with different functionals,[28] and the fact that X-ray distances seem to discard the presence of hydroxyl in the equatorial position, bring out the need of hydrogen bonds at those positions to stabilize the structure. It has been suggested that, in accordance with earlier kinetic studies on synthetic models that pointed out the role of protonation to render more labile oxo/hydroxo ligands of vanadium complexes,[60] the first stage of the catalytic cycle must be protonation of the anionic resting form,[27,30] a task presumably accomplished by His_1.[17,18] In agreement with this suggestion, we found that the protonation of **1** forms a neutral complex with apical water and equatorial hydroxyl, **2**. Note that, in the process **1** → **2**, the V−N bond length shortens and the opposite happens for the apical O−V value, indicating that this bond weakens in the resting state as a preceding step to the subsequent release of water.

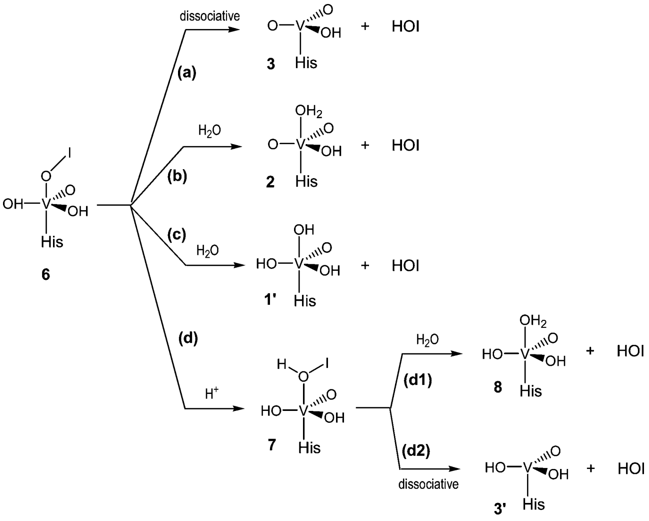
In the study of Zampella et al. on the catalytic activity of VBrPO with DFT calculations performed on a model isolated imidazole−vanadate complex,[30] **2** was assumed as initial structure to obtain the peroxo state of the cofactor. Our calculations (carried out following a different methodology) agree with most of their suggestions and add some supplementary details to the catalytic processes illustrated in Scheme 2 for the VIPO case. In brief, our results predict that the release of the apical water molecule from **2** with the formation of **3** is exoergonic: $\Delta G = -4.1$ in the gas phase and ∼−10 kcal mol^{-1} when solvent is incorporated (Table 2). Species **3** is a tetrahedral intermediate with only three oxygen atoms and a short V−N bond length. It is accepted that, in the next catalytic step, the incoming hydrogen peroxide displaces the apical water and one equatorial oxygen, leading to the neutral peroxo form **4**.[17,18,27,29,30] Peroxide is thus coordinated in a side-on manner in the equatorial plane, distorting the vanadium cofactor and producing a square-based pyramidal oxo−peroxo−vanadium(v) intermediate in the peroxo form of VClPO.[21] It is interesting to mention that our quantum geometry for the isolated intermediate, **4**, is rather similar to that of the cofactor in the X-ray structure (1IDU) of the peroxo form of VClPO.[21]

Table 2. Standard Free Energies at 298 K, $\Delta_R G°$, for Processes Involved in the Catalytic Cycle of VIPO in the Gas Phase ($\varepsilon = 1$) and Obtained in PCM Calculations with $\varepsilon = 4$, 40, and 78.39[a]

process[b]	$\Delta_R G°$ ($\varepsilon = 1$)	$\Delta_R G°$ ($\varepsilon = 4$)	$\Delta_R G°$ ($\varepsilon = 40$)	$\Delta_R G°$ ($\varepsilon = 78.39$)
2 → **3** + H_2O	−4.11	−8.41	−10.5	−10.6
5 + I^- → **6**	−146	−62.6	−37.1	−35.8
(a) **6** → **3** + HOI	2.05	−7.25	−11.4	−11.7
(b) **6** + H_2O → **2** + HOI	6.16	1.16	−0.96	−1.09
(c) **6** + H_2O → **1′** + HOI	1.34	−0.39	−1.42	−1.49
(d1) **7** + H_2O → **8** + HOI	−3.98	−5.28	−5.26	−5.25
(d2) **7** → **3′** + HOI	−2.41	−8.21	−10.2	−10.3

[a] B3LYP/cc-pVTZ calculations. Values in kcal mol^{-1}. [b] Labels refer to Schemes 2 and 3.

Scheme 3. Pathways to the Release of Hypoiodous Acid from Axially Coordinated Hypoiodite Intermediate **6**



This peroxo crystal structure shows that His_1 is no longer hydrogen-bonded while Lys makes direct contact with one peroxo oxygen atom (see Scheme 1). Experimental data[26,60] and theoretical studies[27,29] suggest that protonation of the peroxo moiety is a crucial factor in the activation of peroxo−vanadium complexes and further reaction with halide. Our calculations predict that protonation of **4** should in fact occur at the equatorial oxygen atom of the peroxo group, leading thus to structure **5**. It must be stressed that just this oxygen is hydrogen-bonded to Lys in the X-ray structure, which suggests that this amino acid might play a role in activating the bound peroxide. It is accepted that this protonation is an essential feature of the catalytic mechanism as far as it would increase the potential of the oxo−peroxo−vanadium(v) intermediate for further halide oxidation.[18] Replacement of Lys353 in VClPO with alanine resulted in a considerably reduced catalytic efficiency, which led to the suggestion that the positively charged lysine should polarize the bound peroxide and hence favor nucleophilic attack by the substrate.[61]

The oxo−peroxo species **5** then oxidizes halide, the second substrate of the cycle, by two electrons forming thus an oxidized halogen that is formally at the X^+ oxidation state. However, detection or isolation of the oxidized halogen intermediate is hampered due to its further reaction either with organic substrates (see below) or with a second H_2O_2 molecule forming singlet O_2.[10,11,15,17,18] Therefore, details on completion of the catalytic cycle which entails releasing of the species "X^+" and recovering of the initial resting form must rely exclusively on theoretical work. Previous DFT investigation on bromide[29,30] and our present computational study on iodide show that halide binding involves the remaining unprotonated peroxo oxygen atom. Our calculations predict that the reaction of iodide with **5** to form **6** is strongly exoergonic (-36 kcal mol^{-1} in aqueous phase: Table 2). Geometry data for the hypoiodite adduct **6** are in excellent agreement with the equivalent hypobromite structure obtained by Zampella et al.[30] except the axial O−X bond length, obviously longer for I (1.96 Å) than for Br (1.86 Å in ref 30).

To close the catalytic cycle of VIPO, the axially coordinated hypoiodite must be replaced by a water molecule (**6** → **1** in Scheme 2), a process that can in principle occur according to distinct reaction channels illustrated in Scheme 3. We calculated ΔG under different environment effects (Table 2) for these reactions, all of them involving release of HOI. A dissociative pathway (a) with the formation of **3** is only slightly endoergonic in the gas phase but becomes more favored as increasingly polar media are introduced. This result is similar to that obtained by Zampella et al.[30] for the analogous process with bromine, and also in agreement with this report, we were unable to find low-barrier transition states linking structures **6** and **3**. Direct reactions of **6** with a water molecule to recover resting forms, either **2** in pathway b or an alternative structure **1′** in pathway c, are predicted to be favored very little by our calculations. In agreement again with the bromine case,[30] no low-barrier transition states for these reactions have been found either for the iodine system. We obtained instead stable **6**−water complexes (stabilization energies greater than 30 kcal mol^{-1}) with very strong hydrogen bonds between oxygens of water and vanadate. A similar result was found when a water molecule was introduced in the case of the dissociative pathway mentioned above.

A different possibility not considered before might be protonation of **6** in a similar way to the activation of **1** and **4**. A computational exploration of distinct possible products in pathway d yielded compound **7** as the most stable structure arising from protonating apical oxygen bonded to iodine. Final release of hypoiodous acid from **7** is predicted to be favored either by water displacing HOI, pathway d1, or by following the dissociative pathway d2, both processes becoming more exoergonic upon inclusion of environmental effects, although with little variation with the dielectric constant. Pathway d1 was found to be a barrierless process so that replacement of HOI by water would also be kinetically favored. Because adding a water molecule to **3′** yields **8**, deprotonations of this product allow recovery of initial vanadate structures, either **1** or **2**, in the resting state of the enzyme. We propose that the protonation of **6** and further deprotonation of **8** to close the catalytic cycle of VIPO might be assisted by His476 (His_3 in Scheme 1). This proposal should be supported by the location of the imidazole ring of this histidine with respect to the apical position of structure **6** or **7** (Figure 6). After inserting **7** in the active site of VIPO, the hydrogen bond network of VHPOs (Scheme 1) is essentially kept, with both hydrogen atoms in the equatorial position in the quantum geometries of vanadate forming O2−Nb (Gly 482) and O1−Nη2 (Arg 549) hydrogen bonds (Figure 6A). In addition, we found that apical HOI is properly located to interact with His476. Moreover, exploratory calculations showed that internal rotation of the HOI moiety around the apical O−V bond is essentially barrierless (B3LYP/cc-pVTZ energies differ by a few tenths of a kilocalorie per mole), which suggests that the conformation of hypoiodous acid in the vanadate cofactor must be

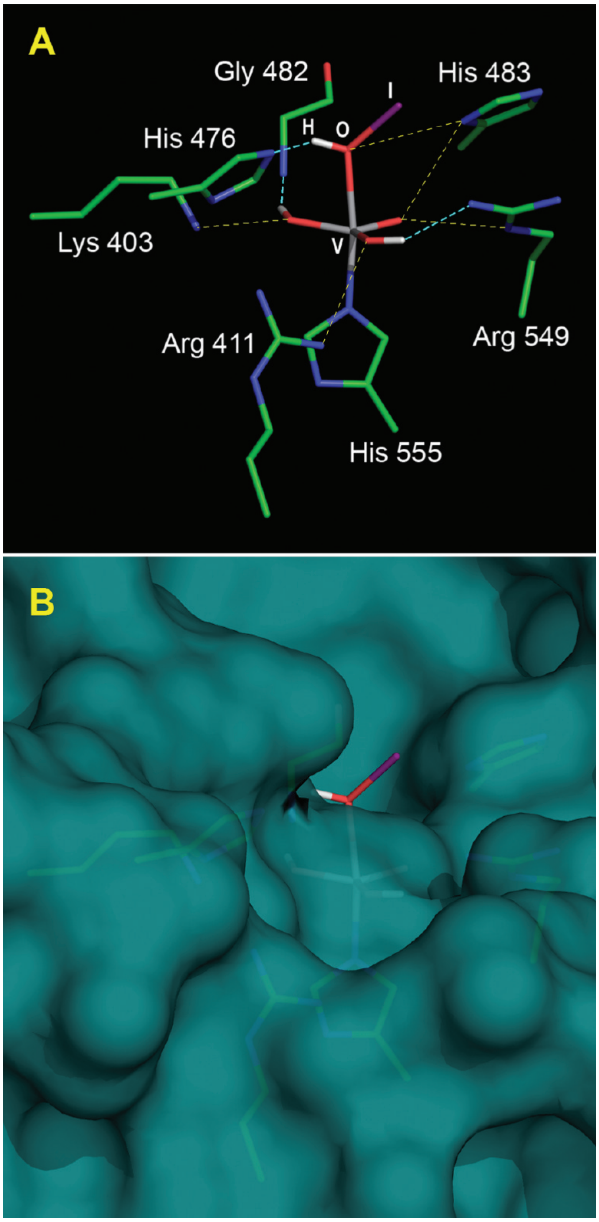


Figure 6. (A) Residues in the active site of VIPO interacting with HOI−vanadate complex **7**. Yellow dashed lines represent hydrogen bonds between non-hydrogen atoms. Cyan dashed lines represent hydrogen bonds involving hydrogen atoms in the quantum optimized geometry of **7**. (B) Close-up view of VIPO molecular surface at the entrance to the substrate channel rendered at the same orientation in A.

determined by hindrance effects on the iodine atom arising from residues in the active site. In fact, the protein surface shows that the HOI group is positioned at the outer region of the cavity entrance with the bulky iodine atom positioned at the greater external part of the cleft (Figure 6B). Replacement by water and further release of hypoidous acid suggested by pathway d in Scheme 3 should be thus facilitated.

Iodination Reactions with Organic Substrates. If an organic substrate is present, it will react with the species "X^+" generated in the mechanism producing a halogenated compound. Experimental studies indicated that hypohalous acid formed in the catalytic cycle of VHPOs acts as a halogenating agent such as HOX[62] or, in the case of bromide and iodide, as X_3^-.[32,63] Reports on VBrPO activity suggest that the nature of the "Br^+" intermediate either as enzyme-bound or as a freely diffusible species apparently depends on the type of organic substrate.[15,62,64] Although there is no direct evidence on the migration of HOBr from the VBrPO active site, a new fluorescence microscopy-based method will permit the monitoring of that process, shedding light on where halogenation of organic substrate actually occurs.[65] The apparent lack of organic substrate specificity in VBrPO was initially interpreted to mean that the enzyme produces a diffusible "Br^+" species that could then carry out bromination reactions.[15,18,34,35] However, recent kinetics studies suggest that the active site channel would hold organic compounds (terpenes and lactones, mainly) in place.[15,18] The role of the site channel in substrate selectivity and the reason why some halogenated products predominate in one alga but not in another related alga with similar VHPO are still open questions.

With regard to iodoperoxidase, no studies on the iodination of organic substrates have been carried out yet. Recent competition experiments indicate that chlorine does not bind to *L. digitata* VIPO, whereas bromide does but in a nonproductive manner.[16] Given the close evolutionary relationship between VIPO and VBrPO, the specificity for iodide of the former led to conjecture that the ancestral VHPO enzyme in brown algae would have been a BrPO and that the loss of bromination in favor of iodination capability had given rise to a novel biochemical function.[16] This agrees with the recently proposed role of VIPO concerning the production of iodocarbons in kelps. On the one hand, these compounds have known microbial toxicity;[11,19] on the other, iodide accumulation acts as a potent inorganic antioxidant, and I^- incorporated in iodocarbons would be easily regenerated by nucleophilic substitution with Cl^-, Br^-, or HO^-.[11] Elucidating details on VIPO-catalyzed production of iodocarbons still needs much experimental work; hence, even conjecturing about that issue is obviously far beyond the scope of this work. We focus here on computing free energies of iodination reactions of common organic molecules by HOI, I_2, and I_3^-. We aim merely to provide the first thermodynamical information regarding the "I^+" iodinating species. It is evident that solvent effects and species in solution different from those dealt with here had to be considered in further studies on iodination reactions associated with the catalytic activity of VIPO. We are currently undertaking this work, which will be the subject of forthcoming papers.

Table 3. Properties Computed with the aug-cc-pVTZ Basis Set (AREP-Optimized Set for Iodine Atom) for Iodine Species[a]

property	MP2[b]	CCSD(T)[b]	theoretical	experimental
I_2				
r_e	2.662	2.690	2.683[c]	2.666[d]
ω_e	226.2	211.6	215.8[c]	214.5[d]
D_e	1.91	1.80	1.13[c]	1.54[d]
I^-				
E.A.	3.21	3.15	2.74[e]	3.06[f]
$I_2 + I^- \rightarrow I_3^-$ reaction				
$\Delta_R H^\circ_0$	−134.5	−124.6		−126 ± 6[g]
I_3^-				
$\Delta_f H^\circ_0$	−257.0	−247.1		−248.5[g]
$\Delta_f H^\circ_{298}$	−253.8	−243.9	171.3[e]	−252 ± 6[g]
$\Delta_f G^\circ_{298}$	−267.8	−257.9		−266[g]
r_e	2.936	2.967	3.002[h]	
ω_e	58.74 (Π_u)	56.35	54.06,[e] 58.20[h]	
	117.7 (Σ_g)	111.2	106.5,[e] 107.3[h]	112[i]
	147.8 (Σ_u)	136.7	145.1,[e] 139.6[h]	139.6[i]

[a] r_e: bond length. ω_e: harmonic vibrational frequencies. D_e: dissociation energy. E.A.: electron affinity. $\Delta_R H^\circ_0$: standard enthalpy of reaction at 0 K. $\Delta_f H^\circ_T$: standard enthalpy of formation at T K. $\Delta_f G^\circ_{298}$: standard Gibbs free energy of formation at 298 K. r_e in eV, ω_e in cm^{-1}, D_e and E.A. in eV, and ΔH° and ΔG° in kJ mol^{-1}. [b] This work. [c] CCSD(T)/cc-pVTZ valence-only ECP/CPP calculations.[51] [d] Reference 66. [e] PW91/cc-pVTZ valence-only ECP calculations.[67] [f] Reference 54. [g] Reference 68. [h] QCISD(T)/TZ all-electron calculations.[69] [i] Reference 69.

If R−H represents an organic compound, the iodination reactions studied are the following:

$$HOI + R{-}H \rightarrow R{-}I + H_2O \qquad (R1)$$

$$I_2 + R{-}H \rightarrow R{-}I + HI \qquad (R2)$$

$$I_3^- + R{-}H \rightarrow R{-}I + HI + I^- \qquad (R3)$$

Before discussing $\Delta_R G^0_{298}$ values obtained as explained in the Methods, Table 3 presents a reliability test of our methodology (that concerns especially performance of the AREP/cc-pVTZ basis set for iodine) by displaying some molecular properties for iodine species related with I_3^-. Gibbs free energies of iodination reactions are listed in Table 4. CH_3I, CH_2I_2, and iodopropene were chosen to represent volatile iodocarbons while CH_2ICHO, CH_2ICOOH, and iodophenol were selected to represent nonvolatile compounds. Both methane derivatives are the most abundant alkyl iodides over oceans, and their release by phytoplankton and algae has been long known.[1] Propene and acetaldehyde iodinated derivatives were chosen, as it was found that they are major factors for regulating reactive halogen chemistry in the MBL.[70] Iodoacetic acid was proposed as one of the possible organic compounds identified in a recent experimental study on iodine speciation in rain and aerosols.[71] Finally, phenol is the side chain of tyrosine, an amino acid that plays a crucial role in biological processing of iodine in all living organisms.

$\Delta_R G^0_{298}$ values in Table 4 agree in predicting a markedly exoergonic iodination reaction with HOI in contrast to reactions with I_2 and I_3^-, which are strongly endoergonic.

Table 4. Standard Free Energies at 298 K, $\Delta_R G°$, for Iodination Reactions R1, R2, and R3[a]

		$\Delta_R G°$ MP2			$\Delta_R G°$ CCSD(T)		
R−H	R−I[b]	R1	R2	R3	R1	R2	R3
CH_4	CH_3I	−88.18	+47.95	+91.89	−84.34	+50.24	+100.5
CH_4	CH_2I_2	−31.26	+104.9	+148.8	−17.16	+117.4	+167.7
$CH_2{=}CH{-}CH_3$	$CH_2{=}CI{-}CH_3$	−97.64	+38.50	+82.43	−129.3	+5.29	+55.55
$CH_2{=}CH{-}CH_3$	*Z*- $CHI{=}CH{-}CH_3$	−93.36	+42.77	+86.62	−126.7	+7.88	+58.14
$CH_2{=}CH{-}CH_3$	*E*-$CHI{=}CH{-}CH_3$	−89.85	+46.28	+90.21	−123.7	+10.85	+61.11
$CH_2{=}CH{-}CH_3$	(*g*)- $CH_2{=}CH{-}CH_2I$	−89.06	+47.07	+91.00	−124.5	+10.11	+60.37
$CH_2{=}CH{-}CH_3$	(*e*)- $CH_2{=}CH{-}CH_2I$	−83.88	+52.25	+96.18	−118.9	+15.67	+65.93
$CH_3{-}CHO$	(*g*)- $CH_2I{-}CHO$	−55.59	+80.54	+124.5	−117.1	+17.53	+67.79
$CH_3{-}CHO$	(*e*)- $CH_2I{-}CHO$	−49.47	+86.66	+130.6	−110.9	+23.71	+73.97
$CH_3{-}COOH$	(*g*)- $CH_2I{-}COOH$	−16.44	+119.7	+163.6	−117.0	+17.59	+67.85
$CH_3{-}COOH$	(*e*)- $CH_2I{-}COOH$	−14.68	+121.5	+165.4	−115.1	+19.47	+69.73
PhOH	*o-trans*-IPhOH	−73.92	+62.21	+106.1	−77.40	+57.18	+107.4
PhOH	*m-trans*-IPhOH	−76.35	+59.78	+94.81	−79.81	+54.77	+105.0
PhOH	*p*-IPhOH	−75.66	+60.48	+104.4	−79.91	+54.68	+104.9
PhOH	*m-cis*-IPhOH	−76.73	+59.40	+103.3	−80.15	+54.43	+104.9
PhOH	*o-cis*-IPhOH	−85.52	+50.61	+94.54	−88.68	+45.90	+96.16

[a] Values in kJ mol^{-1}. [b] (*g*) and (*e*) indicate *gauche* and *eclipse* conformations, respectively. For $CH_2I{-}CHO$ and $CH_2I{-}COOH$ conformations, refer to the relative position of iodine atom and carbonyl oxygen.

Significant differences between MP2 and CCSD(T) results are noticed in some cases, especially for acetic acid and acetaldehyde reactions due to the rather distinct $\Delta_f H^0{}_{298}$ values obtained for these molecules (Table S2 in the Supporting Information). This notwithstanding, MP2 and CCSD(T) results display similar trends for differences between reactions R1−R3. If one considers the magnitude of free energies in Table 4, the thermodynamical efficiency of HOI compared with the other iodinating reagents seems beyond question. Note besides that both methods also agree in predicting the small differences in favor of more stable isomers: ~3 kJ mol^{-1} for the *Z* isomer of $CHI{=}CH{-}CH_3$, ~5 kJ mol^{-1} for the *gauche* conformations of both $CH_2{=}CH{-}CH_2I$ and CH_2ICHO, and ~2 kJ mol^{-1} for the *gauche* conformation of CH_2ICOOH. MP2 and CCSD(T) energies agree again in distinguishing quantitatively the stability of iodophenol isomers: *o-trans-* is the less favored and *o-cis-* the more favored with a significant difference greater than 8 kJ mol^{-1} with respect to the remaining isomers, a result that may be temptatively explained by the stabilizing effect associated with an intramolecular I···H−O hydrogen bond in this isomer.

Conclusions

The active site of the structure modeled for VIPO superposes well with active sites of available crystal structures of chlorine and bromine VHPOs. The hydrogen bond network settled by residues defining the scaffold that anchors the vanadate cofactor is conserved in VIPO, except the position of Ser in VClPO and VBrPO that is occupied by Ala in VIPO. Differences noticed in the spatial location of conserved Trp and Pro along with substitution of His for Phe (also present in VBrPO) are other distinctive structural features of the active site in VIPO.

Electrostatic potential at vanadium-binding site cavities of VHPOs is much stronger than at any other region of protein surfaces due to the presence of four electronegative oxygens of vanadate and four basic residues. However, small differences in the local electrostatic potential at the site pockets reveal a negative domain in VIPO at a region where VClPO and VBrPO have positive or neutral potentials. This region corresponds to the position of Ala in VIPO that substitutes Ser in the other VHPOs. Hence, we propose that this mutation might be associated with small changes of electrostatic potential related with binding iodine instead of the more electronegative bromine or chlorine.

Significant differences in the local topography of cavity entrances are found. VClPO has a deep entrance with nearly buried large regions, whereas VIPO exhibits a great part of the surface around the cavity entrance accessible to the solvent. These differences affect especially hydrophobic patches corresponding to conserved Trp and Pro and the surface region of His476 that is largely exposed in VIPO, whereas it is nearly buried in VClPO. The wide open cleft in VIPO opposite the small deep channel in VClPO suggests that the topography of the access cavity leading to the active site might be a relevant factor regarding halide selectivity in these enzymes.

Quantum calculations were performed to obtain structures and energies of imidazole−vanadate complexes intended to model the cofactor throughout the catalytic cycle of VIPO. Stages corresponding to activation of the initial resting form, subsequent displacement of apical water and binding of hydrogen peroxide, protonation at the equatorial position to activate the peroxo form, and further oxidization of iodide with the binding of I at an apical position parallel those reported before for VBrPO with a different methodology. To close the catalytic cycle, a water molecule replaces axially coordinated hypoiodite recovering the initial resting form. We suggest a pathway consisting of protonation of apical oxygen bonded to iodine and further displacement of HOI by water either at a dissociative reaction or by direct replacement, both exoergonic processes. Inserting these intermediates at the VIPO structure indicates that His476 might assist this protonation, as it is properly located to interact with the apical OI group. Internal rotation of the resulting HOI around the apical O−V bond is found to be essentially free so that the HOI conformation should be determined only by hindrance effects on the active site. The HOI group is located at the outer region of the cavity entrance with the bulky iodine atom positioned at a great external cleft region which facilitates release of hypoiodous acid.

Gibbs free energies of iodination reactions of common organic molecules by HOI, I_2, and I_3^-, the three iodinating "I^+" reagents released by the catalytic activity of VIPO, were obtained in *ab initio* MP2 and CCSD(T) calculations. Upon selecting organic substrates to represent both volatile and nonvolatile iodocarbons, our calculations show that iodination by HOI is greatly favored as it gives strongly exoergonic reactions contrarily to what happens with I_2 or I_3^-.

Acknowledgment. This research has been carried out with funding from the Spanish Ministry of Education, Project FIS2007-61686, Ministry of Science, Project BIO2009-07050, and Comunidad de Madrid, Grant EIADES S-0505/AMB/0296. O.G. acknowledges financial support from CSIC, "JAE-Doc" program.

Supporting Information Available: Ribbon diagram of VIPO protein. Modeled structure of VIPO protein in PDB format. AREP-optimized cc-pVTZ and aug-cc-pVTZ basis sets for iodine atom. $\Delta_f H^0{}_0$, $\Delta_f H^0{}_{298}$, $H^0{}_{298}-H^0{}_0$, and $S^0{}_{298}$ values for species involved in reactions R1−R3. MP2 geometries and MP2 and CCSD(T) absolute energies calculated with aug-cc-pVTZ basis sets for species involved in reactions R1−R3. B3LYP/cc-pVTZ geometries and absolute energies of complexes **1**−**8** plus **1′** and **3′**. B3LYP/cc-pVTZ values of Gibbs free energies and energy corrections obtained in PCM calculations with $\epsilon = 4$, 40, and 80 for all the species involved in the catalytic cycle of VIPO (Schemes 2 and 3). This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) (a) Chameides, W. L.; Davis, D. D. *J. Geophys. Res.* **1980**, *85*, 7383–7393. (b) Solomon, S.; Garcia, R. R.; Ravishankara, A. R. *J. Geophys. Res.* **1994**, *99* (D10), 20491–20500. (c) Vogt, R.; Sander, R.; von Glasow, R.; Crutzen, P. J. *J. Atmos. Chem.* **1999**, *32*, 375–395. (d) McFiggans, G.; Plane, J. M. C.; Allan, B. J.; Carpenter, L. J.; Coe, H.; O'Dowd, C. D. *J. Geophys. Res. Atmos.* **2000**, *105*, 14371–14385. (e) Calvert, J. G.; Lindberg, S. E. *Atmos. Environ.* **2004**, *38*, 5105–5116.

(2) (a) Saiz-Lopez, A.; Plane, J. M. C. *Geophys. Res. Lett.* **2004**, *31*, L04112. (b) Saiz-Lopez, A.; Plane, J. M. C.; McFiggans, G.; Williams, P. I.; Ball, S. M.; Bitter, M.; Jones, R. L.; Hongwei, C.; Hoffmann, T. *Atmos. Chem. Phys. Discuss.* **2005**, *5*, 5405–5439. (c) Saiz-Lopez, A.; Plane, J. M. C.; McFiggans, G.; Williams, P. I.; Ball, S. M.; Bitter, M.; Jones, R. L.; Hongwei, C.; Hoffmann, T. *Atmos. Chem. Phys.* **2006**, *6*, 883–895.

(3) Saiz-Lopez, A.; Boxe, C. S. *Atmos. Chem. Phys. Discuss.* **2008**, *8*, 2953–2976.

(4) (a) Hoffmann, T.; O'Dowd, C. D.; Seinfeld, J. H. *Geophys. Res. Lett.* **2001**, *28*, 1949–1952. (b) O'Dowd, C. D.; Jimenez, J. L.; Bahreini, R.; Flagan, R. C.; Seinfeld, J. H.; Hameri, K.; Pirjola, L.; Kulmala, M.; Jennings, S. G.; Hoffmann, T. *Nature* **2002**, *417*, 632–636. (c) Jimenez, J. L.; Bahreini, R.; Cocker, D. R.; Zhuang, H.; Varutbangkul, V.; Flagan, R. C.; Seinfeld, J. H.; O'Dowd, C. D.; Hoffmann, T. *J. Geophys. Res.* **2003**, *108* (D10), 4318. (d) Burkholder, J. B.; Curtius, J.; Ravishankara, A. R.; Lovejoy, E. R. *Atmos. Chem. Phys.* **2004**, *4*, 19–34. (e) Saunders, R. W.; Plane, J. M. C. *Environ. Chem.* **2005**, *2*, 299–303.

(5) (a) Saiz-Lopez, A.; Chance, K.; Liu, X.; Kurosu, T. P.; Sander, S. P. *Geophys. Res. Lett.* **2007**, *34*, L12812. (b) Saiz-Lopez, A.; Mahajan, A. S.; Salmon, R. A.; Bauguitte, J. B.; Jones, A. E.; Roscoe, H. K.; Plane, J. M. C. *Science* **2007**, *317*, 348–351. (c) Schönhardt, A.; Richter, A.; Wittrock, F.; Kirk, H.; Oetjen, H.; Roscoe, H. K.; Burrows, J. P. *Atmos. Chem. Phys. Discuss.* **2007**, *7*, 12959–12999.

(6) Whitehead, D. C. *Environ. Int.* **1984**, *10*, 321–339.

(7) (a) Moore, R. M.; Groszko, W. *J. Geophys. Res.* **1999**, *104*, 11163–11171. (b) Alicke, B.; Hebestreit, K.; Platt, U. *Nature* **1999**, *397*, 572–573. (c) Allan, B. J.; Plane, J. M. C.; McFiggans, G. *Geophys. Res. Lett.* **2001**, *28*, 1945–1948. (d) Whalley, L. K.; Furneaux, K. L.; Gravestock, T.; Atkinson, H. M.; Bale, C. S. E.; Ingham, T.; Bloss, W. J.; Heard, D. E. *J. Atmos. Chem.* **2007**, *58*, 19–39. (e) Wada, R.; Beames, J. E.; Orr-Ewing, A. J. *J. Atmos. Chem.* **2007**, *58*, 69–87.

(8) (a) McFiggans, G.; Coe, H.; Burgess, R.; Allan, B. J.; Cubison, M.; Alfarra, M. R.; Saunders, R.; Saiz-Lopez, A.; Plane, J. M. C.; Wevill, D. J.; Carpenter, L. J.; Rickard, A. R.; Monks, P. S. *Atmos. Chem. Phys.* **2004**, *4*, 701–713. (b) Carpenter, L. J.; Liss, P. S.; Penkett, S. A. *J. Geophys. Res.-Atmos.* **2003**, *108*, 4256.

(9) Carpenter, L. J. *Chem. Rev.* **2003**, *103*, 4953–4962.

(10) Leblanc, C.; Colin, C.; Cosse, A.; Delage La Barre, S.; Morin, P.; Fievet, B.; Voiseux, C.; Ambroise, Y.; Verhaeghe, E.; Amouroux, D.; Donard, O.; Tessier, E.; Potin, P. *Biochimie* **2006**, *88*, 1773–1785.

(11) Küpper, F. C.; Carpenter, L. J.; McFiggans, G. B.; Palmer, C. J.; Waite, T. J.; Boneberg, E. M.; Woitsch, S.; Weiller, M.; Abela, R.; Grolimund, D.; Potin, P.; Butler, A.; Luther III, G. W.; Kroneck, P. M. H.; Meyer-Klaucke, W.; Feiters, M. C. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 6954–6958.

(12) Küpper, F. C.; Schweigert, N.; Gall, E. A.; Legendre, J. M.; Vilter, H.; Kloareg, B. *Planta* **1998**, *207*, 163–171.

(13) Palmer, C. J.; Anders, T. L.; Carpenter, L. J.; Küpper, F. C.; McFiggans, G. *Environ. Chem.* **2005**, *2*, 282–290.

(14) Kylin, H. *Z. Physiol. Chem.* **1929**, *186*, 50–84.

(15) Butler, A.; Carter-Franklin, N. *Nat. Prod. Rep.* **2004**, *21*, 180–188.

(16) Colin, C.; Leblanc, C.; Michel, G.; Wagner, E.; Leize-Wagner, E.; Van Dorsselaer, A.; Potin, P. *J. Biol. Inorg. Chem.* **2005**, *10*, 156–166.

(17) Winter, J. M.; Moore, B. S. *J. Biol. Chem.* **2009**, *284*, 18577–18581.

(18) Butler, A.; Sandy, M. *Nature* **2009**, *460*, 848–854.

(19) Manley, S. L. *Biogeochem.* **2002**, *60*, 163–180.

(20) (a) Faulkner, D. J. *Nat. Prod. Rep.* **2002**, *19*, 1–48. (b) Gribble, G. W. *Acc. Chem. Res.* **1998**, *31*, 141–152. (c) Carpenter, L. J.; Liss, P. S. *J. Geophys. Res.* **2002**, *105*, 20539–20547.

(21) (a) Messerschmidt, A.; Wever, R. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 392–396. (b) Messerschmidt, A.; Prade, L.; Wever, R. *Biol. Chem.* **1997**, *378*, 309–315.

(22) Macedo-Ribeiro, S.; Hemrika, W.; Renirie, R.; Wever, R.; Messerschmidt, A. *J. Biol. Inorg. Chem.* **1999**, *4*, 209–219.

(23) Weyand, M.; Hecht, H.; Kiess, M.; Liaud, M.; Vilter, H.; Schomburg, D. *J. Mol. Biol.* **1999**, *293*, 595–611.

(24) Isupov, M. N.; Dalby, A. R.; Brindley, A. A.; Izumi, Y.; Tanabe, T.; Murshudov, G. N.; Littlechild, J. A. *J. Mol. Biol.* **2000**, *299*, 1035–1049.

(25) Carter, J. N.; Beatty, K. E.; Simpson, M. T.; Butler, A. *J. Inorg. Biochem.* **2002**, *91*, 59–69.

(26) (a) Casny, M.; Rehder, D.; Schmidt, H.; Vilter, H.; Conte, V. *J. Inorg. Biochem.* **2000**, *80*, 157–160. (b) Rehder, D. *J. Inorg. Biochem.* **2008**, *102*, 1152–1158.

(27) Kravitz, J. Y.; Pecoraro, V. L.; Carlson, H. A. *J. Chem. Theory Comput.* **2005**, *1*, 1265–1274.

(28) Zampella, G.; Kravitz, J. Y.; Webster, C. E.; Fantucci, P.; Hall, M. B.; Carlson, H. A.; Pecoraro, V. L.; De Gioia, L. *Inorg. Chem.* **2004**, *43*, 4127–4136.

(29) Zampella, G.; Fantucci, P.; Pecoraro, V. L.; De Gioia, L. *J. Am. Chem. Soc.* **2005**, *127*, 953–960.

(30) Zampella, G.; Fantucci, P.; Pecoraro, V. L.; De Gioia, L. *Inorg. Chem.* **2006**, *45*, 7133–7143.

(31) Zhang, Y.; Gascon, J. A. *J. Inorg. Biochem.* **2008**, *102*, 1684–1690.

(32) Butler, A. *Coord. Chem. Rev.* **1999**, *187*, 17–35.

(33) Hasan, Z.; Renirie, R.; Kerkman, R.; Ruijssenaars, H. J.; Hartog, A. F.; Wever, R. *J. Biol. Chem.* **2006**, *281*, 9738–9744.

(34) Itoh, N.; Hasan, A. K.; Izumi, Y.; Yamada, H. *Eur. J. Biochem.* **1988**, *172*, 477–484.

(35) De Boer, E.; Wever, R. *J. Biol. Chem.* **1988**, *236*, 12326–12332.

(36) Peroxibase - The Peroxidase database. http://peroxibase.toulouse.inra.fr/index.php (accessed Jun. 9, 2009).

(37) (a) Swiss-Model. http://swissmodel.expasy.org (accessed Jun. 9, 2009). (b) Arnold, K.; Bordoli, L.; Kopp, J.; Schwede, T. *Bioinform.* **2006**, *22*, 195–201. (c) Kiefer, F.; Arnold, K.; Kunzli, M.; Bordoli, L.; Schwede, T. *Nucl. Acid. Res.* **2009**, *37*, D387–D392.

(38) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004**, *25*, 1605–1612, Chimera, version 1.3; UCSF Chimera. http://www.cgl.ucsf.edu/chimera/ (accessed Sep. 9, 2009).

(39) DeLano, W. L. *PyMOL*, version 1.2r1; DeLano Scientific LLC: Palo Alto, CA, 2009. PyMOL Molecular Viewer. http://www.pymol.org (accessed Oct. 26, 2009).

(40) Pacios, L. F. *Comput. Chem.* **1994**, *18*, 377–385.

(41) Tsodikov, O. V.; Record, M. T.; Sergeev, Y. V. *J. Comput. Chem.* **2002**, *23*, 600–609.

(42) Baker, N. A.; Sept, D.; Holst, J. S.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037–10041.

(43) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(44) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. *Nucl. Acid. Res.* **2004**, *32*, W665–W667.

(45) (a) Davis, M. E.; McCammon, J. A. *Chem. Rev.* **1990**, *94*, 7684–7692. (b) Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144–1149.

(46) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361–36.

(47) Balabanov, N. B.; Peterson, K. A. *J. Chem. Phys.* **2005**, *123*, 064107/1–15.

(48) Pacios, L. F.; Christiansen, P. A. *J. Chem. Phys.* **1985**, *82*, 2664–2671.

(49) LaJohn, L. A.; Christiansen, P. A.; Ross, R. B.; Atashroo, T.; Ermler, W. C. *J. Chem. Phys.* **1987**, *87*, 2812–2824.

(50) (a) Ermler, W. C.; Ross, R. B.; Christiansen, P. A. *Adv. Quantum Chem.* **1988**, *19*, 139–182. (b) Christiansen, P. A. *J. Chem. Phys.* **2000**, *112*, 10070–10074.

(51) Martin, J. M. L.; Sundermann, A. *J. Chem. Phys.* **2001**, *114*, 3408–3420.

(52) Werner, H. J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. *MOLPRO*, version 2006.1; MOLPRO quantum chemistry package. http://www.molpro.net (accessed Sep. 15, 2008).

(53) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

(54) Chase, M. W., Jr. *J. Phys. Chem. Ref. Data* **1998**, *9*, 1–1951.

(55) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *GAUSSIAN03*; Gaussian Inc.: Wallingford, CT, 2004.

(56) Bordoli, L.; Kiefer, F.; Arnold, K.; Benkert, P.; Battery, J.; Schwede, T. *Nature Protocols* **2009**, *4*, 1–14.

(57) Hasan, Z.; Renirie, R.; Kerkman, R.; Ruijssenaars, H. J.; Hartog, A. F.; Wever, R. *J. Biol. Chem.* **2006**, *281*, 9738–9744.

(58) Waller, M. P.; Bühl, M.; Geethalakshmi, K. R.; Wang, D.; Thiel, W. *Chem.—Eur. J.* **2007**, *13*, 4723–4732.

(59) Waller, M. P.; Geethalakshmi, K. R.; Bühl, M. J. *J. Phys. Chem. B* **2008**, *112*, 5813–5823.

(60) Colpas, G. J.; Hamstra, B. J.; Kampf, J. W.; Pecoraro, V. L. *J. Am. Chem. Soc.* **1996**, *118*, 3469–3478.

(61) Hemrika, W.; Renirie, R.; Macedo-Ribeiro, S.; Messerschmidt, A.; Wever, R. *J. Biol. Chem.* **1999**, *274*, 23820–23827.

(62) Martinez, J. S.; Carrol, G. L.; Tschirret-Guth, R. A.; Altenhoff, G.; Little, R. D.; Butler, A. *J. Am. Chem. Soc.* **2001**, *123*, 3289–3294.

(63) Everett, R. R.; Soedjak, H. S.; Butler, A. *J. Biol. Chem.* **1990**, *265*, 15671–15679.

(64) Tschirret-Guth, R. A.; Butler, A. *J. Am. Chem. Soc.* **1994**, *116*, 411–412.

(65) Martinez, V. M.; De Cremer, G.; Roeffaers, M. B.; Sliwa, M.; Baruah, M.; De Vos, D. E.; Hofkens, J.; Sels, B. F. *J. Am. Chem. Soc.* **2008**, *130*, 13192–13193.

(66) Huber, K. P.; Herzberg, G. *Constants of Diatomic Molecules*; Van Nostrand Reinhold: New York, 1979; pp 330−333.

(67) Asaduzzaman, A. M.; Schreckenbach, G. *Theor. Chem. Acc.* **2009**, *122*, 119–125.

(68) Do, K.; Klein, T. P.; Pommerening, C. A.; Sunderlin, L. S. *J. Am. Soc. Mass Spectrom.* **1997**, *8*, 688–696.

(69) Lynden-Bell, R. M.; Kosloff, R.; Ruhman, S.; Danovich, D.; Vala, J. *J. Chem. Phys.* **1998**, *109*, 9928–9937.

(70) Toyota, K.; Kanaya, Y.; Takahashi, M.; Akimoto, H. *Atmos. Chem. Phys.* **2004**, *4*, 1961–1987.

(71) Gilfedder, B. S.; Lai, S. C.; Petri, M.; Biester, H.; Hoffmann, T. *Atmos. Chem. Phys.* **2008**, *8*, 6069–6084.

CT100041X

JCTC Journal of Chemical Theory and Computation

ProMetCS: An Atomistic Force Field for Modeling Protein−Metal Surface Interactions in a Continuum Aqueous Solvent

Daria B. Kokh,*,[†] Stefano Corni,[‡] Peter J. Winn,[§] Martin Hoefling,[||] Kay E. Gottschalk,[||] and Rebecca C. Wade*,[†]

Molecular and Cellular Modeling Group, Heidelberg Institute for Theoretical Studies (HITS gGmbH), Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany, INFM-CNR National Research Center on nanoStructures and BioSystems at Surface (S3), Modena, Italy, Centre for Systems Biology, School of Biosciences, The University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, and Ludwig Maximilians University, Munich, German

Received February 12, 2010

Abstract: In order to study protein−inorganic surface association processes, we have developed a physics-based energy model, the ProMetCS model, which describes protein−surface interactions at the atomistic level while treating the solvent as a continuum. Here, we present an approach to modeling the interaction of a protein with an atomically flat Au(111) surface in an aqueous solvent. Protein−gold interactions are modeled as the sum of van der Waals, weak chemisorption, and electrostatic interactions, as well as the change in free energy due to partial desolvation of the protein and the metal surface upon association. This desolvation energy includes the effects of water−protein, water−surface, and water−water interactions and has been parametrized using molecular dynamics (MD) simulations of water molecules and a test atom at a gold−water interface. The proposed procedure for computing the energy terms is mostly grid-based and is therefore efficient for application to long-time simulations of protein binding processes. The approach was tested for capped amino acid residues whose potentials of mean force for binding to a gold surface were computed and compared with those obtained previously in MD simulations with water treated explicitly. Calculations show good quantitative agreement with the results from MD simulations for all but one amino acid (Trp), as well as correspondence with available experimental data on the adhesion properties of amino acids.

1. Introduction

Protein−surface binding events are of great importance in many bioengineering, biomedical and nanotechnology applications. For example, protein adsorption properties are crucial for the integration of medical implants with tissue, and for the assembly of interfacial protein constructs in nanotechnology, such as sensors, activators, and other functional components at the biological/electronic junction. Over the past decades, extensive experimental investigations on the molecular recognition, binding, and self-assembly of proteins, peptides, and amino acids on inorganic surfaces have been reported (for gold, see refs 1−7), and even combinatorially selected peptides with affinity for specific inorganic materials have been successfully synthesized.[8−10] For some examples of protein adsorption studies, particularly in connection with possible applications, see the reviews in refs 11−14 and references therein.

Because of the high complexity of protein adsorption phenomena and the scarcity of experimental data at the

* Corresponding author e-mail: daria.kokh@h-its.org (D.B.K.) and rebecca.wade@h-its.org (R.C.W.).

[†] Heidelberg Institute for Theoretical Studies.

[‡] INFM-CNR National Research Center.

[§] The University of Birmingham.

[||] Ludwig Maximilians University.

10.1021/ct100086j © 2010 American Chemical Society
Published on Web 04/16/2010

atomistic level, however, the mechanisms by which biomolecules interact with inorganic surfaces are still poorly understood,[14] and until very recently, investigations of protein adsorption properties either had a rather qualitative character or were done on the macroscopic scale. This is why, in recent years, great efforts have been applied to adapt computational methods that are usually employed for molecular modeling in solution to the protein−surface problem. In particular, all-atom empirical force field methods, treating water molecules and the internal coordinates of an adsorbate explicitly, are now widely used to investigate biomolecule−surface binding behavior at the atomistic level[13−24] and have been shown to be able to provide qualitative agreement with experimentally observed adsorption tendencies for some small peptides.[19] However, all-atom molecular modeling methods, with explicit inclusion of water molecules, are extremely computationally demanding and therefore restricted to short time (typically of 10−100 ns) and length scales, while most experimental studies give an averaged behavior of large biomolecules over milliseconds or longer. Most of the atoms in molecular dynamics (MD) simulations with explicit water molecules come from the solvent itself. Furthermore, the presence of explicit water molecules slows protein motions. These two factors can make the computational time needed for the convergence of calculated properties extremely long. Therefore, a possible way to reduce computational time is to use an implicit solvent model that, in combination with an all-atom force field representation of a protein, may provide a reasonable compromise between accuracy and computational cost.

Existing implicit solvent models have primarily been developed for simulation of protein or peptide behavior in solution alone[25] and are generally not appropriate for protein interactions with inorganic interfaces.[13] This was demonstrated, in particular, by Sun and Latour[26] in their comparative analysis of commonly used empirical force-field-based implicit solvent models. It was found that the adsorption free energy of a peptide on a self-assembled monolayer (SAM) may change by up to several tens of kilocalories per mole, depending on which solvent model was used for the calculations. Furthermore, it has been recognized recently that the microscopic properties of the hydration shell vary for different solid surfaces, thereby altering the mechanism of adsorbate−surface interaction. For example, on metal surfaces, the desolvation energy may cause a transition barrier to adsorption due to the energetically unfavorable displacement of the water layer,[24] whereas, for some polar surfaces, peptides may be bound to the structured water layer rather than to the surface itself.[23] Hence, to provide a reliable description of protein−metal association in aqueous solvent, the solvent model should include a microscopic characterization of processes at the protein−surface interface.

In the present paper, we propose an approach for computation of the adsorption free energy of a biomolecule to a gold surface with an implicit solvent model that accounts for the short- and long-range effects of the protein−solvent−metal interactions. We employed the Au(111) surface for modeling because of its importance in the field of protein−surface interactions, both for fundamental studies (well-characterized, stable surface in both air and water) and for potential applications (e.g., contacts in nanobioelectronics and optical detection systems). Moreover, extensive theoretical investigations of small organic molecules adsorbed on gold,[1,7,20,27−29] as well as experimental data on protein and peptide adsorption,[3,5,7] are available and can be used for model optimization and validation. The present energy function is designed for use in Brownian Dynamics (BD) simulations of protein adsorption to surfaces but is not limited to this application.

BD methods in which solute molecules are treated as rigid bodies diffusing in a continuum solvent are commonly applied to simulate diffusion-influenced reactions and have been shown to be successful for computing protein−protein,[30] protein−small molecule,[31] and protein−membrane[32] association kinetics. Similarly, BD methods can be directly applied to a large group of proteins with high internal stability that can adsorb onto inorganic surfaces without appreciable changes in conformation or can form a transient complex before conformational changes occur. This method may also open the way for simulation of protein−protein interactions mediated by solid surfaces or protein self-assembly on inorganic substrates.

Despite the apparent similarity of the protein−protein and protein−solid surface association reactions, they have intrinsic differences in kinetics and in the driving forces for the binding processes. Indeed, the leading interaction in the case of protein−protein association to a diffusional encounter complex often arises from the long-range electrostatic forces, while the short-range effects can be described simply by prohibiting overlap of the exclusion volumes of the proteins. The influence of electrostatics on the interaction between a protein and an uncharged metal surface is much weaker since it arises solely from polarization effects. For a neutral solute molecule without a well-pronounced dipole moment, the image-charge potential must rapidly converge to zero as the distance from the metal increases due to the cancellation of contributions from opposite charges. On the other hand, at small distances from the surface, short-range interactions such as van der Waals forces and small metal−solute molecule charge transfers (that may also involve π electrons), along with the desolvation free energy, dominate over the electrostatic interaction. The construction of such an energy function is facilitated by the fact that the van der Waals interaction between organic molecules and a solid state surface, in particular gold, has recently been parametrized with a set of force field parameters[27] which can be directly implemented at an atomistic level in the energy function. A continuum solvent model able to provide a reliable description of the solvent−protein−solid interface, especially hydrophobic effects, needs to be developed and parametrized. This task is complicated by the fact that there is no well-established microscopic model of the protein−water−metal interface even though the behavior of aqueous solvent itself on the metal (in particular, gold) surfaces has been intensively studied both theoretically[33−40] and experimentally,[33,41−45] and some solution-driven effects in MD simulations of peptide adsorption on metal surfaces have been reported.[19,23,24]

In experimental studies of metal wetting properties, gold surfaces have been described both as hydrophilic (on the basis of contact angle measurements)[43] and rather hydrophobic (from the sublimation kinetics of ice layers).[45] On the other hand, experimental and theoretical evidence indicates that water−Au(111) interactions are weak relative to the hydrogen bonds between water molecules.[33,34,40] To minimize the intermolecular interaction energy in the interfacial region at an uncharged surface, water forms hydrogen-bonded clusters (in some studies described as having an ice-like structure[40,42,44]) in which the water dipoles in contact with the metal surface are oriented in the surface plane or slightly tilted with hydrogen oriented toward the bulk water (oxygen points to the surface).[33,42] MD simulations and *ab initio* calculations show that the water density has a maximum in the vicinity of the Au(111) surface (the first water layer) and some density fluctuations at larger distances caused by screening effects (the hydration shell).[34,37−40] This effect is responsible for some energy penalty due to displacement of the hydration shell upon adsorption, which makes binding to gold in water less favorable than in a vacuum as observed in MD simulations.[24] Furthermore, oriented water molecules on the metal surface, acting as dipoles, induce an electrostatic field, which may affect the behavior of charged and polar molecules.[39] Taking all these data into account, it is reasonable to include in the continuum solvent model both the hydrophobic and the electrostatic effects of the interaction with the hydration shell of the metal, and to employ MD simulations of water molecules to compute their hydrophobic and electrostatic contributions to the desolvation energy, in order to parametrize them.

Finally, to check the designed continuum solvent model, we need a well-characterized test system that not only allows us to verify the reliability of our representation of solvation effects and the derived energy parameters but also helps us to understand the contributions of different interaction mechanisms to the total protein−metal binding free energy. It is reasonable therefore to start with validation on small systems, whose interaction can be studied accurately either by experimental or by theoretical methods or both, and then make use of parameter transferability to apply the method to larger ones. The natural choice of such small systems is a set of amino acids whose binding to metal surfaces has been analyzed using MD simulations[28,29] and can also be related to available experimental studies.[1,5,7,9] For the sake of consistency, we used the same gold surface representation, amino acid structures, and force field parameters as used previously in the MD simulations.[28,29] We also employed the same image-charge model and water force field in the MD simulations with explicit water molecules performed to support the development of the continuum approach presented here. Thus, we note that the ProMetCS model developed is based on the force field employed in MD simulations (the GolP model[27]) and, therefore, inherits the limitations of the latter.

The paper is organized as follows. In the next section (Computational Methods), we describe the procedure used for calculating adsorption free energy. We show how the effects of solvent−metal−solute interactions can be approximated by physics-based energy terms, parametrized using explicit solvent simulations, and how they are implemented in the ProMetCS model. We give details of the MD simulations of the behavior of water molecules in the metal hydration shell, which we have used for the design and parametrization of the desolvation energy term. Finally, we show how the adsorption free energy and the potential of mean force (PMF) obtained from MD simulations can be calculated with the ProMetCS model. In the following section, we present the results of the application of the ProMetCS model to amino acid residues and compare the computed PMF binding energies with those from MD simulations as well as with available experimental data. In Appendix I, we give details of the MD simulations of the behavior of water molecules and test atoms in the surface hydration shell that were used for the design and parametrization of the desolvation energy. In Appendix II, we estimate the influence of the intrinsic electrostatic field of the hydration shell on the adsorption of charged molecules.

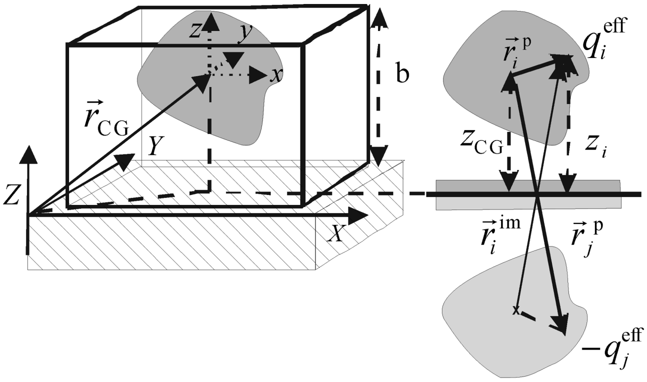


Figure 1. Illustration of the simulation box used for the calculations (left panel) and of the protein-image system employed for computing metal polarization effects (right panel). The low dielectric cavities of the protein and surface are shaded dark gray, and their images are light gray. The gold cluster is shown by a hatched block. Vectors are defined as $\vec{r}_j^{\,p} \equiv (x_i^p, y_i^p, -z_i^p - 2z_{CG})$ and $\vec{r}_i^{\,im} \equiv (x_i^p, y_i^p, z_i^p + 2z_{CG})$ from the geometric centers of the real and image protein, respectively. z_{CG} is the distance between the geometric center of the protein and the surface, and $z_i \equiv z_{CG} + z_i^p$ is the distance between a protein effective charge *i* and the surface.

2. Computational Methods

2.1. Description of System (Setup). The Au(111) surface is described by a gold cluster with atomic layers. A minimum of three layers is necessary (and sufficient) for the accurate calculation of protein−gold van der Waals interactions, as will be shown below. During the calculations, the position of the cluster is fixed with the centers of the atoms in the surface layer at $z = 0$, i.e., in the xy plane of the simulation box, as illustrated in Figure 1. The surface area of the cluster must be larger than the size of the adsorbate in order to account for interatomic interaction effects up to the cutoff employed in calculations (see below). In the present study, a gold cluster with surface dimensions of 100 Å × 100 Å was employed. Since we used the force field parameters for the biomolecule−gold interaction derived in ref 27, the cluster was constructed accordingly (see details below). In calculations of electrostatic and desolvation effects, the gold surface was considered to be a plane.

A distance b from the surface defines the limit of the simulation box where the protein−surface interaction energy is negligible and serves as a reference state for the calculation of the protein adsorption free energy and PMFs. The $z = b$ plane is also used for the generation of the starting positions of the adsorbate for the computation of BD trajectories.

Amino acid residues capped with an acetyl group at the N terminus and a methylamide group at the C terminus, corresponding to those studied in refs 28 and 29, were employed as test adsorbates. Calculations were performed for all 20 natural amino acids with their side chains assigned the standard protonation state at pH 7. Cysteine is known to form a strong bond with the gold surface which cannot be described by the Lennard-Jones-based force field parameters and a rigid gold surface. Therefore, we considered only the protonated form (denoted CysH), which cannot form a strong bond to gold, for the present simulations. To evaluate the effect of conformational variability of the capped amino acids upon binding to gold, we compared the binding properties of several of the most populated binding conformations obtained in MD simulations.[28]

2.2. Implicit Solvent Model: Interaction Energy Function. The protein−metal interaction energy function, U, which implicitly includes solvent effects, is expressed in the ProMetCS model as a sum of three separate contributions:

$$U = E_{\mathrm{LJ}} + U_{\mathrm{EP}} + U_{\mathrm{desolv}} \tag{1}$$

The E_{LJ} energy term describes nonpolar, van der Waals, and weak chemical interactions between a protein and a metal surface. It is parametrized to reproduce experimental binding properties of small organic molecules on gold.[27] It is a sum of classical Lennard-Jones 12−6 terms and will hereafter be denoted as a Lennard-Jones (LJ) term.

U_{EP} is the protein−metal electrostatic interaction free energy in aqueous solvent. (In the general case, it also includes the energy due to the electrostatic interaction between the charges in the protein binding site and the interfacial water potential on the metal surface, see Appendix II; this latter term is neglected in the implementation described in this work.)

The last term in eq 1, U_{desolv}, describes desolvation effects, i.e., the free energy change arising from protein−water, solid surface−water, and water−water interactions. Desolvation effects can be further split into two separate components: the desolvation energy of the protein, $U^{\mathrm{p}}_{\mathrm{desolv}}$, and the desolvation energy of the metal surface, $U^{\mathrm{m}}_{\mathrm{desolv}}$:

$$U_{\mathrm{desolv}} = U^{\mathrm{p}}_{\mathrm{desolv}} + U^{\mathrm{m}}_{\mathrm{desolv}} \tag{2}$$

The first term, the nonpolar (or hydrophobic) protein desolvation energy, is the free energy change of the protein−water system that arises from the replacement of the protein−water interface in the region of the adsorption site by a protein−vacuum interface. The second term in eq 2 represents effects arising from the partial replacement of the metal hydration shell by a protein adsorption site and is given by the free energy change due to insertion of a hydrophobic cavity (which mimics the binding site of the protein) into the hydration shell of the metal surface (note that the change of the protein−metal electrostatic interaction due to surface desolvation is instead included in the electrostatic energy term, U_{EP}).

It should be noted that the entropy contribution to the function represented by eq 1 [specifically, the second and third terms] is limited to the entropy change upon binding of the solvent only. The entropy change due to the restriction of protein motion upon binding the metal surface must be calculated separately. Hence, although U in eq 1 includes some entropic effects, it does not correspond to the complete adsorption free energy. The procedure for calculation of the entire adsorption free energy will be considered at the end of the present section. The three terms contributing to U in eq 1 are now described in more detail.

2.2.1. Lennard-Jones Term: E_{LJ}. van der Waals and weak chemical interactions between the biomolecule and the gold surface are described by the sum of 12−6 Lennard-Jones atom−atom pair potentials corresponding to interactions between each atom i of the biomolecule and each atom j of the gold cluster

$$E_{\mathrm{LJ}} = \sum_{j}\sum_{i} 4\varepsilon_{ij}[(\sigma_{ij}/R_{ij})^{12} - (\sigma_{ij}/R_{ij})^{6}] \tag{3}$$

where R_{ij} is the interatomic distance and

$$\varepsilon_{ij} = \sqrt{\varepsilon_{ii}\varepsilon_{jj}} \text{ and } \sigma_{ij} = \sqrt{\sigma_{ii}\sigma_{jj}}$$

are the (OPLS/AA-like) gold force field (GolP) parameters optimized by Iori et al.[27] for the interaction between organic molecules and a Au(111) surface.

The most important additions introduced in the GolP force field with respect to the standard OPLS/AA force-field[46] can be briefly summarized as follows: (i) The physical position of each Au atom in the upper layer of the gold cluster was replaced by two virtual atoms that occupy hollow sites. This particular representation of the structure of the surface layer has been proposed[27] to reproduce the correct binding position of the adsorbed molecules on the Au(111) surface. (ii) A new generic atom type for the Au atom was introduced, with generic $\varepsilon_{\mathrm{AuAu}}$ and σ_{AuAu} LJ parameters to be used for calculating E_{LJ} for gold−water and gold−protein atom pairs. (iii) Specific LJ parameters for the interaction between Au and the unprotonated N atom in His and the S atoms in CysH/Met were optimized to introduce N−Au and S−Au chemical bonding, respectively. (iv) The ε_{ij} value of carbon atoms in π rings was fitted to reproduce the rather strong interaction between the π electrons of aromatic molecules and the metal surface observed experimentally (if the π ring is oriented parallel to the surface plane). (v) A shell type model describes polarization effects of the gold surface,[47] although the latter feature is not used in the calculations presented here. Details on the derivation of the GolP parameters and a comparison between the adsorption energies calculated with GolP and experimental results for different molecules (typical deviations of less than 5−10% or a few kJ/mol), can be found in ref 27.

The direct pairwise calculation of the E_{LJ} energy between all atoms of the protein and of the metal cluster is too expensive for the large biological molecules usually studied by BD methods. Therefore, a grid-based procedure was implemented in which the LJ interaction energy between the protein and a gold atom is saved on the nodes of a three-dimensional grid with the origin placed at the protein center. The grid size is chosen so that the long-range limit of the protein atom−Au interactions with a cutoff of ∼10 Å is inside the LJ grid. Then, the E_{LJ} interaction energy between the gold surface and a protein can be obtained by summation over all Au atoms of the gold cluster.

The balance between the repulsive and attractive parts of the LJ potentials arising from neighboring protein atoms is extremely important at small protein−surface distances. The binding energy is thus very sensitive to the grid spacing. Our test calculations of amino acid adsorption in a vacuum showed, for example, that a spacing of 0.2 Å may lead to an error in binding energy of up to about 3 kJ mol^{-1} compared to the binding energy of amino acids obtained directly by summation of the pairwise terms in eq 3. For comparison, a grid spacing of 0.5 Å results in an error in the binding energy of up to 12 kJ mol^{-1}. Therefore, a grid spacing of 0.2 Å has been used throughout the present study. Both the accuracy of the computed energy and the calculation speed depend on the number of gold layers employed. The optimal number of layers is three, since the energy correction due to adding a fourth layer is smaller than the uncertainty due to the grid discretization.

At short interaction distances (less than the sum of the atomic van der Waals radii in the OPLS force field), a constant positive energy of 100 kJ mol^{-1} was assigned to avoid very strong repulsion and therefore excessively high forces in BD simulations.

2.2.2. Protein−Metal Electrostatic Free Energy in an Aqueous Solution: U_{EP}. The interaction of a fixed set of partial point charges with a flat infinite uncharged metal surface is represented in classical electrostatics by the interaction between real charges q_i and their image charges, $q_i^{\mathrm{img}} = -q_i$, placed symmetrically with respect to the metal surface plane. This approximation was shown to give good agreement with density functional calculations at a surface-charge distance of >2.5 Å.[48] Likewise, the electrostatic field of a fixed charge density in a nonuniform dielectric medium in the presence of the uncharged metal surface can be simulated by introducing an oppositely charged mirror image of the charge system instead of the metal surface. It is important that, to satisfy the boundary conditions (zero surface potential of the metal), electrostatic potentials of the protein and its opposite-charged image should exactly cancel each other at the surface plane. Therefore, not only the spatial distribution but also the dielectric surroundings of the real/image charges should be symmetrical with respect to the metal surface plane. Practically, in the implicit solvent model, a protein interacting with its image system consists of two charge distributions (one distribution for the real protein and one for its image), each immersed in low dielectric cavities surrounded by a high dielectric solvent and separated by a low dielectric cavity that surrounds the metal surface. The latter cavity is introduced since the centers of the surface layer of metal atoms (defining the metal surface plane) are separated from the solvent by the LJ radius for the metal−water interaction.

The image potential is defined as $\Phi^{\mathrm{im}}(\vec{r}_i^{\,\mathrm{im}}) \equiv -\Phi(\vec{r}_j^{\,\mathrm{p}})$, where $\vec{r}_i^{\,\mathrm{im}} \equiv (x_i^{\mathrm{p}},y_i^{\mathrm{p}},z_i^{\mathrm{p}} + 2z_{\mathrm{CG}})$ and $\vec{r}_j^{\,\mathrm{p}} \equiv (x_i^{\mathrm{p}},y_i^{\mathrm{p}}, - z_i^{\mathrm{p}} - 2z_{\mathrm{CG}})$ are vectors from the geometric centers of the image and real protein, respectively, and z_{CG} is the distance between the protein center and the metal surface as illustrated in Figure 1. Hence, we replace the protein−metal electrostatic interaction by a protein-image interaction with an additional low-dielectric cavity between the protein and the image as illustrated in Figure 1.

The electrostatic interaction free energy of two macromolecules (including solvent-related entropic effects only) can be calculated by numerical solution of the Poisson−Boltzmann equation. This however requires considerable computational resources and cannot be done at each time step of a BD simulation. Alternatively, the problem can be quite accurately solved by using the effective charge approximation for macromolecules (ECM) developed for protein−protein interactions.[49] Following this work, we describe the electrostatic interaction free energy between a protein and its image in the presence of the metal cavity as

$$U_{\mathrm{EP}} = U_{\mathrm{p}}/2 + U_{\mathrm{im}}/2 + U_{\mathrm{p-c}} + U_{\mathrm{im-c}} \tag{4}$$

where U_{p} (U_{im}) corresponds to the energy of interaction of the protein (image) charges with the image (protein) electrostatic potential computed in the presence of both protein and image cavities as well as the metal cavity; $U_{\mathrm{p-c}}$ ($U_{\mathrm{im-c}}$) describes perturbation of the protein (image) electrostatic potential by the low-dielectric cavity of the image (protein). The latter term decreases rapidly with the protein-image distance (i.e., with the distance from the metal surface) and will be referred to hereafter as the electrostatic desolvation energy.

The first and the second terms in eq 4 are equal, and so are the third and fourth terms. Thus,

$$U_{\mathrm{EP}} = U_{\mathrm{p}} + 2U_{\mathrm{p-c}} \tag{4a}$$

For the real protein, the effective charges, q_i^{eff}, in a uniform high dielectric medium give the same electrostatic potential outside the protein surface as that computed for the real protein treated as a low dielectric cavity immersed in high dielectric solvent.[49] The electrostatic energy, U_{p}, can then be approximated by the interaction energy of the real protein effective charges q_i^{eff},[49] immersed in a uniform solvent medium, with the electrostatic potential of the protein image

$$U_{\mathrm{p}} = \sum_i \Phi^{\mathrm{im}}(\vec{r}_i^{\,\mathrm{im}})q_i^{\mathrm{eff}} \ (\text{where } \Phi^{\mathrm{im}}(\vec{r}_i^{\,\mathrm{im}}) \equiv -\Phi(\vec{r}_j^{\,\mathrm{p}})) \tag{5}$$

The electrostatic potential, $\Phi(\vec{r}_j^{\,\mathrm{p}})$, of a protein in water was calculated by numerically solving the linearized Poisson−Boltzmann equation using the UHBD (University of Houston Brownian Dynamics) program.[50] The relative dielectric constant of the protein was assigned as 4 and that of the solvent as 78, and the dielectric boundary was defined by the van der Waals radii of the protein atoms. The protein atoms were assigned partial charges from the OPLS force field.[46] The electrostatic potential was computed on a three-dimensional grid centered on the geometric center of the protein. Since the electrostatic potential changes smoothly with $\vec{r}$, it does not require as accurate a representation as the LJ potential, and we have therefore used a grid with a spacing of 0.5 Å in the present calculations. The effective charges q_i^{eff} were positioned on selected atoms of the charged residues (the carboxylate oxygen atoms of Asp and Glu residues, as well as the amine nitrogen atoms of Lys, Arg, and protonated His residues), and their values were derived by fitting the protein electrostatic potential in a 3-Å-thick layer extending outward from the protein's accessible surface computed with a probe of radius 4 Å.[49,51]

The last term in eq 4a is the electrostatic desolvation term of the protein with effective charges q_i^{eff} due to presence of the image protein cavity and the metal surface cavity. This can be accounted for by the introduction of a positive energy term analogous to that proposed in ref 49 as

$$U_{\text{p-c}} = \sum_i (\Phi_{\text{ed}}^{\text{met}}(\vec{r}_i) + \Phi_{\text{ed}}^{\text{im}}(\vec{r}_i)) \times (q_i^{\text{eff}})^2 \qquad (6)$$

A general equation for the electrostatic desolvation potentials, $\Phi_{\text{ed}}(\vec{r})$ [$\Phi_{\text{ed}}^{\text{met}}(\vec{r})$ or $\Phi_{\text{ed}}^{\text{im}}(\vec{r})$], due to a set of spherical low dielectric cavities is given in the dipole approximation in ref 52 as

$$\Phi_{\text{ed}}(\vec{r}_i) = \alpha \frac{\varepsilon_{\text{s}} - \varepsilon_{\text{p}}}{\varepsilon_{\text{s}}(2\varepsilon_{\text{s}} + \varepsilon_{\text{p}})} \sum_j (1 + k\vec{r}_{ij})^2 \exp(-2kr_{ij} \frac{a_j^3}{r_{ij}^4}) \qquad (7)$$

k is the Debye–Hückel parameter, ε_{p} is the protein dielectric constant, ε_{s} is the solvent dielectric constant, a_j is the van der Waals radius of the jth atom of the protein image (or atoms of the metal surface), and r_{ij} is the distance from the jth atom to the effective charge of the protein q_i^{eff}. The scaling factor α was estimated[52] for protein–protein association as $\alpha = 1.67$. Since $\Phi_{\text{ed}}^{\text{met}}(\vec{r}_i)$ is at least 2^4 times larger than $\Phi_{\text{ed}}^{\text{im}}(\vec{r}_i)$ (eq 7), we can omit the effect of the image cavity on the electrostatic field of the protein and keep only the metal cavity terms:

$$U_{\text{EP}}(\vec{r}) = \sum_i \Phi^{\text{im}}(\vec{r}_i^{\text{im}}) q_i^{\text{eff}} + 2 \sum_i \Phi_{\text{ed}}^{\text{met}}(\vec{r}_i) \times (q_i^{\text{eff}})^2 \quad (8)$$

The electrostatic energy given by eq 8 has been derived for the case of nonoverlapping cavities of the protein and the metal surface. To complete this model, we have to consider the case in which the adsorbed molecule penetrates the first hydration layer of the surface, which in the context of the implicit solvent model means that the low-dielectric cavities of the protein and the metal merge. In general, this has two effects: (i) The change in Born solvation energy should be taken into account; this, however, rapidly vanishes with increasing adsorbate size[53] and can be neglected in the case of molecular adsorption (the case of ions will be discussed at the end of the present section). (ii) The metal-charge interaction energy must be scaled appropriately for the transition from high to low dielectric surroundings.

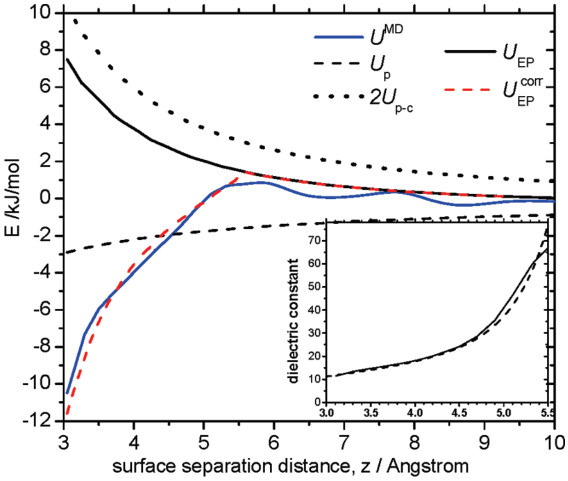


Figure 2. Total electrostatic energy for a test charge atom as a function of distance from the gold surface with explicit (U^{MD}) and implicit ($U_{\text{EP}}^{\text{corr}}$) water models ($U_{\text{p}}$, U_{EP}, and $U_{\text{p-c}}$ are separate contributions to the electrostatic energy, see text for details). Insert: Plot of effective dielectric constant derived from the image-charge potential computed from explicit water simulations (solid line) and approximated by an analytical function (dashed line).

Effect ii can be estimated from the electrostatic energy of an ion in the presence of the metal surface obtained in an MD model in which the solvent is treated explicitly. To this end, we have computed the image-charge energy, U^{MD}, of a test charge atom (with unit charge and $\sigma_{ii} = 2.87$ Å) as the difference in ion energy in explicit-water simulations with and without image-charge effects, see Figure 2. One can see in Figure 2 that, at surface separation distances smaller than $z \sim 5.5$ Å, the electrostatic ion-metal energy computed in the explicit water model, U^{MD}, is much lower than that obtained in the present implicit solvent approximation, U_{EP}. This z value can be considered as the approximate ion-surface distance at which the ion (or an effective charge in a molecule) and surface cavities start to merge. Indeed, this agrees with the Au–water and test charge–water LJ radii (~3 Å and ~2.5 Å, respectively).

The simplest way to account for this effect at small charge-surface distances in the ProMetCS model is to introduce a variable dielectric constant that increases as an effective charge moves away from the surface and reaches the value of $\varepsilon = \varepsilon_{\text{s}}$ when the cavities of the charge and the surface are separated and water molecules are able, at least partially, to screen the charge–metal interaction. Keeping the electrostatic desolvation energy $U_{\text{p-c}}$ unchanged, we fitted the dependence of the dielectric constant on the unit charge-surface separation distance, z, to reproduce the explicit-water electrostatic energy:

$$U_{\text{EP}}^{\text{corr}} \equiv \frac{-1}{2z(4\pi\varepsilon_0)\,\varepsilon(z)} + 2U_{\text{p-c}} \approx U^{\text{MD}} \qquad (9)$$

and therefore,

$$\varepsilon(z) = \frac{1}{2z(4\pi\varepsilon_0)(2U_{\text{p-c}} - U^{\text{MD}})}$$

The computed variable dielectric constant can be approximated by an analytical function $\varepsilon(z) = 4.0 + 0.8z^2 + \exp(z/0.385 - 10.4)$, where z is in Å, for $z < 5.5$ Å (see the insert in Figure 2).

In the case of a set of effective charges, the corrected electrostatic energy, $U_{\mathrm{EP}}^{\mathrm{corr}}$, can be directly applied to the diagonal terms, which correspond to the interaction of an effective charge i with its own image (for charge-surface distances of $z_i < 5.5$ Å):

$$U_{\mathrm{EP}}^{\mathrm{corr}} = U_{\mathrm{EP}} + \sum_i \frac{(q_i^{\mathrm{eff}})^2}{2z_i(4\pi\varepsilon_0)}\left(\frac{1}{\varepsilon_s} - \frac{1}{\varepsilon(z_i)}\right) \tag{10}$$

This approximation is valid for the case of amino acids that are described by one effective charge in the ECM model,[49] but for a system of many effective charges, cross-terms that correspond to the interaction between a charge and the image of another charge should also be taken into account.

The minimum value of the relative dielectric constant is ~10 (see Figure 2), which is consistent with typical values used in modeling of the electrochemical interface.[54] This value can lead to an up to 4-times larger Coulomb energy for a monatomic ion in pure water.

2.2.3. Protein Nonpolar (Hydrophobic) Desolvation Energy. The free energy change of a protein due to its partial desolvation by the gold surface can be described by a nonpolar desolvation energy that is proportional to the solvent accessible surface area (SASA) of a protein and an energy coefficient (Φ_{pd}):[55]

$$U_{\mathrm{desolv}}^{\mathrm{p}} = \sum_m \Phi_{\mathrm{pd}}\mathrm{SASA}_m \tag{11}$$

where the energy potential Φ_{pd} is computed on a three-dimensional grid and is defined as a function of the distance r from the van der Waals surface of a protein:[55]

$$\Phi_{\mathrm{pd}}(r) = \beta\, c\begin{cases} 1 & \text{if} \quad r < a \\ \dfrac{b-r}{b-a} & \text{if} \quad a < r < b \\ 0 & \text{if} \quad r > b \end{cases}$$

The parameters a and b have been optimized[55] by using a standard method for SASA calculations (NACCESS) and are set to 3.1 Å and 4.35 Å, respectively; $c = 0.5$; the coefficient β was set to ~−0.021 kJ mol^{-1} Å^{-2} in the present calculations. It should be noted that the regions of nonzero desolvation energy and LJ binding energy strongly overlap, and this may lead to the relatively smaller hydrophobic desolvation term being dominated by the larger LJ attraction.

2.2.4. Metal Desolvation Energy for Nonpolar Adsorption Sites. To understand the nature of the solvation effects arising from the partial replacement of the metal hydration shell by a biomolecule, we considered the properties of the water in the vicinity of the Au(111) surface that can be derived from MD simulations. We first computed the partial water density as a function of surface water separation distance from a simulation of bulk water in the presence of an Au(111) surface, see Figure 3. The hydration shell consists of two water layers (at 3 Å and 6 Å) with a high partial density of water molecules. The comparison of the density of oxygen and hydrogen inside the first and second layers is higher than that of hydrogen, which indicates a nonuniform orientation of the water molecules, in agreement with other studies.[38,40] We then computed the PMF for one water molecule as a function of the surface water separation distance, see Figure 4. From the PMF, the computed binding free energy for a water molecule is ~−2.8 kJ mol^{-1} and ~−0.6 kJ mol^{-1} for the first and second hydration layers, respectively. The bound water in the first hydration layer is separated by a free energy barrier of ~−4.4 kJ mol^{-1} from the bulk water. The PMF shows that there will be an unfavorable positive energy change of the solvent-metal system when a water molecule is removed from the hydration shell to the bulk.

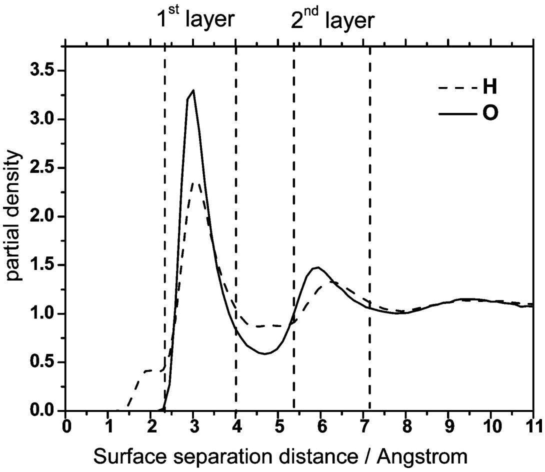


Figure 3. Dependence of the partial density of the water oxygen atoms (solid line) and hydrogen atoms (dashed line) on the distance from the gold surface computed from MD simulations of water in the presence of a gold surface. Densities are normalized to the bulk values; details of calculations are given in Appendix I.

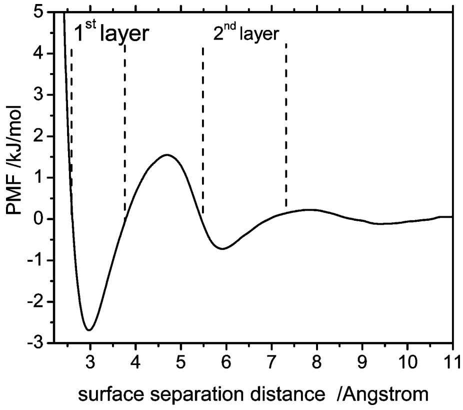


Figure 4. PMF of a water molecule as a function of the distance from the surface computed from MD simulations. Details of calculations are given in Appendix I.

The metal desolvation energy is the free energy change caused by the replacement of the hydration shell of the metal surface by the protein adsorption site. It is, therefore, reasonable to assume that the desolvation energy is proportional to the desolvated area of the metal so that we can use an expression similar to eq 11:

$$U^{m}_{\mathrm{desolv}} = \sum_{i} \Phi_{\mathrm{metd}} S^{\mathrm{desolv}}_{i} \qquad (12)$$

where the coefficient Φ_{metd} is a free energy change for desolvation of a unit surface area of the metal. A proper modeling of the metal desolvation requires Φ_{metd} to depend on the distance of the protein surface atom i from the atoms of the metal surface. At large separations, when the distance between a protein atom i and the metal surface is greater than the LJ cutoff value, $Z_{\max}$, Φ_{metd} must converge to zero. The summation in eq 12 must be carried out over the protein surface atoms, and S^{desolv}_{i} defines a desolvated area of the metal surface associated with the contacting protein atom i.

In the ProMetCS model, the desolvation energy describing replacement by a protein atom i of the first ($z_i < Z_{\mathrm{adw}}$) and the second and higher ($z_i > Z_{\mathrm{adw}}$) hydration layers is given by

$$\Phi_{\mathrm{metd}} = \begin{cases} \Phi^{0}_{\mathrm{metd}} & z_i \le Z_{\mathrm{adw}} \\ \Phi^{0}_{\mathrm{metd}} \exp(-(z_i - Z_{\mathrm{adw}})/\gamma)) & Z_{\mathrm{adw}} < z_i < Z_{\max} \\ 0 & z_i > Z_{\max} \end{cases} \qquad (13)$$

where z_i is the distance between the center of the protein surface atom i and the metal surface, Φ^{0}_{metd} is the desolvation energy per unit area of the first hydration layer, γ describes the decrease in magnitude of the desolvation energy when the second and higher hydration layers are replaced, Z_{adw} corresponds to the position of the first hydration layer as defined above (~3 Å) plus the average LJ radius for the protein−metal atom interaction, which gives $Z_{\mathrm{adw}} \sim 5$ Å, and $Z_{\max} \sim 10$ Å is the cutoff for computing the desolvation term. Using the binding free energies per water molecule derived from the PMF in Figure 4, and assuming that the surface area occupied by one water molecule is ~9 Å^2, we estimate $\Phi_{\mathrm{metd}} \equiv \Phi^{0}_{\mathrm{metd}} \sim 0.31$ kJ mol^{-1} Å^{-2} and $\Phi_{\mathrm{metd}} \sim 0.07$ kJ mol^{-1} Å^{-2} for the first and the second hydration layers, respectively, which leads to an assignment of $\gamma \sim 1.51$ Å.

It should be noted that the desolvation energy U^{m}_{desolv} of eq 12 represents only the part of the free energy change of the metal hydration shell due to replacement of parts of the hydration shell by a noninteracting cavity. Electrostatic effects caused by the interaction of the charges at the adsorption site with oriented water dipoles on the metal surface (see Appendix II) are neglected here. Hence, the value of Φ^{0}_{metd} estimated from the water adsorption energy is only a first approximation that may need further correction.

In order to calculate the desolvation area due to binding of a protein, we placed a two-dimensional grid on the surface plane, centered on the protein. Then, the positions of all protein atom−metal contacts with $z_i < Z_{\max}$ were stored on the grid, and the area defined by the distance around the contact points R_{adw} (defined below) was considered as the desolvation area (illustrated in Figure 5). The total contact areas for atoms with $z_i < Z_{\mathrm{adw}}$ and with $Z_{\mathrm{adw}} < z_i < Z_{\max}$ were calculated separately (they are shown in Figure 5 by the bold solid and dashed lines, respectively). These areas were then multiplied by the corresponding energy coefficients given by eq 13.

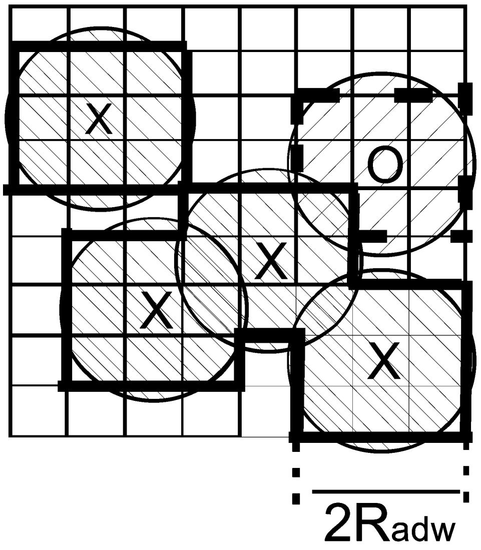


Figure 5. Illustration of the method employed for the calculation of the metal surface area in which water molecules are replaced by the adsorption site of the protein. The crosses and zeros show the positions of the centers of the protein atoms, with $z < Z_{\mathrm{adw}}$ and $Z_{\mathrm{adw}} < z < Z_{\max}$, respectively. The hatched circles with radius R_{adw} show the area each atom is assumed to desolvate. The computed desolvation area is shown by bold lines, the solid and dashed lines corresponding to water desorption from the first and second hydration layers, respectively. See the text for details.

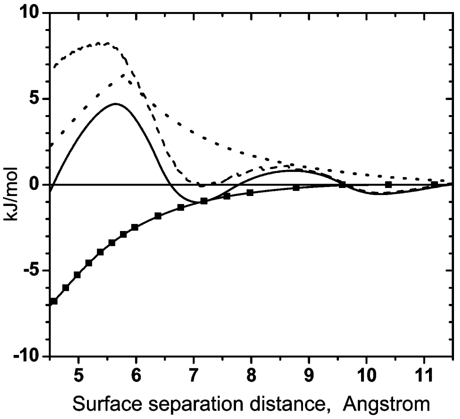


Figure 6. PMF obtained from MD simulations for the test atom (solid line), corresponding LJ potential (squares), and their difference (dashed lines) associated with the desolvation energy. Dotted line, PMF energy computed using the present model (includes both LJ and metal desolvation energies).

The value of R_{adw} was estimated by considering the desolvation energy, S^{desolv}, of a single test atom with "iodine-like" force-field parameters ($\sigma_{ii} = 5.4$ Å, $\varepsilon_{ii} = 0.293$ kJ mol^{-1}) that mimics a small nonpolar functional group of a protein. The PMF of the test atom obtained from MD simulations using a harmonic restraint potential applied along the z axis (the x and y coordinates were fixed during the simulations) is shown in Figure 6, along with the corresponding LJ potential. Since the translational entropy change along the PMF is zero for the present case, the difference between the PMF and LJ energies (dashed line in Figure 6) corresponds to the metal desolvation energy. It shows maxima at the first and second hydration layers at surface separation distances of ∼5.5 Å and ∼8.5 Å, respectively. From the magnitude of the energy maxima, we estimate the desolvation energy change associated with partial replacement of the first hydration layer by an adsorbed atom as $U^{m}_{\text{desolv}} = 8.4$ kJ mol^{-1}. With $\Phi^{0}_{\text{md}} = 0.31$ kJ mol^{-1} Å^2, S_i^{desolv} ∼ 27 Å^2, which can be described by an effective desolvation radius, R_{adw}, of ∼3 Å. The total adsorption energy for the test atom computed as a function of the separation distance with the ProMetCS model with the parameters derived above is shown by the dotted line in Figure 6. The energy function given by eq 13 is by definition not able to reproduce the energy fluctuation at $z_i > Z_{\text{adw}}$, but in the case of a protein, the contribution of this effect is expected to be relatively small. The value of $R_{\text{adw}} = 3$ Å was used throughout all the calculations and appeared to be a good first approximation, as will be shown below.

2.3. Calculation of the Adsorption Free Energy and Potential of Mean Force. We consider an adsorbate as a rigid molecule moving relative to the solid surface. The geometry of the simulation box is the same as described above and shown in Figure 1. An adsorbate energy, defined by eq 1, is generally a six-dimensional function of the protein position and orientation. Three translational degrees of freedom define the position of the molecular center of geometry, x_{CG}, y_{CG}, and z_{CG}, where $\vec{r}_{\text{CG}} = (x_{\text{CG}}, y_{\text{CG}}, z_{\text{CG}})$, and the three rotational coordinates, $\Omega = (\Omega_1, \Omega_2, \Omega_3)$, are represented by Euler angles of the coordinate frame centered at the protein.

The solid surface is usually characterized by a periodic structure, and one can, therefore, expect a periodic variation of interaction energy as the protein position is shifted in the *xy* plane. Without any loss of generality, the molecule motion in the *xy* plane can, therefore, be considered in the area of $\Delta S = \Delta x_{\text{CG}} \Delta y_{\text{CG}}$, where Δx_{CG} and Δy_{CG} define a period of energy variation along the corresponding coordinate. The unit volume of configurational space of the protein−surface system is defined as $\mathrm{d}S\, \mathrm{d}\Omega\, \mathrm{d}z$ ($\mathrm{d}S = \mathrm{d}x_{\text{CG}}\, \mathrm{d}y_{\text{CG}}$, $\mathrm{d}\Omega = \sin \Omega_1\, \mathrm{d}\Omega_1\, \mathrm{d}\Omega_2\, \mathrm{d}\Omega_3$), and the total simulation volume of the configurational space is $8\pi^2 \Delta S b$.

The free energy change upon protein adsorption is then given by[56]

$$\Delta G = -k_{\text{B}} T \ln\left[\frac{Q_{\text{b}}}{Q_{\text{f}}}\right] = -k_{\text{B}} T \ln\left[\frac{\int_{\text{b}} \mathrm{d}z \int_{\Delta S,\Omega} \mathrm{d}\Omega\, \mathrm{d}S \exp(-U(\vec{r}_{\text{CG}}, \Omega, S)/k_{\text{B}}T)}{b 8\pi^2 \Delta S}\right] \tag{14}$$

where Q_{b} and Q_{f} denote configurational partition functions of an adsorbate in the bound and free states per unit volume, $S = (x,y)$, and

$$Q_{\text{b}} = \frac{\int_{\text{bound}} \mathrm{d}z \int_{\Delta S,\Omega} \mathrm{d}\Omega\, \mathrm{d}S \exp(-U(\vec{r}_{\text{CG}}, \Omega, S)/k_{\text{B}}T)}{V_{\text{b}}} \tag{15}$$

The value $Q_{\text{f}} = 8\pi^2$ represents a uniform distribution of an unbound protein over configurational space, and $V_{\text{b}} = \Delta S b$ is the simulation volume.

Direct calculation of the complete 6-dimensional free energy landscape is difficult, and we have therefore used the system symmetry to reduce the dimensions of the energy matrix. First, due to the periodicity of the interaction potential along the x_{CG} and y_{CG} coordinates, only a small area, ΔS, must be explored. A period of the interaction potential is about the dimension of the Au metal cell. In fact, preliminary calculations showed that the greatest variations in potential occur within an area of 6 × 6 Å with a grid spacing of 0.5 Å; i.e., 13 ×13 grid nodes should be computed for the x_{GC} and y_{GC} coordinates. Variations in potential in the *xy* plane as well as with respect to rotation around *z* (Ω_3 angle) arise only from the short-range energy terms (i.e., LJ and desolvation energy terms) and can therefore be neglected if the smallest separation between protein surface atoms and the metal surface, z_{min}, is larger than the LJ cutoff, $Z_{\text{max}} \sim 10$ Å. Thus, at large distances, only the electrostatic component is important, and only three coordinates, Ω_1, Ω_2, and z_{GC}, must be explored. Moreover, since the electrostatic potential is quite smooth, the grid spacing over z_{GC} can be notably increased at $z_{\text{min}} > Z_{\text{max}}$. After some test calculations, we chose a grid spacing $\mathrm{d}\Omega_1 = \mathrm{d}\Omega_2 = 3°$, $\mathrm{d}\Omega_3 = 6°$, and $\mathrm{d}z_{\text{GC}} = 0.2$ Å at $z_{\text{min}} < Z_{\text{max}}$, and $\mathrm{d}\Omega_1 = \mathrm{d}\Omega_2 = 6°$, $\mathrm{d}\Omega_3 = 12°$, $\mathrm{d}z_{\text{CG}} = 2$ Å at $z_{\text{min}} > Z_{\text{max}}$.

Computation of the free energy change upon binding using standard molecular dynamics simulations is not feasible since it requires very extensive sampling to reach and cross high-energy regions of the underlying energy landscape. To overcome this problem, enhanced sampling techniques, such as those based on the umbrella-sampling concept,[57,58] can be used. To explore a reaction coordinate *z*, a series of simulations can be performed with a biasing harmonic restraint potential defined at each point of interest, z^0_{CG}, along the reaction coordinate: $V(z_{\text{CG}} - z^0_{\text{CG}}) = -1/2k(z_{\text{CG}} - z^0_{\text{CG}})^2$.[58] The biased energy distribution function is given by $\exp(-(U(\vec{r}_{\text{CG}}, S, \Omega) + V(z_{\text{CG}} - z^0_{\text{CG}}))/kT)$, and, for a very sharp harmonic potential, the energy distribution function can be approximately described as

$$\exp(-(U(\vec{r}_{\text{CG}}, S, \Omega) + V(z_{\text{CG}} - z^0_{\text{CG}}))/kT) \approx \delta(z_{\text{CG}} - z^0_{\text{CG}}) \exp(-U(\vec{r}^{\,0}_{\text{CG}}, S, \Omega))/kT)$$

where $\delta(z_{\text{CG}} - z_0)$ is the Dirac delta function and $\vec{r}^{\,0}_{\text{CG}} = (x_{\text{CG}}, y_{\text{CG}}, z^0_{\text{CG}})$. In this case, we are concerned with a local function at fixed $z_{\text{CG}}{}^0$ that describes the Boltzmann distribution over the adsorbate positions in the $x_{\text{CG}}y_{\text{CG}}$ plane and over the adsorbate orientation. Instead of an adsorption free energy given by eq 14, we have a PMF along the reaction coordinate, $z_{\text{GC},}$ with

$$G_{\text{PMF}}(z^0_{\text{CG}}) = -k_{\text{B}} T \ln\left[\frac{\int_{\Delta S\Omega} \mathrm{d}\Omega\, \mathrm{d}S \exp(-U(\vec{r}^{\,0}_{\text{CG}}, \Omega, S)/k_{\text{B}}T)}{8\pi^2 \Delta S}\right] \tag{16}$$

It is important to note that the $G_{\text{PMF}}(z_{\text{CG}})$ given by eq 16 includes the same kinds of entropy contributions as the MD simulations and can, therefore, be directly compared with the MD results.

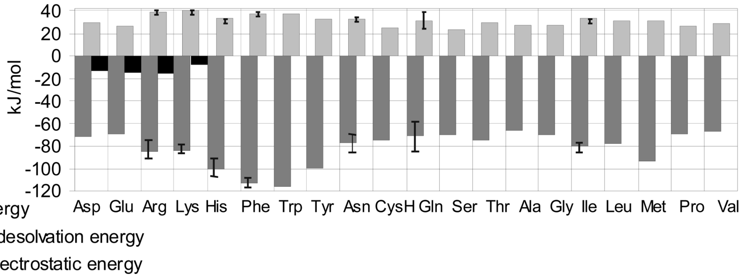


Figure 7. Contribution of the LJ, metal desolvation, and electrostatic (U_{EP}^{corr}) terms to the binding energy of capped amino acids on gold as calculated with the ProMetCS energy function. Error bars show energy deviation for different binding conformations used in simulations.

3. Results and Discussion: Testing of the Model for Adsorption of Capped Amino Acids on the Au Surface

To evaluate the accuracy of the energy model described, we computed the adsorption free energies and the PMFs of capped amino acids and compared the results with those obtained in MD simulations reported recently.[28,29] The main aim of this comparison was the testing of the proposed implicit solvent model against results with an explicit representation of water molecules. However, the solvent representation is obviously not the only difference between the ProMetCS energy function and that used in the MD simulations. To minimize the differences in physical characteristics of the amino acid−gold−water system employed in the two models, we used the same structure of the gold cluster and the same force-field parameters for the LJ energy as used by Hoefling et al.[28,29] Furthermore, in both models, an image-charge approximation was employed for calculation of the electrostatic effects. Finally, we used the most populated binding conformation of each capped amino acid obtained in MD simulations[28,29] in the present simulations. If several conformations with comparable populations were reported, we carried out simulations for all of them separately. However, we did not take into account the change in internal energy of the molecules upon conformational transition during the adsorption process, and this may cause some uncertainty in binding energy as will be discussed below in more detail.

Before presenting the results of the PMF simulations, let us consider the relative contributions of the energy terms of eq 1 to the binding energy of the amino acids. The largest contribution to the binding energy for almost all the amino acids arises from the LJ term, see Figure 7. As can be expected, the LJ energy increases with the number of atoms contacting the surface and, therefore, tends to increase with the size of the amino acid side chain. Due to the empirical design of the GolP force field parameters used in the present study, the binding to gold of Cys, Met, and His is favored if sulfur or nitrogen, respectively, comes close to the surface. Similarly, molecules with π rings (His, Phe, Trp, Tyr) are rather strongly bound to gold if their rings are parallel to the plane of the surface. Indeed, the absolute value of the LJ term shown in Figure 7 demonstrates the largest magnitudes for His, Met, Phe, Tyr, and Trp. In His, we found that binding through the π ring is stronger than attraction via an unprotonated nitrogen atom, and in our calculations, the conformation with the ring parallel to the plane of the surface is more preferable than the tilted one.

The image-charge interaction is quite weak because of the charge-image distances (>6 Å) and the high dielectric constant aqueous medium between them. Indeed, the image-charge energy (first term in eq 8, U_p) is ~ -1.5 kJ mol^{-1} for all charged amino acids. Moreover, as an effective charge approaches the metal surface, induced solvent polarization around the low-dielectric cavities makes the electrostatic interaction effectively repulsive. This effect is described by the positive electrostatic desolvation penalty ($2U_{p-c} \sim +5-7$ kJ mol^{-1} at an ionic strength of 150 mM used in the MD simulations). In fact, the total electrostatic energy becomes negative only when an effective charge penetrates the hydration shell of the metal and its field is not screened by the water molecules any more. This effect, simulated by the variable dielectric constant, leads to an electrostatic energy of about -7 to -14 kJ mol^{-1} for charged residues, which is, however, still notably smaller than the $|E_{LJ}|$ binding energy of up to ca. 115 kJ mol^{-1}.

Whereas the electrostatic contribution to binding to a neutral gold surface is small for capped amino acids, as is the favorable hydrophobic protein desolvation energy ($|U_{desolv}^{p}|$ < 3 kJ mol^{-1}), the positive metal desolvation penalty varies from +20 kJ mol^{-1} to +40 kJ mol^{-1} and provides the largest compensation to the LJ term, see Figure 7. Since, U_{desolv}^{m} is proportional to the capped amino acid−metal contact area, the larger residues are in general characterized by a larger desolvation penalty as well as larger LJ binding energies. From Figure 7, one can also see that residues with long side chains, such as Arg and Lys, have a larger desolvation penalty than more compact residues. On the other hand, comparing all amino acids, the difference in the U_{desolv}^{m} value does not exceed 15 kJ mol^{-1}, while the LJ energy differs by up to $\sim$50 kJ mol^{-1}.

Finally, the binding energy is additionally compensated by the loss in translational and rotational entropy of the molecule upon binding to the surface. The entropy contribution has quite a small dependence on the amino acid type since all the capped amino acids have a well-defined binding position that corresponds to a rather sharp energy minimum. The entropic part due to restriction of rotation and of translation in the xy plane (which is included in the PMF) is about 25 kJ mol^{-1}, whereas the entropy difference due to translation along the z coordinate is $\sim$10 kJ mol^{-1}.

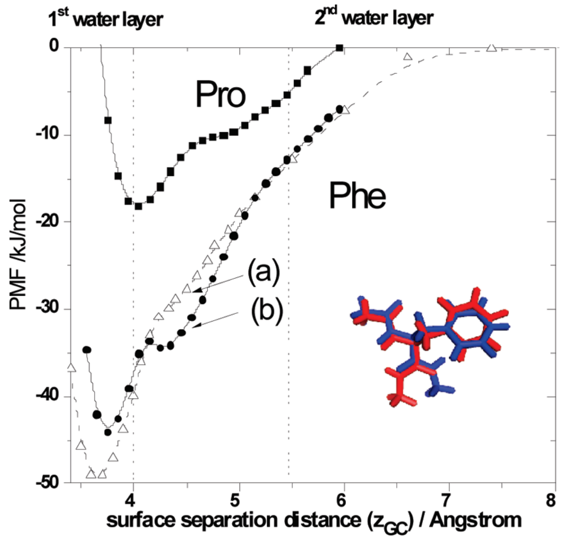


Figure 8. Representative PMF profiles computed using the ProMetCS model for two capped amino acids shown as a function of surface the separation distance: squares, Pro; triangles, circles, Phe. For Phe, PMFs computed for two slightly different conformations are shown: the aromatic ring is in the backbone plane (a, red conformation) and slightly tilted (b, blue conformation). Phe conformations are shown in projection onto the Au surface plane.

Representative PMF profiles for weakly (Pro) and strongly (Phe) bound residues are shown in Figure 8. The shape of the PMF, and the value of its global minimum depend on the conformation of the molecule used in the simulations. For example, the Phe conformation with the aromatic ring and backbone oriented in the same plane is more strongly bound and has only one minimum, whereas tilting the aromatic ring with respect to the backbone plane leads to two minima in the PMF (at ∼3.7 Å and ∼4.4 Å, see Figure 8), corresponding to the orientation of either the side chain or the backbone in the surface plane, respectively.

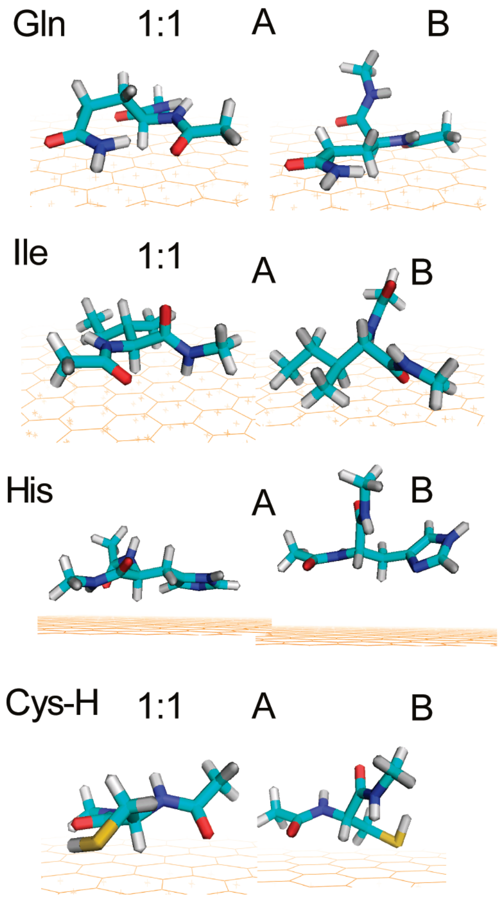


Figure 9. The most populated binding conformations of capped amino acids[28] for which significant dependence of computed PMF binding energy on conformation was observed, shown with their relative populations in the MD simulations. The corresponding PMF binding energies computed with the ProMetCS model are as follows: Gln, −40 (A) and −27 (B) kJ mol^{-1}; Ile, −24.3(A), −16.3 (B) kJ mol^{-1}; His, −48 (A), −31.9 (B) kJ mol^{-1}, where configurations A and B correspond to HIE and HID, respectively; CysH, −43 (A), −33(B) kJ mol^{-1}.

From the above analysis, we conclude that the relative binding strength of the amino acids in the present model is mainly driven by the LJ energy term. Since the LJ binding energy is very sensitive to the positions of the interacting atoms, conformational effects can be very important in the adsorption energy calculations. However, with the rigid-body approximation, the conformation is not adjusted at each simulation step, and therefore, changes in the internal energy of adsorbed molecules are not included in the present simulations. Therefore, to account for the existence of multiple conformations and minimize the uncertainty in the computed energies due to neglecting the change in internal molecular energy upon binding, calculations were carried out for some of the capped amino acids for several of the most populated binding conformations obtained in the MD simulations.[28] For most of the residues, there was only one dominant binding conformation in the MD simulations. For Asn, Arg, Cys, Gln, Glu, Leu, Lys, and Tyr, however, there were two bound conformations with similar populations, and for Ile, there were at least three conformations. The variation in binding energies obtained in the ProMetCS model for the different conformations is less than within ∼5 kJ mol^{-1} for most of these amino acids, except for Gln, Ile, and Cys, whose binding energy variation reaches ∼10−15 kJ mol^{-1}. The most populated binding conformations for the latter residues are shown in Figure 9 along with their relative populations; the corresponding computed PMF binding energies are given in the figure caption. For all of these residues, the most strongly bound conformation has a nearly "flat" geometry (denoted as A in Figure 9) with the side chain as well as part of the backbone oriented parallel to the plane of the surface so that the LJ energy is optimized for the most atoms. In Figure 9, two binding conformations of His, "flat" and "tilted", which were observed in MD simulations[28] for two His forms corresponding to protonation of different nitrogen atoms in the aromatic ring (HIE, HID), are also shown. These have a binding energy difference of ∼16 kJ mol^{-1} in the present calculations with the ProMetCS model, but almost equal binding free energies were computed from the MD simulations.[29] Here, the energy difference may be caused by underestimation of the desolvation penalty of the aromatic ring if it is placed in the surface plane since, as noted above, the U^{m}_{desolv} value is relatively small for aromatic residues.

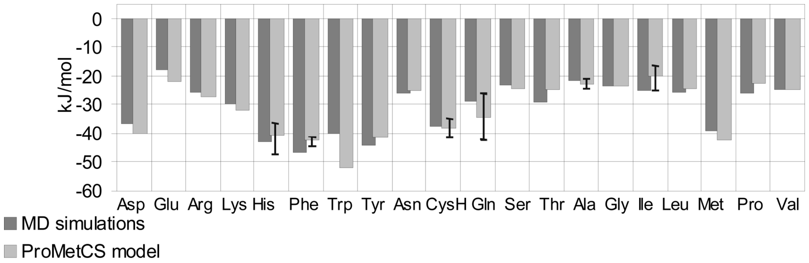


Figure 10. PMF binding free energies of the capped amino acids on the gold surface obtained using the ProMetCS model and MD calculations[29] as calculated with the ProMetCS energy function. Error bars show energy deviation for different binding conformations used in simulations.

Figure 10 shows a comparison of the PMF binding energies (the minimum of the PMF along the *z* coordinate) computed from the MD simulations[29] and with the ProMetCS model. For the residues mentioned above, characterized by several binding conformations in MD simulations that have notably different binding energies in ProMetCS, the PMF energies for the different conformations were averaged with equal weights. With the exception of Trp, the deviation of the PMF binding energies computed with the ProMetCS model from those from MD simulations does not exceed 5 kJ mol^{-1}. The binding energy of Trp is overestimated by ∼13 kJ mol^{-1}, which may be due to two reasons: (i) underestimation of the metal desolvation energy for aromatic residues, which was discussed above and might be especially pronounced in the case of Trp (the notable overestimation of the binding energy for conformation A of His is consistent with this suggestion, see Figure 9); (ii) a conformational energy change upon binding that is not taken into account in these calculations.

Taken into account the uncertainties in the present calculations, we can select capped amino acids with adsorption energies below −40 kJ mol^{-1} and above −25 kJ mol^{-1} and assign them to groups of strongly and weakly bound amino acids, respectively. Thus, His, Met, Phe, Trp, and Tyr belong to the group of strong binders, whereas Ala, Glu, Gly, Ile, Leu, Pro, Ser, and Val can be described as weak binders. The list of "strong binders" agrees with the experimental study of Peelle et al.,[5] in which notable binding was observed only for homopeptides of Cys, His, Met, and Trp. Furthermore, high affinity to the gold lattice has also been suggested for Trp and Tyr by Hnilova et al.[9] However, Phe was not mentioned among the amino acids with high affinity to gold reported in these experimental studies.[5,9]

In general, some overestimation of computed binding energy relative to that observed experimentally might be expected because, in particular, there should be some conformational restrictions on the amino acids in the peptides studied in experiments (i.e., the optimal binding conformation of an amino acid on the surface may be greatly unfavorable in the peptide). Furthermore, the binding to gold of the capping residues used in the calculations may additionally contribute to the computed affinity, leading to some overestimation of binding strength, especially for weakly bound residues. For example, the adsorption free energy of L-phenylalanine derived from electrochemical measurements[7] was characterized as typical for weak chemisorption of small aromatic molecules (from −18 to −37 kJ mol^{-1}, where the larger value is associated with electrostatic binding of the carboxylic group to a positively charged electrode). Considering that, in the present simulations, about 20−25 kJ mol^{-1} of the adsorption energy of the capped Phe molecule come from the capping residues (see adsorption geometry of Phe shown in Figure 8), the binding energy of the Phe residue can be estimated as ∼−25−30 kJ mol^{-1}, which is in the range of experimental values.

4. Summary and Future Directions

In the present paper, we propose an approximation for the calculation of the binding free energy of biomolecules on an atomically flat uncharged Au(111) surface in a continuum aqueous solvent. The interfacial interaction energy is based on an atomistic representation of short-range interactions (van der Waals, weak charge transfer, π orbital interactions) that are approximated by a set of Lennard-Jones potentials, electrostatic interactions described by the image charge method combining with the effective charge approximation, and adsorbate desolvation and metal desolvation free energies. The latter term simulates the solvation free energy change due to the replacement of part of the gold hydration shell by the uncharged binding region of an adsorbate and has been parametrized by using the results of MD simulations of water molecules on gold. MD simulations were also the basis for parametrizing a model of the desolvation effects based on the electrostatic energy. The case when the adsorption site of the biomolecule is charged and interacts with the induced electrostatic field of the oriented water dipoles on the gold surface was also considered. When parametrized using the PMFs for surface binding of negatively and positively charged ions obtained from MD simulations, this effect was found to generally lead to slightly stronger binding of positively charged adsorbates than negatively charged ones (see Appendix II).

The proposed energy model, ProMetCS, has been verified against the recently reported PMFs of capped amino acids obtained from MD simulations.[29] We computed the binding energy of 1−3 of the most populated binding conformations observed in the MD simulations for each amino acid.[28,29] When averaged over these conformations, the computed PMF minimum values (i.e., PMF binding energies) reproduce the results of MD simulations with an error of less than 5 kJ mol^{-1} for all residues except Trp. The trends in amino acid binding to gold are mostly in agreement with available experimental observations of the binding of homopeptides to gold despite the different conditions of the experiments.[5,7,9]

Analysis of the computed binding energies, in particular in comparison with experimental data, gives strong evidence that short-range van der Waals interactions (described by LJ potentials) are a driving force for adsorption of amino acids to a neutral metal surface. The change of the solvation free energy upon adsorption species is positive due to unfavorable distortion of the structure of the water layer on the gold surface. The balance of the short-range LJ attraction and the surface desolvation penalty makes the adsorption energy very sensitive to conformational variations of the adsorbed species and the orientation of the molecule on the gold surface. Therefore, it is advisible to explore a range of adsorbate conformations that are energetically accessible in aqueous solution.

As can be expected, the image-charge electrostatic effects on amino acid−gold interactions are quite small in comparison with the LJ term and the metal desolvation penalty, except in cases where a charged residue penetrates the hydration shell of the metal surface. On the other hand, in adsorption kinetics of large molecules, the electrostatic effects may gain more importance due to their long-range character.

The next step in validation of the ProMetCS model will be to apply it to a set of proteins and compare it with experimental adsorption data on the relative binding properties. Furthermore, due to the time-saving grid-based technique employed in the present model, it can be extended to the simulation of coadsorption and adsorption kinetics.

Acknowledgment. We gratefully acknowledge the support of the European Union (FP6 NEST project no. 028331- Prosurf) and the Klaus Tschira Foundation. We thank Razif Gabdoulline, Michael Martinez, Stefan Richter, Bingding Huang, and Mykhaylo Berynskyy for their assistance with various aspects of this work.

Appendix I. MD Simulations with Explicit Water Molecules

Three MD simulations with explicit water have been performed for this work:

(i) MD simulation of the water−Au(111) interface.

(ii) Calculation of the PMF of “fluorine-like” ions, i.e., an ion/atom with the LJ parameters assigned in the OPLS/AA force field to F^-, but with a +1, 0, and −1 *e* charge.

(iii) Calculation of the PMF of an “iodine-like” neutral atom, i.e., a neutral atom with the OPLS/AA[46] LJ parameters corresponding to those for I^-.

All these calculations were performed with the GROMACS (version 3.3.3 and 4.0.1) software.[59] The simple point charge, SPC, model was used for water (with rigid internal geometry constrained by the RATTLE algorithm), while the force field used for the water−gold, ion−gold, and atom−gold interactions is GolP.[27] The Au surface was simulated by a 5-layer gold slab, using a $7 \times 4\sqrt{3}$ supercell. 3D periodic boundary conditions were used. A second Au slab was placed at ∼3.5 nm from the first in the direction perpendicular to the surface (z), to confine water in a fraction of the periodic box along z (10 nm). In this way, possible spurious effects due to fictitious periodicity along z were minimized. The interslab space is large enough to have a ∼1-nm-thick region of water behaving like bulk SPC water in the middle of the slab (as verified by density profile and oxygen−oxygen correlation functions). The precise value of the interslab space was adjusted for each simulation to yield the bulk density of SPC water at 1 bar of pressure and 300 K at the center of the water layer.

The PME electrostatic model was employed in all the simulations, using GROMACS defaults for the PME parameters. For neutral systems, we performed some tests employing the Yeh and Berkowitz[60] electrostatic corrections proposed for 2D periodic systems treated with 3D periodicity. No significant variation in the reported result was found. For charged systems, no counterions were inserted to neutralize the simulation box to avoid sampling issues.[61] All the simulations were performed in the NVT ensemble, with $T = 300$ K. LJ interactions were cut off at 1.0 nm.

For simulation i, an initial equilibration of 100 ps was followed by a 5 ns production run. The density profiles in Figure 3 were obtained from the resulting trajectory. The PMF of water in Figure 4 was obtained as

$$\mathrm{PMF}(z) = -RT\ln\left(\frac{d(z)}{d_{\mathrm{bulk}}}\right)$$

where $d(z)$ is the water density in the slab centered at z and d_{bulk} is the bulk density.

PMF calculations ii and iii were performed by integration of the average force along the ion-surface separation coordinate,[62] using either umbrella-biased simulations, also called umbrella integration[63] (calculation ii), or a constraint-biased simulation, with a LINCS constraint.[64] In both cases, an initial simulation was performed in which the ion/atom was pulled through the box, along the direction perpendicular to the surface, in 1 ns. From this simulation, 30 snapshots corresponding to 30 different ion-surface distances ranging from 0.25 to 1.2 nm were extracted. From each of these snapshots, a 5-ns-long simulation was started, keeping the ion/atom at the initial distance by using a tight harmonic restraint for ii or a constraint for iii. The thermodynamic restraint/constraint forces were collected during the last 3 ns of the dynamics and averaged to get the opposite of the PMF derivative with respect to the ion-surface distance. By numerically integrating such PMF derivatives, the PMF profiles in Figures 6 and 11 were obtained.

The ability of the GolP force field to reproduce experimental adsorption energies on gold for small molecules has been verified in ref 27, and the soundness of the calculated adsorption free energies of amino acids in solution is shown in ref 29. As a further test of the force field underlying ProMetCS, we calculated the central quantity to characterize the liquid water−gold interaction, i.e., the wetting coefficient k as defined by the relation $k = (\gamma_{\mathrm{sv}} - \gamma_{\mathrm{sl}})/\gamma_{\mathrm{lv}}$ where γ_{sv}, γ_{sl}, and γ_{lv} are the solid−vapor, solid−liquid, and liquid−vapor interface tensions, respectively.[65] The difference $\gamma_{\mathrm{sv}} - \gamma_{\mathrm{sl}}$ was calculated from the 5 ns simulation described above in two ways: by the virial-based expression[65,66] as implemented in GROMACS 4.0.1 and by the energy-based method proposed in ref 20, using the entropic term correction

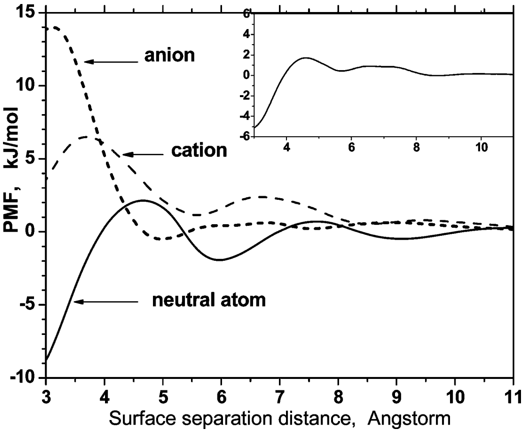


Figure 11. PMF of a test neutral atom (solid curve) and test positively and negatively charged ions (dashed and dotted curves, respectively) as a function of the surface separation distance. Insert depicts half of the difference between the PMF of a cation and of an anion.

proposed there. For the latter calculation, separate simulations of the gold slabs and the water slab were needed, and we performed simulations of 5 ns for each. In both cases, k was then calculated by using the value for γ_{lv} obtained by the virial method on the 5 ns water slab simulation. The virial-based expression yielded $k = 0.95 \pm 0.1$, while the energy-based expression yielded $k = 1.35 \pm 0.03$. For polycrystalline gold, it is known that $k \geq 1$ (i.e., the water contact angle = 0°), and for the Au(111) surface, a value close to or higher than 1 is also expected.[43] Therefore, the calculated k values are compatible with the experimental results. This discussion should not be considered as a detailed study of the water−gold surface tension obtained with GolP, which would require tests with respect to the cell size, the duration of the simulation, and the effects of the LJ cutoff. While such a detailed treatment is outside the scope of this article, it remains that the approximate k value computed here supports the use of GolP results in ProMetCS.

Appendix II. Electrostatic Interaction of an Ion with an Interfacial Water Potential

The electrostatic energy terms described above are derived for a uniform CS and do not take into account the effect related to the ordered water layer that directly contacts the metal surface. To estimate the contribution of this effect to the surface-binding energy of an ion, we first considered adsorption of test ions onto the metal surface. Computed PMFs for the positively and negatively charged test fluorine-like ion s, as well as a corresponding neutral atom, are shown in Figure 11.

The PMF function of an ion can be decomposed into four separate energy terms: (i) the LJ and the image-charge electrostatic interaction energies between the ion and the metal surface; (ii) the positive Born solvation energy given by $q/(8\pi\varepsilon_0 a)(1/\varepsilon_S - 1/\varepsilon_I)$[53] in the case of a charged atom of radius a that is transferred from a high dielectric ε_s to a low dielectric medium with dielectric constant ε_l; (iii) the free energy change of the solvent arising from distortion of the hydration shell of the metal (discussed above); and (iv) the interaction energy of the ion with the electrostatic field of the interfacial water.

Only the last term depends on the sign of the ion's charge and, therefore, can be computed as half of the difference between the PMF of the positively and negatively charged ions plotted in Figure 11. As expected, the resultant function (see insertion in Figure 11) shows fluctuations that roughly correlate with the variation of the oxygen partial density, i.e., with the negative partial charge variation within the hydration shell of the metal surface. One can also see from this plot, that term iv is relatively small (less than ~5 kJ mol^{-1}) and attractive for positively charged ions that are localized at an ion-surface distance of 3−4 Å (i.e., when the ion is inserted into the first hydration layer). On the other hand, the electrostatic field of the surface water layer is preferentially attractive for an anion when it is placed slightly beyond the first hydration layer, at 4−5 Å, which corresponds to the maximum of the hydrogen partial density.

The Born solvation energy is dominant at small distances ($z < 4$ Å) for ions but should be negligible for molecules. A charged fragment of a protein adsorption site would be surrounded by neighboring neutral atoms of the protein. This would mean that the electrostatic effect caused by the hydration shell would be less pronounced than for a bare ion because the charge-water distance would be too large to make the magnitude of the effect significant. Thus, the desolvation effect may be represented solely by the metal desolvation term iii due to the distortion of the hydration shell of the metal. Taking into account all these uncertainties and a modest contribution to the free binding energy, we did not implement the term accounting for this effect in the free energy calculations.

References

(1) Slojkowska, R.; Palys, B.; Jurkiewicz-Herbich, M. *Electrochem. Acta* **2004**, *49*, 4109–4118. Prado, C.; Prieto, F.; Rueda, M.; Feliu, J.; Aldaz, A. *Electrochim. Acta* **2007**, *52*, 3168–3180.

(2) Brown, S. *Nat. Biotechnol.* **1997**, *15*, 269–272.

(3) Willett, R. L.; Baldwin, K. W.; West, K. W.; Pfeiffer, L. N. *Proc. Nat. Acad. Sci. U.S.A.* **2005**, *102*, 7817–7822.

(4) Wei, Y.; Latour, R. A. *Langmuir* **2008**, *24*, 6721–6729.

(5) Peelle, B. R.; Krauland, E. M.; Wittrup, K. D.; Belcher, A. M. *Langmuir* **2005**, *21*, 6929–6933.

(6) Krauland, E. M.; Peelle, B. R.; Wittrup, K. D.; Belcher, A. M. *Biotechnol. Bioeng.* **2007**, *97*, 1009–1020.

(7) Li, H.-Q.; Chen, A.; Roscoe, Sh. G.; Lipkowski, J. *J. Electroanal. Chem.* **2001**, *500*, 299–310.

(8) Tamerler, C.; Duman, M.; Oren, E. E.; Gungormus, M.; Xiong, X.; Kacar, T.; Parviz, B. A.; Sarikaya, M. *Small* **2006**, *11*, 1372–1378.

(9) Hnilova, M.; Oren, E. E.; Seker, U. O.; Wilson, B. R.; Collono, S.; Evans, J. S.; Tamerler, C.; Sarikaya, M. *Langmuir* **2008**, *24*, 12440–12445.

(10) Tamerler, C.; Oren, E. E.; Duman, M.; Venkatasubramanian, E.; Sarikaya, M. *Langmuir* **2006**, *22*, 7712–7718.

(11) Gray, J. J. *Curr. Opin. Struct. Biol.* **2004**, *14*, 110–115.

(12) Harding, J. H.; Duffy, D. M.; Sushko, M. L.; Rodger, P. M.; Quigley, D.; Elliot, J. A. *Chem. Rev.* **2008**, *108*, 4823–4854.

(13) Latour, R. A. *Biointerphases* **2008**, *3*, FC2–FC12.

(14) Cohavi, O.; Corni, S.; De Rienzo, F.; Di Felice, R.; Gottschalk, K. E.; Höfling, M.; Kokh, D.; Molinari, E.; Schreiber, G.; Vaskevich, A.; Wade, R. C. *J. Mol. Recog.* **2010**, *23*, 259−262.

(15) Nakanashi, K.; Sakiyama, T.; Imamura, K. *Biosci. Bioing.* **2001**, *91*, 233–244.

(16) Horinek, D.; Serr, A.; Geisler, M.; Pirzer, T.; Slotta, U.; Lud, S. Q.; Garrido, J. A.; Scheibel, T.; Hugel, T.; Netz, R. R. *Proc. Nat. Acad. Sci. U.S.A.* **2008**, *105*, 2842–2847.

(17) Raut, V. P.; Agashe, M. A.; Stuart, S. J.; Latour, R. A. *Langmuir* **2005**, *21*, 1629–1639.

(18) Raffaini, G.; Ganazzoli, F. *Langmuir* **2004**, *20*, 3371–3378. Makrodimitris, K.,; Masica, D. L.; Kim, E. T.; Gray, J. J. *J. Am. Chem. Soc.* **2007**, *129*, 13713–13722.

(19) Heinz, H.; Farmer, B. L.; Pandey, R. B.; Slocik, J. M.; Patnaik, S. S.; Pachter, R.; Naik, R. R. *J. Am. Chem. Soc.* **2009**, *131*, 9704–9714.

(20) Heinz, H.; Vaia, R. A.; Farmer, B. L.; Naik, R. R. *J. Phys. Chem. C* **2008**, *112*, 17281–17290.

(21) Braun, R.; Sarikaya, M.; Schulten, K. J. *Biomater. Sci. Polymer Ed.* **2002**, *13*, 747–757.

(22) Kantarei, N.; Tamerler, C.; Sarikaya, M.; Halilogla, T.; Doruker, P. *Polymer* **2005**, *46*, 4307–4313.

(23) Ghiringhelli, L. M.; Hess, B.; van der Vegt, N. F. A.; Delle Site, L. *J. Am. Chem. Soc.* **2008**, *130*, 13460–13464.

(24) Verde, A. V.; Acres, J. M.; Maranas, J. K. *Biomacromolecules* **2009**, *10*, 2118–2128.

(25) Feig, M.; Brooks, C. L., III. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.

(26) Sun, Y.; Latour, R. *J. Comput. Chem.* **2006**, *27*, 1908–1922.

(27) Iori, F.; Di Felice, R.; Molinari, E.; Corni, S. *J. Comput. Chem.* **2009**, *30*, 1465–1476.

(28) Hoefling, M.; Iori F.; Corni, S.; Gottschalk, K. E. *ChemPhysChem* **2010**, [Online] DOI: 10.1002/cphc.200900981.

(29) Hoefling, M.; Iori, F.; Corni, S.; Gottschalk, K. E. *Langmuir* **2010**, in press.

(30) Gabdoulline, R. R.; Wade, R. C. *Biophys. J.* **1997**, *72*, 1917–1929.

(31) Shkel, I. A.; Kim, S. Two Dimensional Reaction of Biological Molecules Studied by Weighted Ensemble Brownian Dynamics *Proceedings of the 4th Pacific Symposium on Biocomputing*, (PSB '99) Hawaii, January 4−9, 1999; Altman, R. B., Dunker, A. K., Hunter, L., Klein, T. E., Eds. http://psb.stanford.edu/psb-online/ (accessed Apr 20, 2010); pp 468−479.

(32) Goba, C.; Geyer, T.; Helms, V. *J. Chem. Phys.* **2004**, *121*, 457–464.

(33) Tiel, P. A.; Madey, Th. E. *Surf. Sci. Rep.* **1987**, *7*, 211–285. Henderson, M. A. *Surf. Sci. Rep.* **2002**, *46*, 1–308.

(34) Meng, Sh.; Wang, E. G.; Gao, Sh. *J. Chem. Phys.* **2003**, *119*, 7617–7620. Meng, Sh.; Wang, E. G.; Gao, Sh. *Phys. Rev. B* **2004**, *69*, 195404–17.

(35) Michaelides, A.; Ranea, V. A.; de Andres, P. L.; King, D. A. *Phys. Rev. Lett.* **2003**, *90*, 216102–216106.

(36) Criscenti, L. J.; Cygan, R. T.; Kooser, A. S.; Moffat, H. K. *Chem. Mater.* **2008**, *20*, 4682–4693.

(37) Ju, Sh.-P. J. *Chem. Phys.* 2005*122*, 094718−094718−6.

(38) Chang, Ch.-I.; Lee, W.-J.; Young, T.-F.; Ju, Sh.-P.; Chang, Ch.-W.; Chen, H.-L.; Chang, J.-G. *J. Chem. Phys.* **2008**, *128*, 154703−154703−9.

(39) Torrie, G. M.; Kusalik, P. G.; Patey, G. N. *J. Chem. Phys.* **1988**, *88*, 7826–7840. Bérard, D. R.; Kinoshita, M.; Ye, X.; Patey, G. N. *J. Chem. Phys.* **1994**, *101*, 6271–6280. Bérard, D. R.; Kinoshita, M.; Ye, X.; Patey, G. N. *J. Chem. Phys.* **1995**, *102*, 1024–1033.

(40) Guidelli, R.; Schmickler, W. *Electrocheim. Acta.* **2000**, *45*, 2317–2338.

(41) Barten, D.; Kleijn, J. M.; Duval, J.; v. Leeuwen, H. P.; Lyklema, J.; Cohen Stuart, M. A. *Langmuir* **2003**, *19*, 1133–1139. Hillier, A. C.; Kim, S.; Bard, A. J. *J. Phys. Chem.* **1996**, *100*, 18808–18817.

(42) Ataka, K.; Yotsuyanagi, T.; Osawa, M. *J. Phys. Chem.* **1996**, *100*, 10664–10672. Nihonyanagi, S.; Ye, Sh.; Uosaki, K.; Dreesen, L.; Humbert, Ch.; Thiry, P.; Peremans, A. *Surf. Sci.* **2004**, *573*, 11–16.

(43) Schrader, M E. *J. Colloid Interface Sci.* **1984**, *100*, 372–380.

(44) Sendner, Ch.; Horinek, D.; Bocquet, L.; Netz, R. R. *Langmuir* **2009**, *25*, 10768–10781.

(45) Löfgren, P.; Ahlström, P.; Chakarov, D. V.; Lausmaa, J.; Kasemo, B. *Surf. Sci.* **1996**, *367*, L19–L25. Smith, R. S.; Huang, C.; Wong, E. K. L.; Kay, B. D. *Surf. Sci.* **1996**, *367*, L13–L18.

(46) Jorgensen, W. L.; Tirado-Rivers, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

(47) Iori, F.; Corni, S. *J. Comput. Chem.* **2008**, *29*, 1656–1666.

(48) Finnis, M. W. *Surf. Sci.* **1991**, *241*, 61–72. Sahni, V.; Bohnen, K.-P. *Phys. Rev. B* **1985**, *31*, 7651–7661.

(49) Gabdoulline, R. R.; Wade, R. C. *J. Phys. Chem.* **1996**, *100*, 3868–3878.

(50) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; McCammon, J. A. *Comput. Phys. Commun.* **1995**, *9*, 57–95.

(51) Gabdoulline, R. R.; Wade, R. C. *Methods* **1998**, *14*, 329–341.

(52) Gabdoulline, R. R.; Wade, R. C. *J. Mol. Biol.* **1999**, *291*, 149–162.

(53) Davis, M. E. *J. Chem. Phys.* **1994**, *100*, 5149–5159.

(54) Hamann, C. H.; Hamnet, A.; Vielstich, W. *Electrochemistry*; Wiley: New York, 1998.

(55) Gabdoulline, R. R.; Wade, R. C. *J. Am. Chem. Soc.* **2009**, *131*, 9230–9238.

(56) Ben-Tal, N.; Honig, B.; Bagdassarian, C. K.; Ben-Shaul, A. *Biophys. J.* **2000**, *79*, 1180–1187.

(57) Harvey, S. C.; Prabhakaran, M. *J. Phys. Chem.* **1987**, *91*, 4799–4801.

(58) Beutler, T. C.; van Gunsteren, W. F. *J. Phys. Chem.* **1994**, *100*, 1492–1497.

(59) Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(60) Yeh, I. C.; Berkowitz, M. L. *J. Chem. Phys.* **1999**, *111*, 3155–3162.

(61) Donnini, S.; Mark, A. E.; Juffer, A. H.; Villa, A. *J. Comput. Chem.* **2005**, *26*, 115–122.

(62) Trzesniak, D.; Kunz, A.-P. E.; van Gunsteren, W. F. *ChemPhysChem* **2007**, *8*, 162–169.

(63) Kaestner, J.; Thiel, W. *J. Chem. Phys.* **2005**, *123*, 144104. Van Eerden, J.; Briels, W. J.; Harkema, S.; Feil, D. *Chem. Phys. Lett.* **1989**, *164*, 370–376.

(64) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(65) Sedlmeier, F.; Janecek, J.; Sedner, C.; Bocquet, L.; Netz, R. R.; Horinek, D. D. *Biointerphases* **2008**, *3*, FC23–FC39.

(66) Harris, J. G. *J. Phys. Chem.* **1992**, *96*, 5077–5086.

CT100086J

JCTC Journal of Chemical Theory and Computation

Limits of Free Energy Computation for Protein−Ligand Interactions

Kenneth M. Merz, Jr.*

Department of Chemistry, Quantum Theory Project, 2328 New Physics Building,
PO Box 118435, University of Florida, Gainesville, Florida 32611-8435

Received February 20, 2010

Abstract: A detailed error analysis is presented for the computation of protein−ligand interaction energies. In particular, we show that it is probable that even highly accurate computed binding free energies have errors that represent a large percentage of the target free energies of binding. This is due to the observation that the error for computed energies quasi-linearly increases with the increasing number of interactions present in a protein−ligand complex. This principle is expected to hold true for any system that involves an ever increasing number of inter- or intramolecular interactions (e.g., ab initio protein folding). We introduce the concept of best-case scenario errors ($\mathrm{BCS_{errors}}$) that can be routinely applied to docking and scoring studies and that can used to provide error bars for the computed binding free energies. These $\mathrm{BCS_{errors}}$ form a basis by which one can evaluate the outcome of a docking and scoring exercise. Moreover, the resultant error analysis enables the formation of an hypothesis that defines the best direction to proceed in order to improve scoring functions used in molecular docking studies.

Introduction

Since Paul Dirac noted in 1929, "The fundamental laws necessary for the mathematical treatment of a large part of physics and the whole of chemistry are thus completely known, and the difficulty lies only in the fact that application of these laws leads to equations that are too complex to be solved."[1] Theoretical chemistry has evolved to the point that in some instances tractable equations are utilized to create a computational method that can routinely reach what is termed chemical accuracy or ±1 kcal/mol from experiment.[2,3] This accuracy is achieved for small interacting molecular systems, like the water dimer or other related small molecule complexes.[2,3] The extension of this result to macromolecular systems, however, is less clear. In principle, one would like to believe that, as chemically accurate models are used on ever-larger systems, the same level of accuracy would be possible. It is likely, though, that this is not the case and, indeed, we argue below that the expected errors in energies calculated on macromolecular systems are likely not to reach this level of accuracy. However, what level of accuracy would be expected is unclear. In this note, we describe a "gedanken" experiment for protein−ligand scoring that addresses this very issue by delving deeper into the expected errors in energy computation in macromolecules.

Protein−ligand docking and scoring has been an active field of investigation for the last several decades.[4−8] The concept is that given a small molecule compound we can computationally pose or dock it into a receptor site such that we obtain the correct orientation relative to experiment, while simultaneously predicting a binding free energy in good agreement with experiment. This has proven to be a difficult task,[6,7] which is best captured by: "Accurate prediction of binding affinities for a diverse set of molecules turns out to be genuinely difficult".[5] Indeed, extensive validation studies have shown how challenging this problem is,[9−12] but it is still largely uncertain why. Arguments, including sampling,[13] structural water molecules, tautomeric states, and conformational strain[14] have all been put forward as (partial) explanations. Nonetheless, the way to significantly improve the current state-of-the-art still has to be delineated. Extensive work using free energy perturbation or alchemical methods have shown some promise relative to traditional docking approaches[13,15] but still yield results with fairly large errors

* Corresponding author e-mail: merz@qtp.ufl.edu.

10.1021/ct100102q
Published on Web 04/14/2010

in terms of both the binding orientation and the free energy of binding in prospective or blind studies.[13]

Types of Errors. When deciding upon what type of error model to use, there are two extremes that need to be considered. The first is determinate or systematic errors, which are errors that have a value that can be assigned and corrected for when obtaining insight into the reliability of a measurement. The second extreme is random or indeterminate errors whose sources are not certain, do not have a definite value, but do fluctuate in a random way. In each case, it is possible to propagate the errors over a series of measurements or in our case interactions in, for example, a protein−ligand complex. In the present work, we will assume that we will be accumulating errors via sums of interactions, hence, systematic errors are propagated as a simple sum of the individual errors in the interactions, while random errors are accumulated as the square root of the sum of the squares of the individual errors.[16] The sum of errors used to propagate systematic errors can also be shown to represent the upper limit for random errors as well.[16]

The Variation Principle and Error. The variation principle, given as

$$\frac{\int \Phi H \Phi}{\int \Phi\Phi} \geq E_o \tag{1}$$

where Φ is the wave function, and H is our Hamiltonian. This principle states that as the wave function Φ is improved at a defined Hamiltonian H (e.g., Hartree−Fock) that we should asymptotically approach the ground-state energy E_o appropriate to that Hamiltonian.[17] Hence, for the computation of E_o, we should expect the error to be above the "true" E_o, suggesting that we are dealing with a systematic error. Critically, for the present analysis, if we compute the interactions embodied within a protein−ligand complex using a variational method, then we would expect our computed interaction energy to be above or below the expected value, since we do not know the magnitude of the difference (or error) between the individual E_o's computed for the interacting partners. This, of course, goes away when the equality in eq 1 is satisfied, and the true E_o is reached. Hence, all interactions computed by a series of variational methods (e.g., HF/aug-cc-pVTZ, CASSCF, full CI, etc.) can be described as having a random error relative to our reference interaction energy value. Thus, we conclude that for variational methods, error accumulation appears to behave as if we are working with a systematic or determinate error for the electronic energy, E_o, while for the computation of the interaction energy, these methods behave as if one is dealing with a random error.

Variational methods are very particular in the sense that most modern computational methods are nonvariational. This includes force fields, semiempirical QM (e.g., AM1, PM3, SCCDFTB, etc.), density functional theory (DFT), MPX (X = 2−4) theory, and coupled cluster theory (e.g., CCSD(T)).[17] Hence, in contrast to variational methods, nonvariational methods, in principle, are expected to display random errors (in that we do not know if the computed energy is above or below E_o) for both the computation of E_o and the interaction energy that is propagated as the square root of the sum of squares (see eq 9 below), rather than as a simple sum. Interestingly, the accumulation of systematic errors will yield a larger overall error for E_o than the random analysis given the same absolute magnitudes of the individual errors. The difference between systematic and random uncertainties has to do with cancellation of errors that can happen when one deals with random errors. Nonetheless, as noted above, it can be proven that the sum of error model represents the upper limit for random errors.[16] Hence, the use of more theoretically grounded variational methods over nonvariational methods does not appear to offer a benefit when it comes to error propagation or the computation of interaction energies.

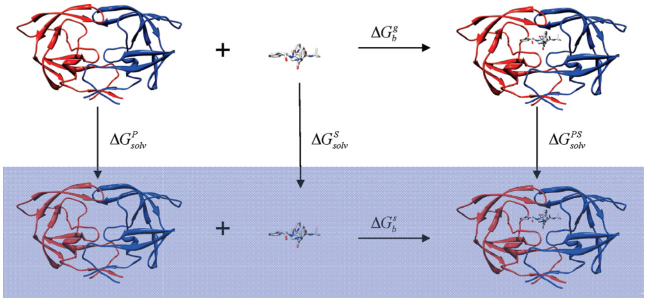


Figure 1. Thermodyamic cycle to estimate the free energy of binding of a drug molecule to a protein receptor. P = Protein, S = small molecule/substrate, and PS = protein-small molecule/substrate complex.

Protein−Ligand Binding Free Energies. Consider the standard thermodynamic cycle shown in Figure 1.[15,18] Using this, we can write the following standard expressions:

$$\Delta G_b^s = \Delta G_b^g + \Delta G_{solv}^{PS} - \Delta G_{solv}^P - \Delta G_{solv}^S \tag{2}$$

where ΔG indicates free energy changes in the gas-phase (ΔG^g) or solution (ΔG^s), and the subscript b denotes binding. Terms associated with changes in the solvation free energy are also indicated by solv. What we want to do next is to break this down into individual components of the total energy, enthalpy, and entropy in a compact form, which we briefly outline below. Expanding out the free energy of binding in the gas-phase term and consolidating the solvation free energies, we can write:

$$\Delta G_b^s = (E_{total}^{PS} + H_{corr}^{PS}) - TS^{PS} - ((E_{total}^P + H_{corr}^P) - TS^P + (E_{total}^S + H_{corr}^S) - TS^S) + \Delta\Delta G_{sol} \tag{3}$$

where the E_{total} terms are the individual electronic energies, H_{corr} are the individual enthalpy correction terms to the electronic energies, and the S terms are the respective the entropies. The latter two terms can be computed using the rigid-rotor harmonic approximation.[19,20] $\Delta\Delta G_{sol}$ is given as:

$$\Delta\Delta G_{sol} = \Delta G_{solv}^{PS} - \Delta G_{solv}^P - \Delta G_{solv}^S \tag{4}$$

and the individual terms can be obtained from explicit or implicit solvation models or experiment.[21,22] Rearranging and collecting terms together that represent the change in the total energy, enthalpy, and entropy, respectively, we obtain:

$$\Delta G_b^s = (E_{\text{total}}^{PS} - (E_{\text{total}}^{P} + E_{\text{total}}^{S})) + (H_{\text{corr}}^{PS} - (H_{\text{corr}}^{P} + H_{\text{corr}}^{S})) - (\text{TS}^{PS} - (\text{TS}^{P} + \text{TS}^{S})) + \Delta\Delta G_{\text{sol}} \quad (5)$$

which, when consolidated, leads us to a compact representation[23] of the four key terms we have to evaluate in order to obtain the free energy of binding of a ligand to a protein in aqueous solution:

$$\Delta G_b^s = \Delta E_{\text{int}}^{PS} + \Delta H_{\text{corr}}^{PS} - T\Delta S^{PS} + \Delta\Delta G_{\text{sol}} \quad (6)$$

where the first three of the individual terms are defined as:

$$\begin{aligned} \Delta E_{\text{int}}^{PS} &= E_{\text{total}}^{PS} - (E_{\text{total}}^{P} + E_{\text{total}}^{S}) \\ \Delta H_{\text{corr}}^{PS} &= H_{\text{corr}}^{PS} - (H_{\text{corr}}^{P} + H_{\text{corr}}^{S}) \quad (7) \\ T\Delta S^{PS} &= \text{TS}^{PS} - (\text{TS}^{P} + \text{TS}^{S}) \end{aligned}$$

For our purposes, a key result of this simple analysis is that the master equation (eq 6) yields a ΔE term, which indicates the change in the electronic energy of a molecule in the gas-phase. This term, along with the second term in eq 6 (the ΔH term), are quantities for which modern electronic structure theory can obtain highly reliable values for using appropriate methodologies.[2,19] Given that our analysis yields terms involving the change in the electronic energy and the vibrational enthalpy correction term upon binding in the gas-phase, we can ask ourselves how can we make use of this in an error analysis? Herein, we make the assumption that the change in the electronic energy and the vibrational enthalpy correction term can be estimated as a linear combination of the individual interactions. We define "individual interaction" not in a pairwise or atom-by-atom sense but in a chemical sense. Hence, two carbon atoms interacting does not satisfy our definition, but an hydroxyl hydrogen bonding to a carbonyl or a valine side chain forming a van der Waals complex with a phenyl ring from an inhibitor does fit our definition. Furthermore, we can envision these interactions as not being isolated in the gas-phase but as being in the context of the protein−ligand environment in a combined quantum mechanical/molecular mechanical (QM/MM) sense.[17] This is an approximation, but as noted by Zhou and Gilson,[24] it has some justification, while doing the same for the entropy term (ΔS) is not due to standard state concentration and translational entropy considerations. Experimentally, this is borne out by recent work of Klebe and co-workers,[25] that shows for a series of thrombin inhibitors the overall free energy and entropy terms show significant cooperative effects, but the enthalpy term alone, within the experimental error bars, shows little or no cooperativity. Thus we can approximate interaction energy and enthalpy terms as

$$\Delta E_{\text{int}}^{PS} + \Delta H_{\text{corr}}^{PS} = \Delta H_{\text{int}}^{PS} \approx \Delta H_{\text{int 1}}^{PS} + \Delta H_{\text{int 2}}^{PS} + \Delta H_{\text{int 3}}^{PS} + ... \quad (8)$$

The score function given in eq 6 has a total of four terms in the expression, and we have discussed two of the four (ΔE and ΔH). For the TΔS, terms we cannot decompose them,[24] which presently leaves us at a loss regarding how to best estimate the expected overall error for this term. Intuitively, we expect the error to be larger given the need for extensive sampling and the accurate computation of translational, rotational, and vibrational components of the total entropy, but ultimately quantitatively expressing this error contribution will be essential.[20,24] Hence, at this juncture, estimating the expected uncertainty in the entropy component would involve significant guesswork and will be left for future analysis.

The approximation of eq 8 assumes that each interaction pair, while "seeing" all other interactions (in a QM/MM sense), behaves independently from the other pairs in terms of their overall interaction energy, vibrational enthalpy correction, and their associated errors. For two "interaction pairs" that are far away from each other, this is not an unreasonable approximation, but for two interaction pairs that are closer to one another, one might imagine this to be more of a significant approximation especially if QM effects, like polarization and charge transfer, play a role. For each of these *intermolecular* interactions, we can compare various computational approaches for obtaining interaction energies or enthalpies with experiment or with "chemically" accurate quantum mechanical methods in order to obtain an error estimate for each interaction. Hence, the hypothesis here is that through the careful analysis of these interactions, we can identify errors in computationally less expensive methods and improve them to derive better simple models that can be used in more extensive drug design applications. Assuming that we are dealing with random errors, we can propagate the errors using the following expression:

$$\text{Error}_{\Delta H_{\text{int}}^{PS}} = [(\text{Error}_{\Delta H\text{int 1}})^2 + (\text{Error}_{\Delta H\text{int 2}})^2 + (\text{Error}_{\Delta H\text{int 3}})^2 + ...]^{1/2} \quad (9)$$

While it is attractive to think about doing this computationally[26] or experimentally, the fact is this is difficult to do, so initially it is better to examine the boundary conditions of our approximations and the resultant ramifications using a gedanken experiment. Let us pick a concrete example of Indinavir (crixivan, $K_i = 0.358$nM or −12.8 kcal/mol binding free energy)[27,28] bound to the HIV-1 protease (PDB ID: 1HSG).[29] The LigPlot[30] diagram is given in Figure 2 highlighting both hydrophobic and hydrophilic contacts. From the LigPlot analysis and a graphical one, 18 hydrophobic contacts were identified (Gly 48 and Gly 49 on chains A and B are involved in carbonyl−H−C interactions),[31] and 6 hydrogen bonds along with 5 hydrogen bonds between the ligand and crystallographic waters for a total of 29 individual protein−ligand contacts in our "pharamcophore" that we imagine, for the sake of our gedanken argument, that we have obtained experimental values for each individual interaction energy (29 total) in the context of the ligand. It may not be possible to do this experimentally, but again we are constructing a gendaken experiment to help us understand the boundary conditions of our error hypothesis. Further let us imagine that we will test the performance of two computational methods (these could be ab initio, semiempirical, or force-field-based calculations) for their ability to compute interaction energies and that they are found to have

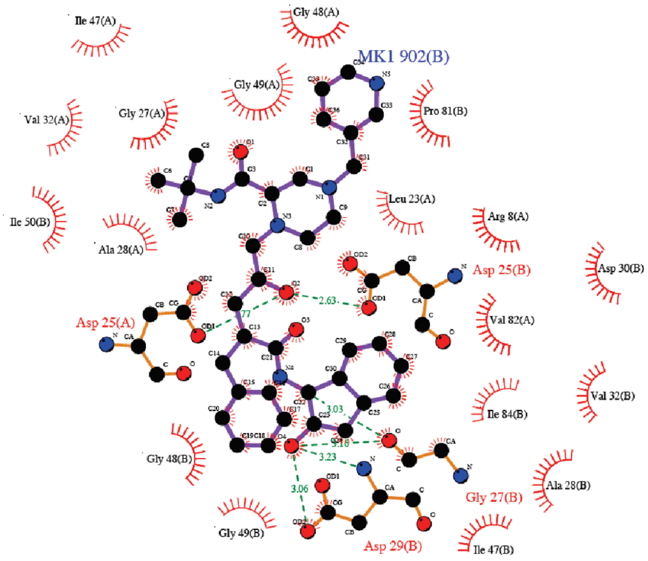


Figure 2. The LigPlot diagram of Indinavir bound to HIV-1 protease.

error bars relative to our experimental values for each enthalpy of interaction of ±1 and ±0.5 kcal/mol, respectively. By each enthalpy of interaction, we mean chemically distinct pairs of molecules that, in part, represent each of the 29 interactions identified in this protein−ligand complex. For example, the side chain of Val33 places a methyl group into the face of an aromatic ring of the inhibitor (in Figure 2, the ring adjacent to the MK1 label). This interaction along with the remaining 28 we assume to each have individual errors of ±1 or ±0.5 kcal/mol with respect to experiment to give us a range of individual errors to evaluate when we do our error propagation. The use of this error range per interaction was chosen because this can be achieved, with great effort, using modern converged QM calculations and, hence, represents error bars that could be realized today.[2] Error varies along the reaction coordinate for the formation of each individual interaction,[26] so by error, we specifically mean the deviation from an established value at the minimum.

In the two cases outlined above, the total error between experiment and the two model Hamiltonians is ±5.4 kcal/mol for the case where we have 29 interactions each with an error of ±1 and ±2.7 kcal/mol for the case where we have 29 interaction each with an error of ±1 kcal/mol both propagating the error, as given by eq 9. An individual error of ±1 to ±0.5 kcal/mol is quite good, but imagine if one of the 29 interactions is off by ±5 kcal/mol with respect to experiment, while the remaining 28 stay at ±1 and ±0.5 kcal/mol, respectively. This leads to predicted errors of ±7.3 and ±5.6 kcal/mol for the two cases. In these cases, the one badly modeled interaction makes the single largest contribution to the error, suggesting that by simply improving this one bad interaction we can significantly improve the approximate model. This points out one of the advantages of this approach in that we can focus on interaction classes that are poorly modeled and expend our parametrization efforts on these problem areas first, in order to realize the largest improvement in simpler models.

The next term that we need to consider is the change in the solvation free energy, $\Delta\Delta G_{\text{sol}}$. From eq 4, we see that the solvation term consists of terms involving the solvation free energy of the ligand, the protein, and the protein−ligand complex. First we will consider the expected error for the solvation free energy of the ligand and then explore how to estimate this for the protein calculations. For small drug-like fragments, the SAMPL prospective[32−34] validation has demonstrated that the deviation from experiment is of the order of ±1.8 to ±2.5 kcal/mol. For the sake of our analysis, we propose that an expected error of ±2.0 kcal/mol for the computed value for the solvation free energy of Indinavir is appropriate.

The protein and protein−ligand cases are trickier to evaluate, but in the present case, we only need to know how many side chains are exposed in the protein and the protein−ligand complex. This has to do with the way in which these methods work. They are surface or reaction-field-based algorithms,[21,22] and the expectation is that the surface residues will have the greatest impact on the solvation free energy, while the buried residues contribute significantly less. In the free protein, we estimate that there are 100 exposed side chains, which is reduced to 88 upon complexation of Indinavir. However, the only part of the protein that matters is the part that becomes buried (or in other words is experiencing a change) upon ligand complexation in an error analysis because the remaining exposed residues contribute a constant error that is canceled out in the final expression for $\Delta\Delta G_{\text{sol}}$. Thus, in order to better estimate the error in the solvation free energy calculations for the protein and protein−ligand complex, we adopt the approach where only the associated buried (exposed) part of the protein is considered in the complexation process. In the case of 1HSG, we estimate that 12 active site residues are buried (or exposed) as the result of the complexation (disassociation) process.

How can we estimate the expected error in our solvation free energy computations? We propose that we can adopt an approach similar to what we did for the interaction energy component of the score function in that we view the solvation process and being broken down into individual solvation free energy calculations for each amino acid side chain.

$$\Delta G_{\text{solv}}^{\text{PS}} \approx \Delta G_{\text{solv}}^{\text{int 1}} + \Delta G_{\text{solv}}^{\text{int 2}} + \Delta G_{\text{solv}}^{\text{int 3}} + \ldots \qquad (10)$$

Hence, the problem becomes a case of estimating the expected error for the solvation free energy of each individual side chain that becomes buried or exposed. A broad range of solvation models have been proposed[21,22] each with its own average error from experiment, but the very best models reach errors on the order of ±1.0 kcal/mol or less for small neutral species and above this for charged molecules. Given that we have a mixture of charged and neutral species in the present case, we will adopt ±1 kcal/mol as a good estimate for the individual errors given in eq 10. We note that we are using a different error for this situation when compared to the error in the solvation free energy of the ligand molecule (estimated as ±2.0 kcal/mol above). The reason for this is that we view the side chains to be simple organic molecules, like benzene (Phe), phenol (Tyr), propane (Val), etc., for which typical solvation models[21,22] give smaller errors than for larger drug-like molecules analyzed, for example, in the SAMPL prospective study.[33] Larger or smaller errors can be adopted, and the reader can do the simple mathematics to see how this affects the outcome of the error estimation. For the protein−ligand case, we assume the error is ∼±0.0 kcal/mol because, as noted above, only the residues that are buried or exposed contribute to the overall error estimation. This is not to say that the surface residues that are solvent exposed in both the free and complexed protein state have no error, it simply acknowledges that because these exposed residues appear in both of these states that their errors largely cancel. For the free protein, we estimate the error to be ∼±3.5 kcal/mol, which is the square root of 12 (12 residues being buried/exposed each with a solvation free energy error of ±1 kcal/mol). Finally, the ligand is proposed to have a solvation free energy error of ±2.0 kcal/mol, as outlined above. To estimate the total error for the $\Delta\Delta G_{\mathrm{sol}}$ term in eq 6, we take the three components (protein−ligand = ∼±0.0, protein ∼±3.5, and ligand ±2.0 kcal/mol), square them, and take the square root to yield a total error estimate of ±4.0 kcal/mol.

Putting all of this together, we can now estimate a lower bound for the expected error range using eq 6 to compute binding free energies. It is a lower limit because we have not attempted to estimate the expected errors in the entropy component of eq 6. The estimated error for the electronic energy and enthalpic term was ±5.4 kcal/mol (assuming ±1 kcal/mol error for each of the 29 individual the computed interaction energies) and ±2.7 kcal/mol (assuming a ±0.5 kcal/mol uncertainty), while for the solvation term, we arrived at a value of ±4.0 kcal/mol. Propagating these two terms again as the square root of the sum of the squares yields values of ±6.7 and ±4.8 kcal/mol as the expected uncertainty computed given the parameters we have chosen. For Indinavir, the experimental binding free energy is −12.8 kcal/mol[27,28] so our estimated errors represent up to one-half of the binding free energy of this molecule. Put another way, if we use a computational scoring function that has the error parameters described above and it predicts that the binding free energy of Indinavir is 1nM or −12.3 kcal/mol, then we would have an uncertainty that would suggest that our predicted binding affinity actually ranges from better than picomolar to submicromolar, which is a wide binding affinity range. In the submicromolar case, you would decide not to further study the molecule in question, while in the better than picomolar instance, the compound would almost certainly be further examined.

Discussion

From this analysis we estimate that the error in our model scoring function with its associated uncertainties would yield not particularly satisfactory results for an absolute binding free energy determination. The choice of Indinavir is a real challenge since this molecule is quite large and has many protein−ligand contacts one has to consider, so for simpler molecules or cases where fewer contacts are present, one would expect the error to decrease. Thus, one important conclusion from this analysis is that as the molecular size increases or the number of contacts increases, we would expect a quasi-linear increase in the expected error. In retrospect, this is not that unexpected of a conclusion.

Does our conclusion bear up against the current results available in the literature? In a detailed recent analysis of docking success Kolb and Irwin concluded that: “When they work do docking screens really discover ligands for the right reasons? For simple models systems with very small ligands, docking appears to work amazingly well.”[7] They go on to discuss that for large drug-like molecules success has been tough to come by. From our analysis, we would conclude that the uncertainty in studies of small molecules interacting with a receptor would be significantly less than that for a molecule like Indinavir. Thus, an important conclusion of Kolb and Irwin’s work and the present analysis is that working with small ligand scaffolds or fragments as done in fragment drug design[35] is a more prudent approach than doing in silico screens through large numbers of higher molecular weight drug-like molecules. Success with smaller ligands is also seen in the extensive and careful work of Gilson and co-workers on host guest systems, where they have been able to achieve errors in the 1−2 kcal/mol range for the estimation of binding free energies.[15,36,37] Shoichet, Dill, and co-workers have also had good success with free energy perturbation methodologies when *prospectively* applied to small aromatic organic molecules binding to an engineered binding site in T4 lysozyme,[38,39] where they realized errors of ∼2 kcal/mol. The most complex system reported on to date are the *retrospective* FKBP studies of Roux and co-workers[40] and Pande and Shirts and co-workers.[41,42] In these studies, errors for the prediction of the binding free energy for a series of 8 FKBP inhibitors was between 1.4−2.5 kcal/mol depending on the choice of protocols utilized. This level of agreement is outstanding, especially for the larger molecules studied in these efforts. In a final *prospective* study carried out by Roux and co-workers[13] on 50 compounds to JNK kinase, the agreement was far less satisfactory, but Roux and his team noted that longer simulations might be needed to obtain converged results.

Overall, our hypothesis is borne out by the available literature, with the possible exception of the FKBP results. However, clearly more work of this sort is needed to sort out the issues raised herein. An interesting question arises with regards to thermodynamic cycle-free energy perturbation (TC-FEP) methods.[43−46] These have been very successful in examining relative free energies for many series of compounds, which appears to fly in the face of our error hypothesis. However, upon careful inspection it is clear that relative methods, like TC-FEP, have distinct advantages that reduce the expected error. Taking the TC shown in Scheme 1, we arrive at the following well-known relationship:

$$\Delta G_{\mathrm{bind}}(\mathrm{IH}) - \Delta G_{\mathrm{bind}}(\mathrm{ICH_3}) = \Delta G_{\mathrm{sol}}(\mathrm{IH} - \mathrm{ICH_3}) - \Delta G_{\mathrm{bind}}(\mathrm{IH} - \mathrm{ICH_3}) \quad (11)$$

This relates that the relative binding free energy can be computed via two alchemical transformations rather than by computing two absolute free energies (ΔG_{bind}(IH) and

Scheme 1. Thermodynamic Cycle

$$\begin{tikzcd}
E + \text{I-H} \arrow[r, "\Delta G_{\text{bind}}(\text{IH})"] \arrow[d, "\Delta G_{\text{sol}}(\text{IH-ICH}_3)"'] & E{\cdot}\text{I-H} \arrow[d, "\Delta G_{\text{bind}}(\text{IH-ICH}_3)"'] \\
E + \text{I-CH}_3 \arrow[r, "\Delta G_{\text{bind}}(\text{ICH}_3)"] & E{\cdot}\text{I-CH}_3
\end{tikzcd}$$

$\Delta G_{\text{bind}}(\text{ICH}_3)$) and taking their difference. For the first term on the right-hand side of eq 11 ($\Delta G_{\text{sol}}(\text{IH}-\text{ICH}_3)$), we simulate the conversion of the free ligand in aqueous solution by converting a hydrogen (H) into a methyl (CH_3) group. This involves the alteration of only one interaction site exposed to solvent, and as a result, the error in this computation is likely to be quite small, but for the sake of argument we will assume it is ±1 kcal/mol. For the conversion of a hydrogen atom into a methyl group within the protein ($\Delta G_{\text{bind}}(\text{IH}-\text{ICH}_3)$), this involves the conversion of only one interaction type, while all remaining ones remain the same. Hence, we can hypothesize that we can realize an error for this conversion of ±1 to ±0.5 kcal/mol from experiment based on previous experience using this method.[43-47] Thus, the fully propagated errors (assuming random errors) are ±1.12 to ±1.4 kcal/mol. The expected error for the absolute free energy of binding computations is expected to be much larger (as outlined above), but through the clever use of a TC, we cancel out many of the errors giving results that can be in excellent agreement with experiment. In a related approach where a ligand "core" is fixed and where R groups are systematically added is another way in which error can be reduced because in this model changes in the R group are incurring errors, while the fixed "core" represents a constant error that can be ignored when looking at the relative efficacy of ligands in this so-called congeneric series. Clearly, developing technologies or approaches based off of relative energy computations can afford significant error reductions assuming that major conformational changes are not realized as a result of the perturbation. If the latter occurs, more interactions come into play increasing the expected uncertainty.

Sampling is one of the two major issues facing computational biology along with accurate energy computation. For complex systems that undergo many conformational changes, thorough sampling is an absolute necessity along with accurate energy computations. In many cases by carrying out a molecular dynamics (MD) simulation or via exhaustive sampling of a protein−ligand complex, better estimates of the binding free energy can be obtained even though only a local sampling in the binding pocket is realized, while no major conformational changes are observed.[23,46,48] Furthermore, carrying out consensus scoring studies across a range of scoring functions can achieve a similar result.[49,50] Using our error analysis, it is clear these effects have their roots in the fact that we are dealing with random errors in typical score functions, as also noted by Wang and Wang[51] in a purely statistical docking/scoring experiment. If we take one docking pose, the positioning may optimize or minimize the expected uncertainty, but both estimates are misleading and by sampling over many local conformations, we can obtain a Gaussian distribution in the error function and thereby minimize the uncertainty.[16,51] Hence, by MD or exhaustive sampling we are achieving the same outcome that replicate analytical measurements do when calibrating an instrument subject to random errors. Based on this idea, it would be always prudent to consensus score over a range of local "poses" (generated by MD simulations, for example) after docking a given protein−ligand complex in order to obtain the best estimate for the computed binding energy for a given score function.[50] If large-scale conformational changes are important, then this would not be particularly beneficial, but sampling a ±1 Å root-mean-square deviation around the initial pose could afford some benefit. This is what consensus scoring[12,48-51] and MM-PBSA-like methods[18,23] have achieved with positive results.

Conclusions

Through simple error propagation analysis we show that as the size of a molecule increases, the expected error in free energy computation for protein−ligand complexation would increase quasi-linearly to a point that the error bar is a large fraction of the expected binding free energy. This conclusion is largely supported by the available literature, which shows that, for small molecules, docking and scoring or absolute free energy computation does yield excellent results, but that as the system size increases, the evidence for success in the use of these methods is sparse. Usually, sampling effects are put forth as the major reason for this behavior, and the present analysis *absolutely* does not eliminate this as a serious issue faced when using these techniques. It is clearly important to sample extensively, but the modeling of complex phenomenon also requires that the energy be computed with extreme accuracy, as shown via our simple error analysis. Moreover, we suggest that sampling local structure around an initial pose would have benefits like those seen in consensus scoring.[49]

Through the use of the principles outlined herein, one can formulate an error estimate for the free energy of binding of any given protein−ligand docking pose. We term this the best-case scenario error ($\text{BCS}_{\text{error}}$). By simply counting up the number of interactions being made as a result of complexation and through an estimate of the burial of residues within an active site, error estimates for three of the four terms given in eq 6 can be produced, which ultimately, via further error propagation yields an estimate of the error in the predicted free energy of binding. In the preceding, we gave error estimates that we think represents the "best case scenario", but one can use any set of values one prefers based on personal biases. The utility of this comes when it comes time for data reduction from a large scale docking effort. Many hits are reported, and via estimates of the expected error for each case in a hit list better "educated" decisions can be made.

The result of the present analysis provides an interesting hypothesis that, in principle, should allow us to better understand how to make the computation of absolute binding free energy more robust from an energetic perspective. The error propagation of the enthalpy of binding and the solvation free energy provides a framework from which we can address the errors expected using advanced quantum chemical techniques to force fields. This approach can be applied to related problems that involve the formation of multiple interactions, like protein-folding and the protein−ligand problem highlighted herein. However, due to the complexity (dependence on standard-state concentration) of the entropic part, the use of the present approach does not afford a clear way in which we can get a better grasp of the errors expected in the computation of entropy. This is a subject for ongoing analysis.

Acknowledgment. K.M.M. thanks the National Institutes of Health (GM044974 and GM066689) for funding the present research. The present work is the result of numerous discussions with a wide range of students and colleagues. Helpful conversations with Dr. Xiao He, Mr. John Faver, Professor Raphael Haftka, Professor Adrian Roitberg, Professor Erik Deumens, Professor Barry Honig, Dr. Marti Head, and Dr. Greg Warren are gratefully acknowledged.

References

(1) Dirac, P. A. M. *Proc. R. Soc. London, Ser. A* **1929**, *123*, 714–733.

(2) Helgaker, T.; Klopper, W.; Tew, D. P. *Mol. Phys.* **2008**, *106*, 2107–2143.

(3) Bartlett, R. J.; Musial, M. *Rev. Mod. Phys.* **2007**, *79*, 291–352.

(4) Brooijmans, N.; Kuntz, I. D. *Annu. Rev. Biophys. Biomolec. Struct.* **2003**, *32*, 335–373.

(5) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. *J. Med. Chem.* **2006**, *49*, 5851–5855.

(6) Morra, G.; Genoni, A.; Neves, M. A. C.; Merz, K. M.; Colombo, G. *Curr. Med. Chem.* **2010**, *17*, 25–41.

(7) Kolb, P.; Irwin, J. J. *Curr. Top. Med. Chem.* **2009**, *9*, 755–770.

(8) Guvench, O.; MacKerell, A. D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 56–61.

(9) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. *J. Med. Chem.* **2006**, *49*, 5912–5931.

(10) Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 601–619.

(11) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. *J. Med. Chem.* **2007**, *50*, 726–741.

(12) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins* **2003**, *52*, 609–623.

(13) Deng, Y. Q.; Roux, B. *J. Phys. Chem. B* **2009**, *113*, 2234–2246.

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

(15) Gilson, M. K.; Zhou, H. X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

(16) Taylor, J. R. *An Introduction to Error Analysis. The Study of Uncertainties in Physical Measurements*; University Science Books: Sausalito, CA, 1997.

(17) Cramer, C. J. *Essentials of Computational Chemistry*; John Wiley & Sons: New York, NY, 2002.

(18) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. *Acc. Chem. Res.* **2000**, *33*, 889–97, 3rd.

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

(20) McQuarrie, D. A. *Statistical Thermodynamics*; Haroer & Row: New York, NY, 1973.

(21) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161–2200.

(22) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3093.

(23) Kuhn, B.; Kollman, P. A. *J. Med. Chem.* **2000**, *43*, 3786–91.

(24) Zhou, H. X.; Gilson, M. K. *Chem. Rev.* **2009**, *109*, 4092–107.

(25) Baum, B.; Muley, L.; Smolinski, M.; Heine, A.; Hangauer, D.; Klebe, G. *J. Mol. Biol.* **2010**, *397*, 1042–1054.

(26) Molnar, L. F.; He, X.; Wang, B.; Merz, K. M. *J. Chem. Phys.* **2009**, *131*.

(27) Dorsey, B. D.; Levin, R. B.; McDaniel, S. L.; Vacca, J. P.; Guare, J. P.; Darke, P. L.; Zugay, J. A.; Emini, E. A.; Schleif, W. A.; Quintero, J. C.; et al. *J. Med. Chem.* **1994**, *37*, 3443–51.

(28) Vacca, J. P.; Dorsey, B. D.; Schleif, W. A.; Levin, R. B.; McDaniel, S. L.; Darke, P. L.; Zugay, J.; Quintero, J. C.; Blahy, O. M.; Roth, E.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 4096–100.

(29) Chen, Z.; Li, Y.; Chen, E.; Hall, D. L.; Darke, P. L.; Culberson, C.; Shafer, J. A.; Kuo, L. C. *J. Biol. Chem.* **1994**, *269*, 26344–8.

(30) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. *Protein Eng.* **1995**, *8*, 127–34.

(31) Vargas, R.; Garza, J.; Dixon, D. A.; Hay, B. P. *J. Am. Chem. Soc.* **2000**, *122*, 4750–4755.

(32) 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*, 769–79.

(33) Guthrie, J. P. *J. Phys. Chem. B* **2009**, *113*, 4501–7.

(34) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 4538–43.

(35) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. *J. Med. Chem.* **2008**, *51*, 3661–80.

(36) Moghaddam, S.; Inoue, Y.; Gilson, M. K. *J. Am. Chem. Soc.* **2009**, *131*, 4012–21.

(37) Chen, W.; Chang, C. E.; Gilson, M. K. *Biophys. J.* **2004**, *87*, 3035–49.

(38) Boyce, S. E.; Mobley, D. L.; Rocklin, G. J.; Graves, A. P.; Dill, K. A.; Shoichet, B. K. *J. Mol. Biol.* **2009**, *394*, 747–63.

(39) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. *J. Mol. Biol.* **2007**, *371*, 1118–34.

(40) Wang, J.; Deng, Y.; Roux, B. *Biophys. J.* **2006**, *91*, 2798–814.

(41) Jayachandran, G.; Shirts, M. R.; Park, S.; Pande, V. S. *J. Chem. Phys.* **2006**, *125*.

(42) Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C. D.; Shirts, M. R.; Sorin, E. J.; Pande, V. S. *J. Chem. Phys.* **2005**, *123*, 084108.

(43) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395–2417.

(44) Knight, J. L.; Brooks, C. L. *J. Comput. Chem.* **2009**, *30*, 1692–1700.

(45) Huang, N.; Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5166–5177.

(46) Foloppe, N.; Hubbard, R. *Curr. Med. Chem.* **2006**, *13*, 3583–3608.

(47) Jorgensen, W. L. *Acc. Chem. Res.* **2009**, *42*, 724–733.

(48) Wang, R.; Lu, Y.; Wang, S. *J. Med. Chem.* **2003**, *46*, 2287–303.

(49) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. *J. Med. Chem.* **1999**, *42*, 5100–5109.

(50) Oda, A.; Tsuchida, K.; Takakura, T.; Yamaotsu, N.; Hirono, S. *J. Chem. Inf. Model.* **2006**, *46*, 380–91.

(51) Wang, R.; Wang, S. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422–6.

CT100102Q