JCTC Journal of Chemical Theory and Computation

# Properties and Permeability of Hypericin and Brominated Hypericin in Lipid Membranes

Emma S. E. Eriksson,[†,‡,||] Daniel J. V. A. dos Santos,[§] Rita C. Guedes,[§] and Leif A. Eriksson*[,†,||]

*Örebro Life Science Center, School of Science and Technology, Örebro University, Fakultetsgatan 1, 70182 Örebro, Sweden, Modelling and Simulation Research Center, Örebro University, Örebro, Sweden, Department of Pharmacy, University of Lisbon, 1649-019 Lisbon, Portugal, and School of Chemistry, National University of Ireland, University Road, Galway, Ireland*

Received May 25, 2009

**Abstract:** The promising photosensitizing properties of hypericin, a substituted phenanthroperylene quinone naturally found in Saint John's wort, has led to the proposal that it can be utilized in photodynamic therapy. Structurally modified derivatives are at the present time being investigated to generate a more effective hypericin photosensitizer. Neither the detailed mechanism behind the powerful action of hypericin, arising as a result of light excitation, nor the intracellular localization and transportation is still fully understood. In the present work, molecular dynamics simulations have been performed to study the properties and the permeability of hypericin and modifications thereof, substituted with one or four bromine atoms, in a dipalmitoylphosphatidylcholine lipid membrane. The molecules were found to accumulate in the most dense region of the lipids due to competing interactions with the hydrophobic lipid interior and the polar aqueous environment. This was confirmed by analyzing the radial distribution functions and by the density profiles of the system components. Calculated free energy profiles display large negative changes in free energy for the transport process of the molecules into the lipids, which also support this finding. Permeability coefficients show overall fastest diffusion in the membrane system for hypericin containing one bromine.

## 1. Introduction

**1.1. Properties of Hypericin.** Hypericin (Figure 1) is a phenanthroperylene quinone substituted with hydroxyl and alkyl groups that was first isolated from Saint John's wort (*Hypericum perforatum*) in 1911.[1] However, long before that, this plant was used in therapy as an antidepressant and in wound healing. More recently it has been shown that hypericin possesses toxicity against viruses such as hepatitis B[2], herpes,[3,4] and human immunodeficiency virus (HIV).[3,5−7] The molecule also displays antitumor activity, demonstrated both in vitro[8−10] and in vivo.[11−15] Both the antiviral and antitumor properties have been observed in the presence of light, and the chromophoric system along with the hydroxyl and alkyl substitution makes the molecule an efficient photosensitizer. This implies that the molecule might be used in photodynamic therapy (PDT), a three-component method in which a combination of light, administrated drug (a photosensitizer), and oxygen is required. PDT was first used in the 1970s and is now a promising treatment method of cancer and viral diseases. Tumors that are possible targets of this kind of therapy need to be more or less superficially located, either in the skin tissue or close beneath, to enable light penetration. Tumors located in cavities, such as the sinuses or stomach, are also treated using directed light.

* Corresponding author. E-mail: leif.eriksson@nuigalway.ie.

[†] Örebro Life Science Center, School of Science and Technology, Örebro University.

[‡] Modelling and Simulation Research Center; Örebro University.

[||] School of Chemistry, National University of Ireland.

[§] iMED.UL, CECF, Department of Pharmacy, University of Lisbon.

10.1021/ct9002702 CCC: $40.75 © 2009 American Chemical Society
Published on Web 11/09/2009

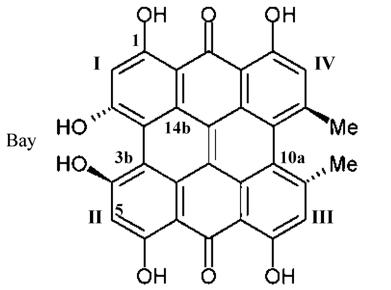


***Figure 1.*** The hypericin molecule (Hy). The molecular axis is defined by the vector uniting carbon atoms 10a to 3b, and the molecular plane is defined by carbon atoms 10a, 3b, and 14b.[16] Bromine substitution was modeled at position I (Hy−Br) and positions I−IV (Hy−4Br).

Hypericin has a high ability to generate singlet oxygen and other reactive oxygen species (ROS) when irradiated by light. The quantum yield for singlet oxygen formation was at first estimated to be as high as 0.73,[17,18] but the number has later been revised to be as low as 0.36 in ethanol and to be less than 0.02 in water.[19] In liposomes, the quantum yields have been measured to be 0.35[20] and 0.43.[21] Given that these quantum yields are too low to explain the strong photodynamic action of the molecule, it can be concluded that there must be some additional mechanism involved.

Brominated hypericin at some of the positions (denoted with I−IV in Figure 1) are of considerable interest as they meet certain fundamental requirements for a photosensitizer. Brominated hypericins have displayed potential phototoxic activity, e.g., against the herpes simplex virus and the influenza virus.[22] It was also shown that some of these enhance the quantum yield of singlet oxygen and increase the quantum efficiency of superoxide formation compared with that of unsubstituted hypericin, as a result of enhanced intersystem crossing between the first excited singlet and triplet states.[23,24] A recent study from our group shows that one possible reaction of brominated hypericin after excitation is reduction followed by dissociation, generating a negatively charged bromine as a leaving group and a hypericin radical capable of binding to biological molecules.[25]

**1.2. Transportation and Carriers of Hypericin.** For drug molecules to reach possible cellular targets, they must transfer across the plasma membrane of the cell. The intracellular location and the way there, often mediated by the assistance of a carrier, depends on the properties of the molecules. This is an important field of study since the ROS generated from the photoreactions have short lifetimes and can only cause oxidative damage in the nearest surroundings. Several experimental studies suggest possible intracellular targets for photosentizisers, but for hypericin, the exact target and the transportation thereto are still not fully understood. One possible way for the molecule to enter the cell is through diffusion.[8] Another possible pathway to cellular transport is by accumulation in low-density lipoproteins (LDL)[15,26,27] and to a lower extent in high-density lipoproteins and human serum albumin[15,28−32] in human plasma when administrated into the bloodstream. Interaction with these biological molecules also helps to solubilize the highly hydrophobic hypericin molecule and to prevent aggregation, which otherwise would suppress virucidal activity and inhibit photodynamic properties. Another way to avoid aggregation and to solubilize the hypericin molecules is to encapsulate them into other appropriate drug-carriers, such as liposomes.[33] Tetra-brominated hypericin has been shown to exhibit higher binding constants to liposomes than hypericin as well as a higher singlet oxygen quantum yield compared to hypericin when bound to liposomes.[34]

**1.3. Intracellular Location of Hypericin.** The penetration of photosensitizers into various cell compartments, especially the nucleus, and their intracellular concentrations are important properties when considering cytotoxic activity. Although the exact intracellular location of hypericin is still unclear, the hydrophobic character of the molecule indicates accumulation in cytoplasmatic membranes, such as the endoplasmic reticulum and the Golgi apparatus, in which the molecule has also been found.[35−37]

Cholesterol carried by LDL is, upon entering the cell, directed to the lysosomes in which it is hydrolyzed. Hypericin encapsulated in LDL has been confirmed by several studies to end up located in the lysosomes.[38,39] Several model systems imply initial lysosomal damage caused by hypericin which triggers the mitochondrial apoptosis pathway.[40] Hypericin has been reported to accumulate in mitochondria,[38,41] and some pathways involve breakdown of the mitochondrial membrane.[42−44]

Studies also show an accumulation in the cell membrane.[8,45−47] One of these studies shows that only after long-term incubation, the molecules can penetrate the membrane and eventually reach the nucleus,[45] which has been pointed out as another possible target for hypericin.[45−47] Hypericin has been shown to interact with DNA, preferably with guanine and adenine nucleotide bases through formation of hydrogen bonds between position N7 of the purines and the hydroxyl groups of hypericin.[48−50]

Membrane lipid peroxidation can be another powerful consequence of photoinduced intracellular damage caused by hypericin.[51−53] Photosensitized hypericin has the ability to decrease the plasma membrane potential (depolarization) as well as the activity of $Na^+$, $K^+$-ATPase in liposome models;[53] which might have serious effects on the condition of a cell. Also in the treatment of various virus infections, membranes seem to be a potential target for hypericin, since only lipid-coated viruses are inactivated by the treatment.[54,55]

The widespread in findings of specific sites of intracellular localization of hypericin is probably due to the usage of different model systems, incubation time, and constitution of the incubation medium.

In order to obtain more insight on the action of hypericin in a biological environment, we here report the behavior of hypericin (Hy) with no, one (Hy−Br; position I, Figure 1) or four bromines (Hy−4Br; positions I−IV, Figure 1) inside a lipid dipalmitoylphosphatidylcholine (DPPC) bilayer, applying molecular dynamics simulations.

## 2. Theoretical Methodology

The molecular dynamics program GROMACS (version 3.3)[56,57] was used throughout the study. The membrane model used was an already equilibrated DPPC bilayer

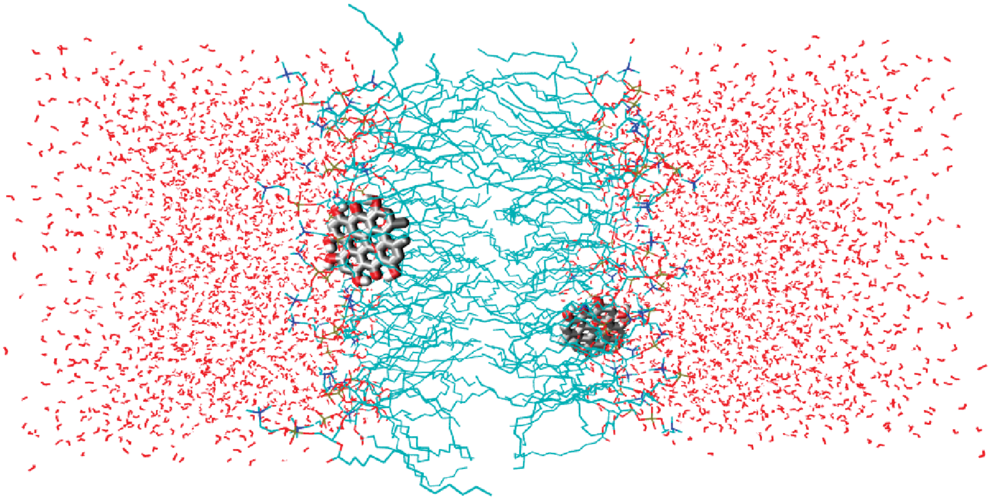


***Figure 2.*** Snapshot from a simulation showing two hypericin molecules in a DPPC bilayer.

consisting of 64 lipids and 3 846 water molecules.[58,59] Three independent simulations were performed, one for each neutral hypericin derivative (Hy, Hy−Br, and Hy−4Br). The geometries of the hypericin molecules were generated from geometry optimizations of neutral hypericin derivatives in the quantum chemistry program Gaussian03[60] at the B3LYP/6-31G(d,p) level of theory. Mulliken atomic charges obtained from the geometry optimizations were assigned to the molecules. The GROMACS force field was used throughout. The topology of hypericin was obtained using the PRODRG software[61] through its web server (http://davapc1.bioch.dundee.ac.uk/prodrg/), using the PDB coordinates obtained in the quantum optimizations. As bromine is not parametrized in the force field, Lennard-Jones and ligand parameters for chlorine were used instead. For the DPPC phospholipids, a standard united-atom force field was applied,[62] and for water, we used the SPC model.[63] The starting point for each simulation was the resulting coordinates from previous 20 ns simulations using different conditions in which two hypericin molecules of each derivative were inserted into the membrane model, one in the outer region of the water phase and one in the middle of the lipid phase of the bilayer (this was part of an initial set of exploratory simulations to look for the general partition behavior of the molecules, and all the unconstrained equilibrium calculations were done using systems that contained the molecules already inside the lipid bilayer). These simulations resulted in both molecules locating in the lipid bilayer part of the system after a short equilibration, due to the hydrophobic character of the molecules. The systems were first equilibrated for 10 ns to uncorrelate those from the previous simulations, and followed by 50 ns productions in which the system trajectories were collected every 0.8 ps. During the simulations, none of the hypericin molecules moved out into the water phase or across the bilayer middle. The two molecules were located during the entire simulations in the lipid bilayer region, on opposite water/lipid interface sides, and were never close enough to interact strongly with each other (see Figure 2).

All simulations were performed using a time step of 2 fs and using the isothermal−isobaric ensemble at $T = 323$ K and $p = 1$ bar. The temperature and the pressure were held constant using a Nosé−Hoover thermostat[64,65] with a coupling constant of 0.1 ps and a semi-isotropic Parrinello−Rahman barostat[66,67] with a coupling constant of 1 ps. A particle mesh Ewald scheme[68,69] was used to calculate the electrostatic interactions with a 10 Å cutoff for the real space. The same cutoff was used for the short-range van der Waals interactions (Lennard-Jones terms). Bond lengths were constrained using the LINCS algorithm.[70]

Analysis was performed on the equilibration runs to check for equilibration convergence and on production runs from which all reported data was obtained. Data analysis programs written in C++ were used when no other program was available in the GROMACS package to calculate various properties reported herein.

A potential of mean force formalism was used to calculate free energy profiles for hypericin molecules across the lipid bilayer (the direction of the $z$-axis). The $z$-component of the force, $F_z$, acting on the molecule at certain constrained distances between the molecule and the bilayer center-of-mass was collected at different positions along the $z$-axis. The free energy for the transfer process between $z_i$ and $z_f$ is written as

$$\Delta G = G_{z_f} - G_{z_i} = -\int_{z_i}^{z_f} \langle F_z \rangle_z \mathrm{d}z \qquad (1)$$

where the bracket means an average over the forces collected at each constrained distance. To calculate the free energy profile for the translocation of each molecule, 41 constrained simulations were performed in which the hypericin molecule was located at different positions that differ by 0.1 Å along the $z$-axis direction. The starting points for the simulations were sampled from the previous unconstrained simulations. To sample the points in the middle of the bilayer, where the molecule was never located during the unconstrained simulations, a weak force was used to push the molecule toward the lipid bilayer, choosing the value of the force to make the least perturbation possible on the bilayer system. Each point in water was equilibrated for 1 ns, and a production run of 4 ns was used. Inside the lipid bilayer, an increase in the sampling was needed due to the slower motion of the molecules and, therefore, each point was equilibrated for at least 4.6 ns, and a production run of 10 ns followed. The force acting on the hypericin center-of-mass was collected at every time step during the production run. A SHAKE algorithm[71] was used to constrain the distance between the center-of-mass of the bilayer and the hypericin molecules (the molecules were constrained in the $z$-direction but allowed to rotate).

The permeability is defined as the current density divided by the concentration gradient across the membrane. The procedure developed by Marrink and Berendsen[72] was adopted to calculate the permeability coefficients, based on the fluctuation dissipation theorem and on using the deviation of the instantaneous force, $F(z,t)$, from the average force acting on the molecule obtained during the constrained dynamics:

$$\Delta F(z, t) = F(z, t) - \langle F(z, t)\rangle \tag{2}$$

The local time-dependent friction coefficient, $\xi$, can be calculated from the following autocorrelation function:

$$\xi(z, t) = \langle \Delta F(z, t)\Delta F(z, 0)\rangle/RT \tag{3}$$

where $T$ is the absolute temperature and $R$ is the gas constant. By integrating the friction coefficient, one can obtain the diffusion coefficient, $D$:

$$D(z) = RT/\xi(z) = (RT)^2/\int_0^\infty \langle \Delta F(z, \mathrm{t})\Delta F(z, 0)\rangle \mathrm{d}t \tag{4}$$

This function was fitted to a double exponential using a nonlinear fitting procedure[72] in order to integrate the autocorrelation of the force fluctuations:

$$C(t) = A_0 \exp(-t/\tau_0) + A_1 \exp(-t/\tau_1) \tag{5}$$

This illustrates that the molecules move inside the lipid bilayer in two distinct time scales, corresponding to the two decay times, $\tau_0$ and $\tau_1$, one fast and one slow.

The permeability coefficient, $P$, can be calculated by integrating over the local resistances across the membrane, $R(z)$. $R(z)$ is obtained by dividing the exponential of the previously calculated free energies, $\Delta G(z)$, by the diffusion coefficients, $D(z)$:

$$1/P = \int R(z)\mathrm{d}z = \int_{z_\mathrm{i}}^{z_\mathrm{f}} \frac{\exp(\Delta G(z)/kT)}{D(z)}\mathrm{d}z \tag{6}$$

## 3. Results and Discussion

Already during initial equilibration (unpublished data), the hypericin molecules moved into the lipids and remained there for the rest of the simulation. In Figure 2, we display a snapshot from the simulation. The probability of finding the molecules in the interface between the lipids and the water phase is high, as illustrated in Figure 3A for Hy−Br. In Figure 3B, the equilibrium distributions of all three hypericin derivatives are displayed. The figures show clearly that the probability to find the hypericin molecules is high close to the densest region of the membrane, i.e., close to the polar lipid head groups. The molecules are highly hydrophobic but also have some amphiphilic character due to the many hydroxyl groups. These groups have the ability to interact with water, which explains why such large and inflexible molecules accumulate in the most dense region of the membrane, where they are within the lipid phase yet in contact with waters that penetrate into the bilayer. The final density profile of Hy−4Br is wider compared to Hy and Hy−Br, suggesting that this molecule is moving closer to the bilayer middle (the same applies to Hy−Br when compared to Hy). This affects the possibility to interact with surrounding water, as is discussed below in connection with radial distribution functions. Hypericin displays the most narrow density profile. None of the three molecules moves further out than 2.5 nm from the bilayer center. The density maximum is located closer to the bilayer center with an increasing level of bromination. A related study performed on the action of psoralen derivatives in lipid membranes showed similar accumulation in the lipid region, although slightly closer to the bilayer center (away from the densest region of the membrane), despite the smaller sizes of those compounds.[73] Generally similar partition profiles were also obtained for noncharged, hexyl ester and ethyl ester 5-aminolevulinic acids.[74]

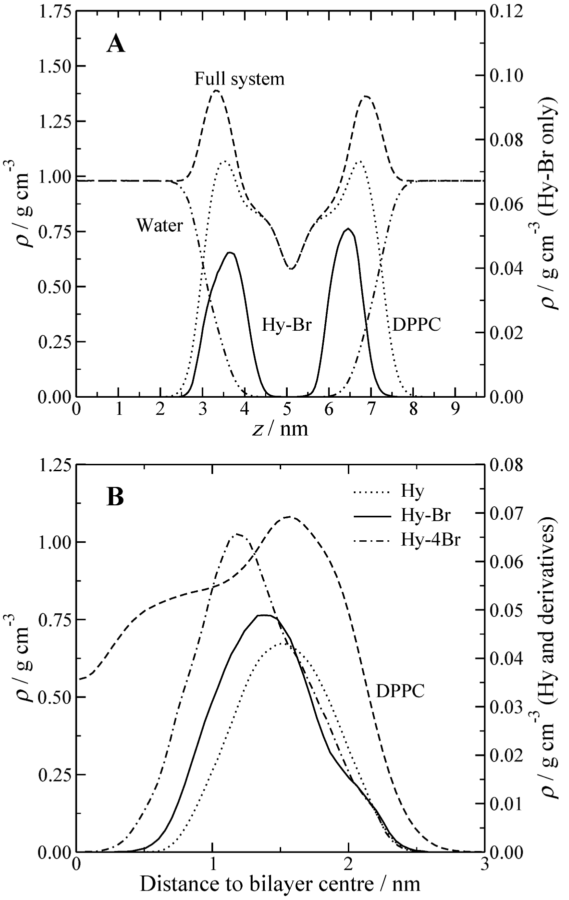


***Figure 3.*** (A) Density profile for Hy−Br in the DPPC bilayer. (B) Resulting distribution for the three different hypericin derivatives as function of distance to the bilayer middle and compared with the DPPC density profile. The latter were obtained using the bilayer center-of-mass as the folding point.

A molecular axis tilt angle was defined as the vector uniting carbon 10a to carbon 3b (Figure 1), since those atoms belong to the inner and rigid part of the hypericin molecule and can give a clear idea about the orientation of the bay area of the molecule. An important aspect to account for is

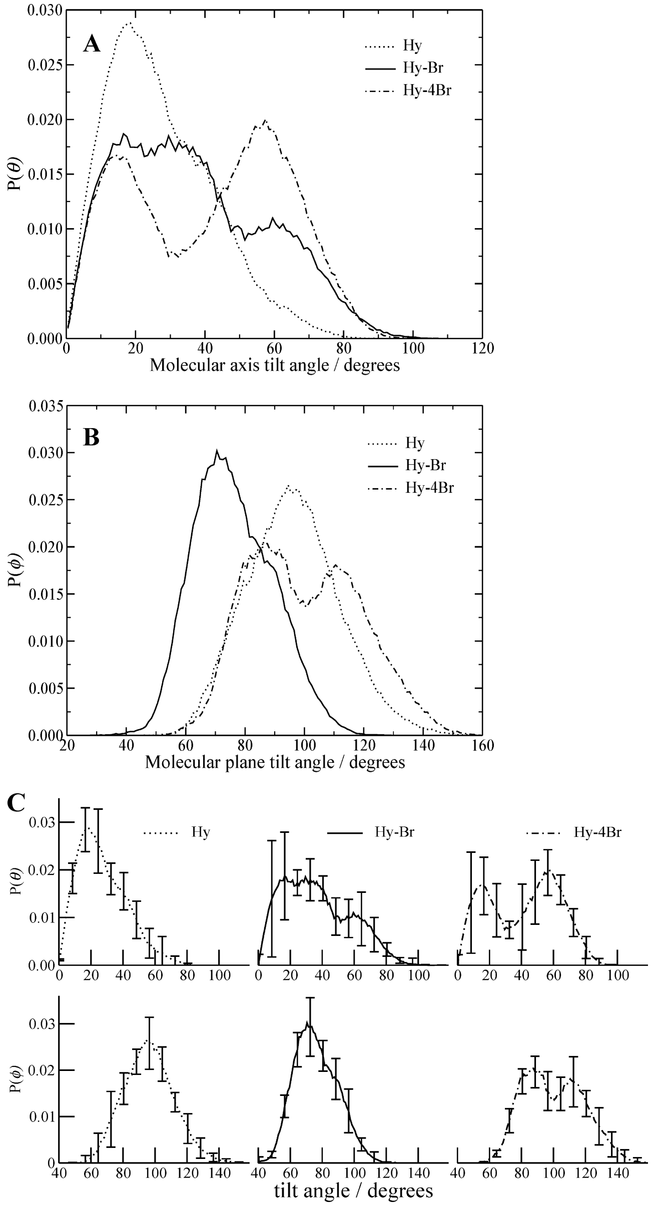


***Figure 4.*** (A) The angle between the vector normal to the bilayer and the molecular axis vector pointing from atom $C_{10a}$ to atom $C_{3b}$. (B) Molecular plane tilt angle, defined by the angle formed by the bilayer normal and the vector normal to the molecular plane. (C) Same plots as A and B for the molecular axis tilt (upper graphs) and molecular plane tilt (lower graphs) with the error bars this time displayed separately for clarity.

the fact that the bilayer is symmetric, with two interfaces between lipids and water, generating two normal vectors pointing in opposite directions. This was considered by subtracting normal vectors pointing toward negative values by 180°. The molecular axis tilt angle distributions for the hypericin derivatives are plotted in Figure 4A. The results are interpreted such that if the angle is 0° or 180°, then the molecular axis is parallel to the bilayer normal (the *z*-direction), and if the angle is 90°, then the molecular axis is perpendicular to the bilayer normal.

For hypericin, the distribution is much sharper than for the other two molecules and has a clear maximum located around 18°. For Hy−Br and Hy−4Br, the distributions are wider and bimodal, presenting two maxima, one located close to 16° (16−40° for Hy−Br) and another close to 58°. The major difference between the molecules is that an increased bromine content results in larger angle values, which are more probable. The first thing to note from these values is that the bay area of the molecule containing alcohol groups is oriented toward the water interface and the opposite part of the molecule containing the two methyl groups is located toward the inner and apolar part of the bilayer, in accordance with the hydrophobic nature of these.

A simple angle calculation shows that the molecules tend to maximize the interaction of the OH groups with the water interface. If the molecular axis tilt angle was 0°, then only two groups would be close to water, but by rotating ∼20°, the other two OH groups also surrounding the carbonyl oxygen are oriented closer to water. An angle of about 50° indicates that the bromine atoms become closer to water, and thus, in the molecules containing bromine atoms, there is a competitive balance between orienting both the OH and Br groups toward water. It is important to note that the probability to find an angle larger than or equal to 90° is essentially zero for all three molecules. This means that the methyl groups are never closer to the water interface than the OH groups.

The molecular plane was defined by carbons 10a, 3b, and 14b, and the plane tilt angle $\phi$ was defined by the angle formed by the bilayer normal and the vector normal to the molecular plane. The molecular plane tilt angle distributions for the hypericin derivatives are plotted in Figure 4B. The results are interpreted as follows: if the angle is 0° or 180°, then the molecular plane is aligned with the bilayer plane, and if the angle is 90°, then the molecular plane is perpendicular to the bilayer plane. Since hypericin and its derivatives have a disk-like shape, one would expect the molecule plane to be perpendicular to the bilayer plane, since this way less interfacial area is needed to fit each molecule inside the bilayer. This implies that plane tilt angles close to 0° or 180° would be difficult or impossible to find, but instead the plane tilt angle should be close to and somewhere around 90°. This is clearly seen in Figure 4B, where the tilt angle for all three molecules can vary between 40° to 150° with maxima close to 90°, especially for Hy and Hy−4Br.

The plane tilt angle maximum for Hy−Br is around 70°, that is, 20° off the molecule plane being perpendicular with the bilayer plane. For Hy we have a maximum around 95°, 5° away from an alignment with the perpendicular to the bilayer plane. For Hy−4Br we have a bimodal distribution, which means that the angles sampled preferentially around two maxima: one around ∼85° and another around ∼110°. Both maxima are centered around distributions of the same amplitude. One should take in consideration that, due to symmetry reasons, the relative position of the molecule is the same, if the plane tilt angle has the same divergence from 90°. This means that the first maximum of Hy−4Br ($\phi$ = 85°) has the same relative position in relation to the interface as Hy ($\phi$ = 95°). The other maximum of Hy−4Br, located close to 110°, has a deviation of 20° from being perpendicular to the bilayer plane, the same amount as for Hy−Br ($\phi$ = 70°). From these data, one can conclude that hypericin preferentially orients itself such that the molecular plane is aligned along the *z*-axis of the bilayer. The addition of one bromine favors the molecular plane tilting, and with the addition of four bromine, the molecular symmetry is regained and the molecule orients in either position.

The error bars, considered as the standard deviation of the mean value:[75]

$$\sigma_{\mathrm{av}} = \frac{\sigma}{\sqrt{(N/D - 1)}} \tag{7}$$

were calculated and are displayed in Figure 4C. Each error bar in the tilt distributions was calculated by dividing the production run in blocks of 1 ns ($N = 50$) and by calculating the standard deviation of the mean value of the blocks ($\sigma$). Autocorrelation functions ($C_n$) using these 50 values were used to calculate the decorrelation time $D$ (time corresponding to $C_n$ being equal to zero). With all this data, the error bars can be calculated according to the previous equation.

A simple test to see if the tilt axes distributions sample (or not) bimodal distributions is to split the 50 ns production runs in blocks and to analyze each block separately. If the tilt axis has a bimodal distribution, one can see in each block a clear preference in each molecule for one or the other bimodal zones and in between blocks the changing of probabilities of each bimodal zone because some molecule (or molecules) change the tilt and start to sample the other zone. The error bars also give an indication of this fact. For instance, the tilt molecular axis of the Hy−4Br molecule (Figure 4C, top right) can be found to preferentially sample values around 16° or 58°, and the error near 30° is lower because the two distributions have similar probabilities in that area. Moreover, the larger error bar for brominated hypericin is due to the fact that these molecules have the tendency to sample two different regions of the angle space, in opposition to hypericin that has only a single and sharp distribution.

Summarizing the molecular axis and molecular plane tilt angle distributions, we can conclude that hypericin and its derivatives are oriented in such a way to have the hydroxyl groups close to the water interface, with possible rotation to allow the bromine atoms to also become oriented toward this interface, and that the molecular plane is essentially perpendicular to the bilayer plane.

Radial distribution functions between oxygen atoms on the hypericin derivatives and hydrogen atoms in the surrounding water (Figure 5A) and between polar hydrogen atoms on the hypericin derivatives and oxygen atoms in the surrounding water (Figure 5B) were calculated. The first peak in both figures (at ~0.18 nm) corresponds to a hydrogen bond. The following peak in Figure 5A corresponds to the second hydrogen in the same water molecule or a second solvation shell, whereas the second peak in Figure 5B corresponds to a second solvation shell of water. After this, there is an increase in amplitude of the radial distribution functions as more and more water molecules are included in the shells of higher order. The fact that hydrogen bonds are found implies that the molecules do interact with water. As the bay hydroxyl groups are located very close to the interface between the lipids and the water, these are the likely atoms/groups involved in hydrogen-bond formation. The radial distribution functions decrease in the order Hy > Hy−Br > Hy−4Br, both in the case of hydrogen and oxygen interactions. The probability to find hydrogen bonds between oxygen on the hypericin molecules and hydrogen atoms in water is higher than between hypericin hydrogens and surrounding oxygen. This is due to the two lone pairs on oxygen that, hence, enables formation of two hydrogen bonds. Hydrogens bonds are detected for all molecules, involving hydrogen as well as oxygen on the hypericin molecules, except in the case involving hydrogen on Hy−4Br for which a peak is hardly visible. As discussed in connection with the density profiles above, Hy−4Br moves closer to the bilayer center compared to Hy and Hy−Br, which reduces the interaction with water and, thereby, results in lower radial distribution functions. Hypericin displays a more narrow density profile, positioned in the region closest to the interface between the lipids and water, and thus, the radial distribution functions are the highest for hypericin.

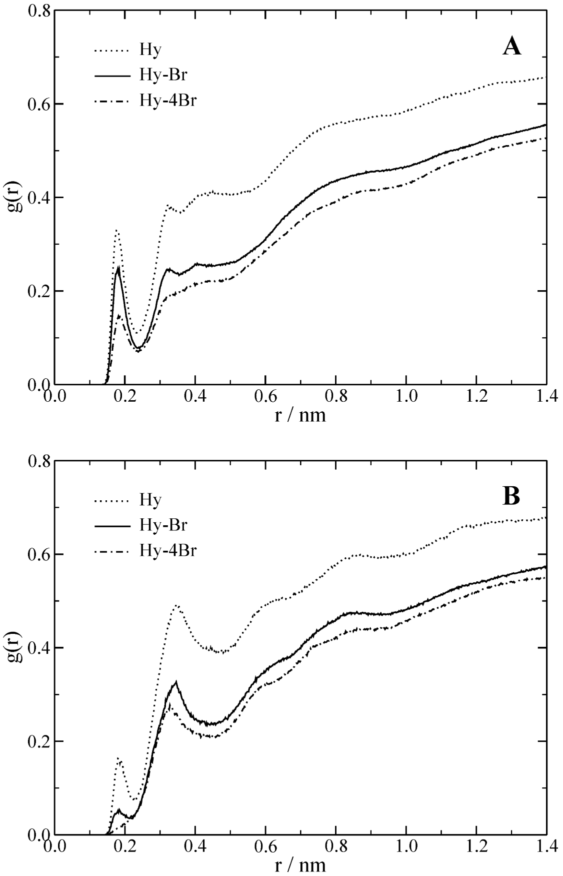


***Figure 5.*** Radial distribution functions between (A) oxygen atoms on the hypericin derivatives and hydrogen atoms in the surrounding water; and (B) between polar hydrogen atoms on the hypericin derivatives and oxygen atoms in the surrounding water.

The mean-square displacement (MSD)[76] reveals details about the movements of the molecules inside the bilayer. The MSD is defined by

$$\mathrm{MSD}(t) = \langle |\vec{r}(t) - \vec{r}(0)|^2 \rangle \tag{8}$$

where $\vec{r}(0)$ and $\vec{r}(t)$ are the positions of a particle at time $t =$ 0 and at a certain time $t$. The brackets indicate a time average

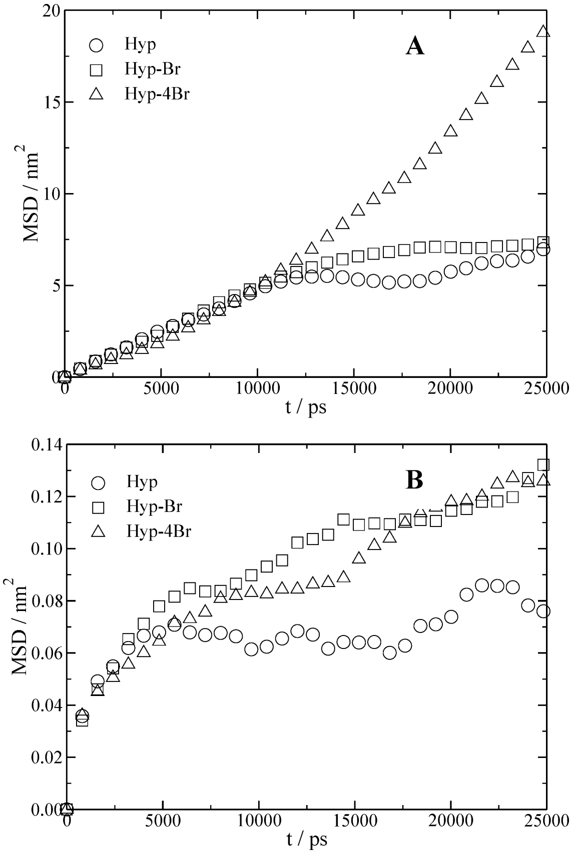


*Figure 6.* MSD in the (A) *xy*-plane and in the (B) direction normal to the bilayer.

over all similar particles and over different time origins along the simulation. The Einstein relation allows for the calculation of the diffusion coefficient, $D$, at sufficiently long simulation times:[76]

$$D = \lim_{t\to\infty} \frac{1}{2dt}\langle |r_i(t) - r_i(0)|^2 \rangle \qquad (9)$$

where $d$ is the dimensionality of the space. This way, one can obtain the MSD for the molecules moving in the bilayer plane ($d = 2$) and moving along the bilayer normal ($d = 1$), respectively. The MSD provides a measure of the average distance a molecule travels in the system, and the growth rate of the MSD depends on how often the molecule collides, i.e., a measure of the ease of diffusion of the molecule.

Like other molecules diffusing in confined media, the hypericin molecules never reach the Einsteinian limit of proper diffusion within the limited time of the simulation, and anomalous diffusion occurs where MSD is proportional to $t^n$, with $0 < n < 1$.[77] The implication is that a direct comparison with experimental diffusion coefficients cannot be made. However, based on the MSD, one can state which molecules have a higher or a lower diffusive regime. The MSD in the bilayer plane and along the normal of the bilayer ($z$-direction) are displayed in Figure 6A and B, respectively. From Figure 6A, it is clear that the movement in the bilayer plane is not significantly changed by adding a bromine atom to hypericin, whereas the addition of four bromine atoms allows the molecule to move more easily. Although Hy−4Br is heavier, the MSD primarily reflects the hydrogen-bond capability, in accordance with the findings for the radial distribution functions that showed a clear decrease in hydrogen-bonding interaction between Hy−4Br and water.

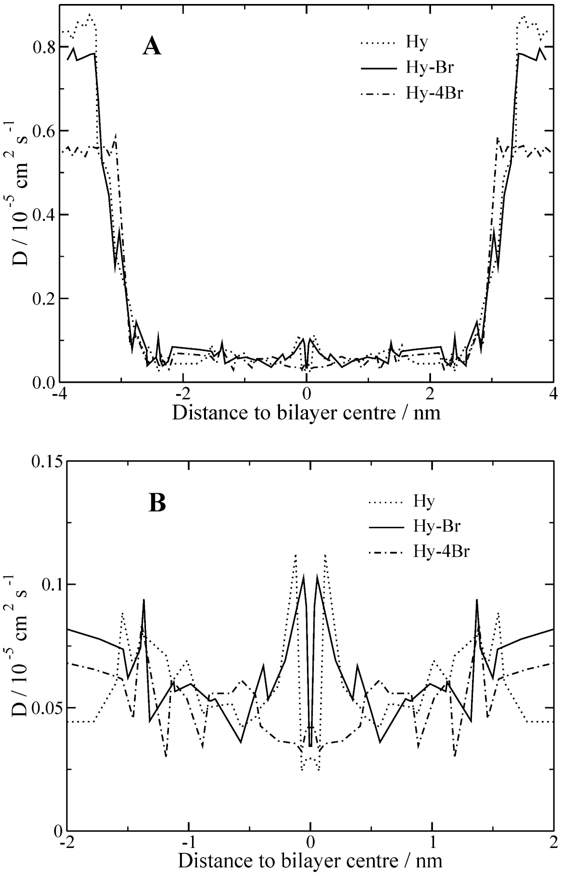


*Figure 7.* (A). Local diffusion coefficients of the hypericin derivatives in the DPPC bilayer, as functions of the distance to the bilayer middle. (B) Magnification of the region within 2 nm from the bilayer center.

One should be cautious in interpreting the MSD along the bilayer normal since this movement, as opposed to the movement in the bilayer plane, is finite, and hence, the MSD should level off independently of any characteristic of the molecule under consideration. Anyway this property is displayed in Figure 6B and shows that the addition of at least one bromine atom increases the diffusion of the molecules. This again shows that the higher ability of hypericin to form hydrogen bonds with water makes its average position closer to the lipid−water interface (i.e., the time spent at the interface is larger) and hinders both the movement of the molecule in the direction normal to the bilayer and in the bilayer plane. Comparing both movements in the bilayer, it is also clear that the diffusion in the bilayer plane is much higher than that in the $z$-direction, for all molecules.

The local diffusion coefficients across the bilayer were calculated by integrating the fitted autocorrelation functions (eq 4), and the dependence on the distance to the bilayer center is displayed in Figure 7A and B (the latter showing only the region within 2 nm from the bilayer center). Figure 7A shows a significant difference between the diffusion coefficients in the lipid region of the bilayer and in the water phase. The diffusion of the molecules in water is faster than

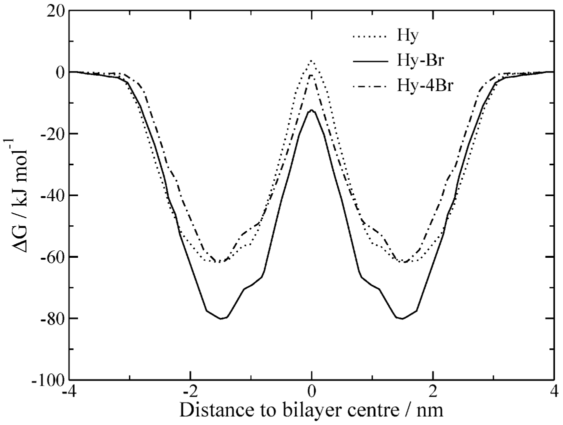


***Figure 8.*** Free energy profiles for the hypericin derivatives inside the DPPC bilayer.

inside the lipids, since the water molecules can rearrange faster around the hypericin molecules as they move around. When hypericin is inside the bilayer, space must be created by the much slower moving lipid molecules before it can move to another position. Inside water, the diffusion coefficients of the molecules decrease with increased bromine content. This seems to be caused by the different molecular size and especially by the different molecular weights of the molecules, where the heavier and the larger molecule moves more slowly.

Looking closer at the diffusion coefficients within the lipid region of the bilayer (Figure 7B), it can be seen that Hy−4Br moves slower than the other two molecules, also in the very middle of the bilayer where the density is at its lowest. The other two molecules show a minor increase in diffusion very close to the bilayer center, with very similar profiles.

Free energy profiles for the transport process from water and into the lipids, as functions of the distance to the bilayer center, were calculated using the potential of mean force formalism outlined above (Figure 8).[78] In similar calculations, using the same technique but only 2 ns in the production runs (we used 4 ns in water and 10 ns in the lipid bilayer),[79] they obtained errors in the free energy that ranged from about 0.7 to 4 kJ/mol in the bilayer middle, where the errors were found to be larger and, in a way, that was not clearly dependent on the size of the molecules studied. The profiles show a local minimum in the region 1−2 nm from the bilayer center, near the polar headgroup region, when moving from the water phase into the lipids. Having passed this minimum, the free energy increases when the molecules move toward the bilayer middle. Hy is the only one of the three molecules which shows a positive change in free energy at the very middle of the bilayer, with an increase of ∼3.5 kJ/mol compared to furthest out in the water phase. Hy−4Br shows a decrease of ∼1.2 kJ/mol, and Hy-Br shows a decrease of ∼12.3 kJ/mol in the bilayer middle. Hy−Br also displays a deeper free energy minimum, ∼18 kJ/mol lower than for Hy and Hy−4Br, in the region close to the polar headgroups. This indicates that Hy−Br is the most likely to accumulate inside the lipids, and, since the diffusion in the lipid region is low, the molecules are more likely to reside in the polar headgroup region of the bilayer.

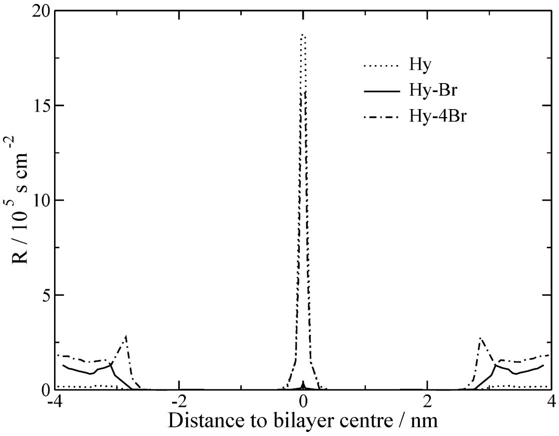


***Figure 9.*** Local resistance profiles of the different hypericin derivatives in the DPPC bilayer, as functions of the distance to the bilayer middle. The hypericin profile was scaled down by a factor 7.

***Table 1.*** Permeability Coefficients Inside the DPPC Bilayer (cm s$^{-1}$)

| molecule | permeability coefficients |
|---|---|
| Hy | $4.21 \times 10^{-4}$ |
| Hy−Br | $4.94 \times 10^{-3}$ |
| Hy−4Br | $1.51 \times 10^{-3}$ |

From the calculated free energy profiles and the local diffusion coefficients across the lipid bilayer, the local resistance was calculated using eq 6, and the resulting profiles for the three molecules are displayed in Figure 9. To display all three molecules more clearly, the Hy resistance profile was scaled down by a factor 7. For all three molecules, an increase in resistance was found in the bilayer middle, although for Hy−Br this is much less than for Hy and Hy−4Br. For Hy and Hy−4Br, the peaks in the center are considerably higher than the increased resistance seen in the water phase, whereas for Hy−Br the opposite is noted. The free energy plays a dominant role in the appearance of the resistance profiles, and the increase in free energy in the middle of the bilayer is in accordance with the increased resistance in this region.

Permeability coefficients were calculated by integrating the resistance profiles across the bilayer, and from those, it can be concluded that the permeation decreases in the order Hy−Br > Hy−4Br > Hy (Table 1). As mentioned before, the permeability strongly depends on the free energy, and the decrease in permeation follows the increase in energy. Hy−Br displays a significantly lower free energy, both close to the polar headgroups of the lipids and in the center of the lipids, than the other two which contributes to an easier and, therefore, faster permeation. An experimental study has shown that halogenation of drug molecules enhances permeation by increasing the permeability coefficients and enhances the free energy of transfer into the lipid membrane, as compared to the nonsubstituted molecules.[80] Those results are in agreement with our findings that the brominated hypericins display higher permeability coefficients than that of the nonsubstituted hypericin.

Estimated permeability coefficients from experiments of hypericin across monolayers of Caco-2 cells are in the order of $10^2-10^3$ less than those calculated herein.[46] The difference when comparing experimental and calculated permeability coefficients is probably due to that natural membranes contain many more components, such as proteins and cholesterol, which can either enhance or suppress the permeability through the membrane. The present computational study was performed with a simplified membrane model and with the aim to study passive permeation only, and how the permeability is affected by different levels of bromination.

## Conclusions

Three hypericin derivatives were studied in order to reveal their distribution and specific properties in a DPPC lipid membrane, using classical molecular dynamics simulations. All three molecules showed a strong preference to accumulate in the densest region of the membrane, close to the polar headgroups. Hypericin accumulates closest to the interface between the lipids and water. This is also manifested by radial distribution functions which show the highest number of hydrogen bonds between hypericin and water. Local diffusion coefficients show, as expected, high-diffusion rates in the water phase compared to that of the lipids, due to the large size and hydrophobic character of the molecules. Calculated permeability coefficients suggest a faster overall diffusion for Hy−Br. This finding is also supported by the free energy profiles which displays a more negative change in free energy for the transport process of Hy−Br moving from water into the lipids. For all three hypericin derivatives, the free energy profiles display minima within 1−2 nm from the bilayer center, in the same region as where the molecules were found to accumulate according to the density profiles. According to the present results, we can expect more of Hy and Hy−4Br to accumulate within the membrane, suggesting a larger possibility of direct photodamage caused by those. Hy−Br has a higher capability to translocate across the membrane and would potentially have a larger probability to penetrate the membrane and, thus, reach other targets in the interior of a cell.

**Acknowledgment.** Financial support from the Modelling and Simulation Research Center (ESEE), the Swedish Research Council, the Faculty of Science and Technology at Örebro University, and the National University of Ireland, Galway (LAE), is gratefully acknowledged. We also acknowledge generous grants of computing time at the National Supercomputing Center (NSC) in Linköping. The Portuguese Fundação para a Ciência e a Tecnologia is also acknowledged for financial support.

## References

(1) Cerny, C. *Z. Phys. Chem.* **1911**, *73*, 371–382.

(2) Moraleda, G.; Wu, T. T.; Jilbert, A. R.; Aldrich, C. E.; Condreay, L. D.; Larsen, S. H.; Tang, J. C.; Colacino, J. M.; Mason, W. S. *Antivir. Res.* **1993**, *20*, 235–247.

(3) Lopezbazzocchi, I.; Hudson, J. B.; Towers, G. H. N. *Photochem. Photobiol. Sci.* **1991**, *54*, 95–98.

(4) Hudson, J. B.; Lopezbazzocchi, I.; Towers, G. H. N. *Antivir. Res.* **1991**, *15*, 101–112.

(5) Meruelo, D.; Lavie, G.; Lavie, D. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 5230–5234.

(6) Lenard, J.; Rabson, A.; Vanderoef, R. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 158–162.

(7) Degar, S.; Prince, A. M.; Pascual, D.; Lavie, G.; Levin, B.; Mazur, Y.; Lavie, D.; Ehrlich, L. S.; Carter, C.; Meruelo, D. *AIDS Res. Hum. Retroviruses* **1992**, *8*, 1929–1936.

(8) Thomas, C.; Pardini, R. S. *Photochem. Photobiol. Sci.* **1992**, *55*, 831–837.

(9) Andreoni, A.; Colasanti, A.; Colasanti, P.; Mastrocinque, M.; Riccio, P.; Roberti, G. *Photochem. Photobiol. Sci.* **1994**, *59*, 529–533.

(10) VanderWerf, Q. M.; Saxton, R. E.; Chang, A.; Horton, D.; Paiva, M. B.; Anderson, J.; Foote, C.; Soudant, J.; Mathey, A.; Castro, D. J. *Laryngoscope* **1996**, *106*, 479–483.

(11) Thomas, C.; Pardini, L.; Pardini, R. S. In 3rd Biennial Meeting of the International Photodynamic Association; Buffalo, NY, 1990.

(12) Chung, P. S.; Rhee, C. K.; Kim, K. H.; Paek, W.; Chung, J.; Paiva, M. B.; Eshraghi, A. A.; Castro, D. J.; Saxton, R. E. *Laryngoscope* **2000**, *110*, 1312–1316.

(13) Liu, C. D.; Kwan, D.; Saxton, R. E.; McFadden, D. W. *J. Surg. Res.* **2000**, *93*, 137–143.

(14) Chen, B.; de Witte, P. A. *Cancer Lett.* **2000**, *150*, 111–117.

(15) Chen, B.; Xu, Y.; Roskams, T.; Delaey, E.; Agostinis, P.; Vandenheede, J. R.; de Witte, P. *Int. J. Cancer* **2001**, *93*, 275–282.

(16) Freeman, D.; Konstantinovskii, L.; Mazur, Y. *Photochem. Photobiol. Sci.* **2001**, *74*, 206–210.

(17) Jardon, P.; Lazortchak, N.; Gautron, R. *J. Chim. Phys. Phys.-Chim. Biol.* **1987**, *84*, 1141–1145.

(18) Racinet, H.; Jardon, P.; Gautron, R. *J. Chim. Phys. Phys.-Chim. Biol.* **1988**, *85*, 971–977.

(19) Darmanyan, A. P.; Burel, L.; Eloy, D.; Jardon, P. *J. Chim. Phys. Phys.-Chim. Biol.* **1994**, *91*, 1774–1785.

(20) Bouirig, H.; Eloy, D.; Jardon, P. *J. Chim. Phys. Phys.-Chim. Biol.* **1992**, *89*, 1391–1411.

(21) Ehrenberg, B.; Anderson, J. L.; Foote, C. S. *Photochem. Photobiol. Sci.* **1998**, *68*, 135–140.

(22) Hudson, J. B.; Delaey, E.; De Witte, P. A. *Photochem. Photobiol. Sci.* **1999**, *70*, 820–822.

(23) Delaey, E.; Zupko, I.; Chen, B.; Derycke, A.; Van Laar, F.; De Vos, D.; De Witte, P. *Int. J. Oncol.* **2003**, *23*, 519–524.

(24) Guedes, R. C.; Eriksson, L. A. *J. Photochem. Photobiol., A* **2006**, *178*, 41–49.

(25) Eriksson, E. S. E.; Guedes, R. C.; Eriksson, L. A. *Int. J. Quantum Chem.* **2008**, *108*, 1921–1929.

(26) Kascakova, S.; Refregiers, M.; Jancura, D.; Sureau, F.; Maurizot, J. C.; Miskovsky, P. *Photochem. Photobiol. Sci.* **2005**, *81*, 1395–1403.

(27) Mukheriee, P.; Adhikary, R.; Halder, M.; Petrich, J. W.; Miskovsky, P. *Photochem. Photobiol. Sci.* **2008**, *84*, 706–712.

(28) Sjoholm, I.; Ekman, B.; Kober, A.; Ljungstedtpahlman, I.; Seiving, B.; Sjodin, T. *Mol. Pharmacol.* **1979**, *16*, 767–777.

(29) Falk, H.; Meyer, J. *Monatsh. Chem.* **1994**, *125*, 753–762.

(30) Kohler, M.; Gafert, J.; Friedrich, J.; Falk, H.; Meyer, J. *J. Phys. Chem.* **1996**, *100*, 8567–8572.

(31) Miskovsky, P.; Jancura, D.; Sanchez-Cortes, S.; Kocisova, E.; Chinsky, L. *J. Am. Chem. Soc.* **1998**, *120*, 6374–6379.

(32) Miskovsky, P.; Hritz, J.; Sanchez-Cortes, S.; Fabriciova, F.; Ulicny, J.; Chinsky, L. *Photochem. Photobiol. Sci.* **2001**, *74*, 172–183.

(33) Vemuri, S.; Rhodes, C. T. *Pharm. Acta Helv.* **1995**, *70*, 95–111.

(34) Roslaniec, M.; Weitman, H.; Freeman, D.; Mazur, Y.; Ehrenberg, B. *J. Photochem. Photobiol., B* **2000**, *57*, 149–158.

(35) Vandenbogaerde, A. L.; Delaey, E. M.; Vantieghem, A. M.; Himpens, B. E.; Merlevede, W. J.; de Witte, P. A. *Photochem. Photobiol. Sci.* **1998**, *67*, 119–125.

(36) Uzdensky, A. B.; Ma, L. W.; Iani, V.; Hjortland, G. O.; Steen, H. B.; Moan, J. *Laser Med. Sci.* **2001**, *16*, 276–283.

(37) Delaey, E. M.; Obermueller, R.; Zupko, I.; De Vos, D.; Falk, H.; de Witte, P. A. M. *Photochem. Photobiol. Sci.* **2001**, *74*, 164–171.

(38) Ali, S. M.; Chee, S. K.; Yuen, G. Y.; Olivo, M. *Int. J. Mol. Med.* **2002**, *9*, 257–270.

(39) Siboni, G.; Weitman, H.; Freeman, D.; Mazur, Y.; Malik, Z.; Ehrenberg, B. *Photochem. Photobiol. Sci.* **2002**, *1*, 483–491.

(40) Guicciardi, M. E.; Leist, M.; Gores, G. J. *Oncogene* **2004**, *23*, 2881–2890.

(41) Theodossiou, T. A.; Noronha-Dutra, A.; Hothersall, J. S. *Int. J. Biochem. Cell Biol.* **2006**, *38*, 1946–1956.

(42) Vantieghem, A.; Assefa, Z.; Vandenabeele, P.; Declercq, W.; Courtois, S.; Vandenheede, J. R.; Merlevede, W.; de Witte, P.; Agostinis, P. *FEBS Lett.* **1998**, *440*, 19–24.

(43) Vantieghem, A.; Xu, Y.; Vandenabeele, P.; Denecker, G.; Vandenheede, J. R.; Merlevede, W.; de Witte, P. A.; Agostinis, P. *Photochem. Photobiol. Sci.* **2001**, *74*, 133–142.

(44) Berlanda, J.; Kiesslich, T.; Oberdanner, C. B.; Obermair, F. J.; Krammer, B.; Plaetzer, K. *J. Environ. Pathol. Toxicol. Oncol.* **2006**, *25*, 173–188.

(45) Miskovsky, P.; Sureau, F.; Chinsky, L.; Turpin, P. Y. *Photochem. Photobiol. Sci.* **1995**, *62*, 546–549.

(46) Sattler, S.; Schaefer, U.; Schneider, W.; Hoelzl, J.; Lehr, C. M. *J. Pharm. Sci.* **1997**, *86*, 1120–1126.

(47) English, D. S.; Doyle, R. T.; Petrich, J. W.; Haydon, P. G. *Photochem. Photobiol. Sci.* **1999**, *69*, 301–305.

(48) Miskovsky, P.; Chinsky, L.; Wheeler, G. V.; Turpin, P. Y. *J. Biomol. Struct. Dyn.* **1995**, *13*, 547–552.

(49) SanchezCortes, S.; Miskovsky, P.; Jancura, D.; Bertoluzza, A. *J. Phys. Chem.* **1996**, *100*, 1938–1944.

(50) Kocisova, E.; Chinsky, L.; Miskovsky, P. *J. Biomol. Struct. Dyn.* **1998**, *15*, 1147–1154.

(51) Senthil, V.; Jones, L. R.; Senthil, K.; Grossweiner, L. I. *Photochem. Photobiol. Sci.* **1994**, *59*, 40–47.

(52) Hadjur, C.; Richard, M. J.; Parat, M. O.; Jardon, P.; Favier, A. *Photochem. Photobiol. Sci.* **1996**, *64*, 375–381.

(53) Chaloupka, R.; Obsil, T.; Plasek, J.; Sureau, F. *Biochim. Biophys. Acta Biomembr.* **1999**, *1418*, 39–47.

(54) Tang, J.; Colacino, J. M.; Larsen, S. H.; Spitzer, W. *Antivir. Res.* **1990**, *13*, 313–325.

(55) Hudson, J. B.; Towers, G. H. N. *Pharmacol. Therapeut.* **1991**, *49*, 181–222.

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

(57) Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(58) Marrink, S. J.; Berger, O.; Tieleman, P.; Jahnig, F. *Biophys. J.* **1998**, *74*, 931–943.

(59) Biocomputing at the University of Calgary, Structures and Topologies. http://moose.bio.ucalgary.ca/index.php?page=Structures_and_Topologies. Accessed December 1, 2007), File: dppc64.pdb.

(60) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Jr., J. A. M.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov,B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian03; Rev E.01*; Gaussian, Inc.: Pittsburgh PA, 2003.

(61) Schuttelkopf, A. W.; van Aalten, D. M. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 1355–63.

(62) Biocomputing at the University of Calgary, Structures and Topologies. http://moose.bio.ucalgary.ca/index.php?page=Structures_and_Topologies. Accessed December 1, 2007), Files: lipid.itp and dppc.itp.

(63) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel Publishing Company: Dordrecht, Netherlands, 1981.

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

(65) Hoover, W. G. *Phys. Rev.* **1985**, *31*, 1695–1697.

(66) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(67) Nose, S.; Klein, M. L. *Phys. Rev. Lett.* **1983**, *50*, 1207–1210.

(68) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(69) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(70) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(71) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.

(72) Marrink, S. J.; Berendsen, H. J. C. *J. Phys. Chem.* **1994**, *98*, 4155–4168.

(73) dos Santos, D. J. V. A.; Eriksson, L. A. *Biophys. J.* **2006**, *91*, 2464–2474.

(74) Erdtman, E.; dos Santos, D. J. V. A.; Lofgren, L.; Eriksson, L. A. *Chem. Phys. Lett.* **2008**, *463*, 178–182.

(75) dos Santos, D. J. V. A.; Muller-Plathe, F.; Weiss, V. C. *J. Phys. Chem. C* **2008**, *112*, 19431–19442.

(76) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, U.K., 1990.

(77) Müller-Plathe, F.; Rogers, S. C.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1992**, *199*, 237–243.

(78) Paci, E.; Ciccotti, G.; Ferrario, M.; Kapral, R. *Chem. Phys. Lett.* **1991**, *176*, 581–587.

(79) Bemporad, D.; Essex, J. W.; Luttmann, C. *J. Phys. Chem. B* **2004**, *108*, 4875–4884.

(80) Gerebtzoff, G.; Li-Blatter, X.; Fischer, H.; Frentzel, A.; Seelig, A. *ChemBioChem* **2004**, *5*, 676–684.

CT9002702

JCTC Journal of Chemical Theory and Computation

# Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties

Riccardo Baron,*[†,‡] Philippe H. Hünenberger,[§] and J. Andrew McCammon[†,‡,||,⊥]

*Department of Chemistry & Biochemistry, Center for Theoretical Biological Physics, Department of Pharmacology, and Howard Hughes Medical Institute, University of California San Diego, La Jolla, California 92093-0365 and Laboratory of Physical Chemistry, ETHZ, ETH Hönggerberg, CH 8093, Zürich, Switzerland*

Received July 18, 2009

**Abstract:** The convergence properties of the absolute single-molecule configurational entropy and the correction terms used to estimate it are investigated using microsecond molecular dynamics simulation of a peptide test system and an improved methodology. The results are compared with previous applications for systems of diverse chemical nature. It is shown that (i) the effect of anharmonicity is small, (ii) the effect of pairwise correlation is typically large, and (iii) the latter affects to a larger extent the entropy estimate of thermodynamic states characterized by a higher motional correlation. The causes of such deviations from a quasi-harmonic behavior are explained. This improved approach provides entropies also for molecular systems undergoing conformational transitions and characterized by highly frustrated energy surfaces, thus not limited to systems sampling a single quasi-harmonic basin. Overall, this study emphasizes the need for extensive phase-space sampling in order to obtain a reliable estimation of entropic contributions.

## 1. Introduction

Entropy is a key property to understand a wide variety of physical, chemical, and biochemical phenomena. However, the estimation of absolute entropies and entropy differences from computer simulations is a long-standing problem[1−9] and one of the current challenges in computational chemistry.[10−15]

The calculation of reliable absolute entropies from molecular dynamics (MD) simulations is intrinsically difficult because the absolute entropy is a measure of the overall extent of phase space (PS) *accessible* to a molecular system. However, absolute single-molecule entropies can be estimated based on an analytical approximation to the configurational probability distribution corresponding to the PS *accessed* by a simulated system.[2] The underlying theory, assumptions, approximations, and alternative practical implementations have been recently reviewed.[10,11] The relationship among quasi-harmonic (QH), essential-mode, and normal-mode analyses has also been investigated.[11] For an extensive review of the subject, not limited to the QH approach, see also refs 11 and 16−19 and references therein.

The difference between the *true* entropy of a simulated system and its QH estimate arises from (i) *anharmonicities* (i.e., non-Gaussian behavior) in the probability distributions along individual QH modes and (ii) *correlations* among the probability distributions associated with different QH modes (beyond the pairwise linear correlations accounted for). These effects are neglected in standard QH analysis[10,11] and (nearly)

* Corresponding author phone: +1-858-534-2913; e-mail: rbaron@mccammon.ucsd.edu.

[†] Department of Chemistry & Biochemistry, University of California San Diego.

[‡] Center for Theoretical Biological Physics, University of California San Diego.

[§] ETH Hönggerberg.

[||] Department of Pharmacology, University of California San Diego.

[⊥] Howard Hughes Medical Institute, University of California San Diego.

10.1021/ct900373z CCC: $40.75 © 2009 American Chemical Society
Published on Web 11/02/2009

always lead to a negative entropy contribution.[11] A method to correct for both artifacts was recently described.[11] Point ii is of particular relevance when trying to estimate entropy differences between two conformational states of a molecular system because error cancellation cannot be guaranteed a priori.[11,20] By taking into account correlation effects of increasing order, entropy estimates based on corrected QH analysis aim at capturing the entropy corresponding to the entire PS sampled (see Figure 1 in ref 11). Thus, this approach is not limited to systems sampling single QH basins and allows capturing conformational transitions.

In the present article, we expand the previous study in ref 11. A general formulation is proposed to account for correction terms of increasing order, and its practical implementation and limitations are discussed. We review previous studies employing this novel approach on an array of (bio)molecular systems providing a solid basis for its application and demonstrating the importance of these correction terms in the evaluation of absolute entropy and entropy differences. Using microsecond MD simulation of a test system, we analyze the convergence properties of the absolute single-molecule entropy and of the correction terms used to estimate it. The results emphasize that sufficient PS sampling is required for a reliable estimation of entropic contributions because convergence of both the QH upper bound and the required correction terms should be achieved.

## 2. Methods

**2.1. QH Analysis.** QH analysis aims to account for motions in the overall extent of PS *accessible* to a molecular system at thermodynamic equilibrium. It relies on approximating the configurational probability distribution as a multivariate Gaussian, the momenta of which can be estimated, e.g., from molecular dynamics (MD) or Monte Carlo simulations.

More precisely, for a given choice of generalized coordinate system $\boldsymbol{q}$ (of dimension $M' = 3N$, $N$ being the number of atoms), its input quantity is the covariance matrix $\underline{\mathbf{C}}_{\boldsymbol{q}}$ characterizing the atom-positional fluctuations (and their correlations) around an average configuration $\bar{\boldsymbol{q}}$. Assuming a canonical ensemble and fluctuations resulting from an underlying harmonic potential of the form

$$\tilde{\mathcal{V}}_h(\boldsymbol{q}) = \frac{1}{2}(\boldsymbol{q} - \tilde{\boldsymbol{q}}_o)^T\, \tilde{\underline{\mathbf{H}}}_{\boldsymbol{q}}(\boldsymbol{q} - \tilde{\boldsymbol{q}}_o) \tag{1}$$

where $\tilde{\underline{\mathbf{H}}}_{\boldsymbol{q}}$ is an effective Hessian matrix and $\tilde{\boldsymbol{q}}_\mathrm{o}$ an effective equilibrium configuration, it follows that[11]

$$\tilde{\boldsymbol{q}}_o = \bar{\boldsymbol{q}} \text{ and } \tilde{\underline{\mathbf{H}}}_{\boldsymbol{q}} = \beta^{-1}\underline{\mathbf{C}}_{\boldsymbol{q}}^{-1} \tag{2}$$

Note that the corresponding harmonic model only strictly produces the correct average configuration $\bar{\boldsymbol{q}}$ and covariance matrix $\underline{\mathbf{C}}_{\boldsymbol{q}}$ for generalized coordinate systems where the mass-metric tensor $\underline{\mathbf{A}}_{\boldsymbol{q}}$ is configuration independent.[11]

In this study, we only consider the specific case of single-molecule entropy (i.e., the entropy of individual distinguishable atoms in a covalently bound molecule) based on MD simulation trajectories. As detailed elsewhere,[21] *single-molecule entropy* differs from *molecular entropy* in that the former estimate only accounts for intermolecular correlation in terms of the effect of the solvent on the single-molecule dynamics.

In practice, the QH analysis of an MD trajectory involves the following steps.[11]

First, the average configuration $\bar{\boldsymbol{q}}$ and the covariance matrix $\underline{\mathbf{C}}_{\boldsymbol{q}}$ in the chosen coordinate system are evaluated as

$$\bar{\boldsymbol{q}} = \langle \boldsymbol{q} \rangle \text{ and } \underline{\mathbf{C}}_{\boldsymbol{q}} = \langle (\boldsymbol{q} - \bar{\boldsymbol{q}}) \otimes (\boldsymbol{q} - \bar{\boldsymbol{q}}) \rangle \tag{3}$$

The equilibrium configuration $\tilde{\boldsymbol{q}}_\mathrm{o}$ and Hessian matrix $\tilde{\underline{\mathbf{H}}}_{\boldsymbol{q}}$ of the effective underlying harmonic model are then defined according to eq 2.

Second, the (symmetric) metric-tensor-weighted covariance matrix is diagonalized

$$\underline{\mathbf{V}}_{\boldsymbol{q}}^T\underline{\mathbf{A}}_{\boldsymbol{q}}^{1/2}\underline{\mathbf{C}}_{\boldsymbol{q}}\underline{\mathbf{A}}_{\boldsymbol{q}}^{1/2}\underline{\mathbf{V}}_{\boldsymbol{q}} = \underline{\mathbf{F}}_{\boldsymbol{q}} \tag{4}$$

where $\underline{\mathbf{V}}_{\boldsymbol{q}}$ is a $M \times M$-dimensional (orthogonal) matrix the columns of which represent the $M'$ components of the eigenvectors $\{\boldsymbol{\nu}_{\boldsymbol{q},m} |\, m = 1, ..., M'\}$ (called QH modes) of the metric-tensor-weighted covariance matrix and $\underline{\mathbf{F}}_{\boldsymbol{q}}$ is a diagonal matrix containing the corresponding eigenvalues. These eigenvalues are related to the associated angular frequencies of the underlying effective harmonic model as (see eqs 2 and 4)

$$\omega_m = (\beta F_{\boldsymbol{q},m})^{-1/2}, m = 1, 2, ..., M \tag{5}$$

The sum of the eigenvalues in $\underline{\mathbf{F}}_{\boldsymbol{q}}$ is equal to the total mean-square metric-tensor-weighted fluctuation of the system, i.e.

$$\begin{aligned} Tr[\underline{\mathbf{F}}_{\boldsymbol{q}}] &= Tr[\underline{\mathbf{A}}_{\boldsymbol{q}}^{1/2}\underline{\mathbf{C}}_{\boldsymbol{q}}\underline{\mathbf{A}}_{\boldsymbol{q}}^{1/2}] \\ &= \langle [\underline{\mathbf{A}}_{\boldsymbol{q}}^{1/2}(\boldsymbol{q} - \bar{\boldsymbol{q}})] \cdot [\underline{\mathbf{A}}_{\boldsymbol{q}}^{1/2}(\boldsymbol{q} - \bar{\boldsymbol{q}})] \rangle \end{aligned} \tag{6}$$

so that the eigenvalues can be interpreted as contributions of individual QH modes to this quantity (a larger value indicating a larger contribution to the total fluctuation of the molecule).

Third, the simulated trajectory is projected onto the QH modes, i.e., one considers the transformed coordinates $\boldsymbol{b}_{\boldsymbol{q}}$ defined as

$$\boldsymbol{b}_{\boldsymbol{q}} = \underline{\mathbf{V}}_{\boldsymbol{q}}^T\underline{\mathbf{A}}_{\boldsymbol{q}}^{1/2}(\boldsymbol{q} - \bar{\boldsymbol{q}}) \tag{7}$$

These so-called QH coordinates satisfy the properties[11]

$$\langle \boldsymbol{b}_{\boldsymbol{q}} \rangle = \mathbf{0} \text{ and } \langle \boldsymbol{b}_{\boldsymbol{q}} \otimes \boldsymbol{b}_{\boldsymbol{q}} \rangle = \underline{\mathbf{V}}_{\boldsymbol{q}}^T\underline{\mathbf{A}}_{\boldsymbol{q}}^{1/2}\underline{\mathbf{C}}_{\boldsymbol{q}}\underline{\mathbf{A}}_{\boldsymbol{q}}^{1/2}\underline{\mathbf{V}}_{\boldsymbol{q}} = \underline{\mathbf{F}}_{\boldsymbol{q}} \tag{8}$$

Because $\underline{\mathbf{F}}_{\boldsymbol{q}}$ is diagonal, eq 8 enforces that the individual components $\{\boldsymbol{b}_{\boldsymbol{q},m} |\, m = 1, ..., M'\}$ of the QH coordinates $\boldsymbol{b}_{\boldsymbol{q}}$ are pairwise linearly uncorrelated, which, however, does not imply the absence of higher order (i.e., pairwise supralinear and higher order) correlations.

We previously motivated the choice of a Cartesian vs internal coordinate system.[11] If a Cartesian coordinate system $\boldsymbol{r}$ is employed[6,8] (after removal of the overall translational and rotational motion from the sampled trajectory[22]), the mass-metric tensor $\underline{\mathbf{A}}_{\boldsymbol{r}}$ is identical to the mass matrix $\underline{\mathbf{M}}$ (thus configuration independent, so that eq 2 is exactly satisfied).

In this case, the QH analysis relies on the diagonalization of the mass-weighted Cartesian covariance matrix, i.e.

$$\underline{\mathbf{D}}_r = \underline{\mathbf{M}}^{1/2}\underline{\mathbf{C}}_r\underline{\mathbf{M}}^{1/2} \tag{9}$$

in place of $\underline{\mathbf{A}}_q^{1/2}\ \underline{\mathbf{C}}_q\ \underline{\mathbf{A}}_q^{1/2}$ in eq 4.

In the absence of geometric constraints, the corresponding eigenvalue matrix $\underline{\mathbf{F}}_r$ contains $3N - 6$ nonzero and 6 vanishing elements. If $N_c$ geometrical constraints are present in the system (e.g., bond-length constraints), these will map to an identical number of zero eigenvalues (see Appendix A in ref 23 for a derivation in the mathematically similar context of essential-mode analysis). Thus, the number of QH modes with nonzero eigenvalues is $M = 3N - N_c - 6$, where $M' = 3N$. When using a generalized coordinate system excluding overall translation and rotation variables, one has $M' = M = 3N - N_c - 6$. Note that the QH coordinates have units of mass$^{1/2}$ × length.[11]

**2.2. Entropies and Correction Terms.** Single-molecule entropies can be obtained as follows.[11] In terms of QH coordinates, the configurational probability distribution associated with the effective harmonic model of eq 2 corresponds to that of $M$ independent harmonic oscillators. Thus, the associated entropy $S_{\mathrm{o}}$ can be calculated analytically. Assuming a canonical ensemble and a configuration-independent mass-weighted metric tensor, this leads to[11]

$$S_{\mathrm{o}} = \sum_{m=1}^{M} s(\omega)\cdot((\beta E_{r,m})^{-1/2}) \tag{10}$$

where $s(\omega)$ is the canonical entropy of a one-dimensional harmonic oscillator with angular frequency $\omega$. The classical expression $s_{cl,o}(\omega)$ and the quantum-mechanical expression $s_{qm,o}(\omega)$ for this quantity are

$$s_{cl,o}(\omega) = k_{\mathrm{B}}(1 - \ln \beta\hbar\omega) \tag{11}$$

and

$$s_{qm,o}(\omega) = k_B\left[\frac{\beta\hbar\omega}{e^{\beta\hbar\omega} - 1} - \ln(1 - e^{-\beta\hbar\omega})\right] \tag{12}$$

where $\hbar = h(2\pi)^{-1}$ is the reduced Planck's constant, leading to eq 10 to corresponding total estimates $S_{cl,o}$ and $S_{qm,o}$, respectively.

In practice, even if the underlying trajectory was generated at the classical level, the QH entropy must be evaluated using the quantum-mechanical oscillator formula because in the high-frequency limit the classical entropy of a one-dimensional harmonic oscillator diverges to the unphysical limit of $-\infty$ rather than to the physical limit of zero.[8,11] However, the QH entropy estimate $S_{qm,o}$ is not the absolute configurational entropy of a single molecule but an upper bound for this quantity due to the presence of QH mode anharmonicities and correlations not accounted for in the effective harmonic model of eq 2. Corresponding correction terms can be formulated exactly at the classical level using an approach previously described[11] and briefly summarized below.

In the canonical ensemble, assuming a configuration-independent mass-metric tensor, the exact classical single-molecule entropy reads[11]

$$S_{cl} = -k_B\left[\frac{M}{2}\left(1 - \ln\frac{\beta h^2}{2\pi}\right) - \int \mathrm{d}\boldsymbol{b}_{\boldsymbol{q}}\, p(\boldsymbol{b}_{\boldsymbol{q}})\ln p(\boldsymbol{b}_{\boldsymbol{q}})\right] \tag{13}$$

where $p(\boldsymbol{b}_{\boldsymbol{q}})$ is the probability distribution in the $M$-dimensional space of the QH coordinates $\boldsymbol{b}_{\boldsymbol{q}}$ (eq 7). This expression can be compared with the approximate (classical) QH estimate $S_{cl,o}$ based on eqs 10 and 11, i.e.

$$S_{cl,o} = -k_{\mathrm{B}}\sum_{m=1}^{M}\left(1 - \frac{1}{2}\ln\frac{\beta\hbar^2}{F_{\boldsymbol{q},m}}\right) \tag{14}$$

A series of increasingly accurate estimates $\{S_{cl,K}| K = 0, 1, ..., M\}$ may now be formulated as

$$S_{cl,K} = S_{cl,o} - k_B\left[\frac{K}{2M}C(K,M)\sum_{m=1}^{M}(1 + \ln 2\pi F_{\boldsymbol{q},m}) + \sum_{c=1}^{C(K,M)}\int \mathrm{d}\boldsymbol{b}_{\boldsymbol{q}}^{(c)} p^{(c)}(\boldsymbol{b}_{\boldsymbol{q}}^{(c)})\ln p^{(c)}(\boldsymbol{b}_{\boldsymbol{q}}^{(c)})\right] \tag{15}$$

where $c$ denotes a combination of $K$ QH modes, $C(K,M) = [(M - K)!K!]M!$ for $K > 0$ along with $C(0,M) = 0$ represents the total number of possible combinations $c$ of $K$ modes among the $M$ QH modes and $p^{(c)}(\boldsymbol{b}_{\boldsymbol{q}}^{(c)})$ is the $K$-dimensional probability distribution in the subspace of the QH coordinates $\boldsymbol{b}_{\boldsymbol{q}}^{(c)}$ within $\boldsymbol{b}_{\boldsymbol{q}}$ that are involved in a combination $c$. The derivation of this equation is given in the Appendix in the Supporting Information.

It is easily verified that $S_{cl,K=0} = S_{cl,o}$ (eq 14, i.e., the uncorrected classical QH entropy) and $S_{cl,K=M} = S_{cl}$ (eq 13, i.e., the exact classical entropy). Substituting the classical estimate $S_{cl,o}$ by the corresponding quantum-mechanical estimate $S_{qm,o}$ (eqs 10 with 12) into eq 15 and introducing successive correction terms defined as

$$\Delta S_{cl,K} = S_{cl,K} - S_{cl,(K-1)} \tag{16}$$

leads to a (classically) corrected QH entropy estimate

$$\tilde{S}^{ctd} = S_{qm,o} + \sum_{K=1}^{M}\Delta S_{cl,K} \tag{17}$$

The successive correction terms of eq 17 involve integrals over the probability distributions $p^{(c)}(\boldsymbol{b}_{\boldsymbol{q}}^{(c)})$ in eq 15 with increasing dimensionality $K$. Note that these terms are all individually negative (or vanishing). The first correction term $\Delta S_{cl,1}$ involves one-dimensional (1D) integrals and accounts for anharmonicities in the individual QH modes. The second correction term $\Delta S_{cl,2}$ involves two-dimensional (2D) integrals and accounts for pairwise (supralinear) correlations between the QH modes. For simplicity, these two terms will be renamed $\Delta S_{cl}^{ah}$ and $\Delta S_{cl}^{pc}$, respectively, to match the notation used in other studies.[11,20,24−28]

The following higher order correction terms account for correlations among QH modes beyond the pairwise ones. Although the classical QH entropy estimate $S_{cl,o}$ usually represents a poor approximation to its quantum-mechanical counterpart $S_{qm,o}$, the evaluation of the correction terms at the classical level remains accurate because anharmonicities and correlations principally affect the low-frequency QH modes for which the classical approximation holds.[11]

The successive correction terms in the series of eq 17 are increasingly difficult to evaluate because both (i) the number of terms $C(K,M)$ involved in the evaluation of $\Delta S_{cl,K}$ and (ii) the sparseness in the required multiple-mode probability distributions $p^{(c)}(\boldsymbol{b}_q{}^{(c)})$ increase exponentially with $K$. For this reason, their evaluation is restricted in practice to the first two terms and implies an intrinsic uncertainty on the final estimate compared to the *true* single-molecule entropy (i.e., persisting in the limit of infinite sampling).

However, note that, in a different context, alternative approximate formulations to estimate terms of increasing order mutual information have been proposed and seem to suggest that the first two correction terms in eq 17 are indeed dominant.[29−31] No study heretofore investigated the convergence properties of these terms along a simulation trajectory.

Following from eqs 16 and 17, the expressions for the first two correction terms are

$$\Delta S_{cl}^{ah} = -k_B\left[\frac{1}{2}\sum_{m=1}^{M}(1+\ln 2\pi F_{\boldsymbol{q},m}) + \sum_{m=1}^{M}\int \mathrm{d}b_{\boldsymbol{q},m}\, p^{(m)}(b_{\boldsymbol{q},m})\ln p^{(m)}(b_{\boldsymbol{q},m})\right] \qquad (18)$$

and

$$\Delta S_{cl}^{pc} = -k_{\mathrm{B}}\left[\frac{M-1}{2}\sum_{m=1}^{M}(1+\ln 2\pi F_{\boldsymbol{q},m}) + \sum_{m=1}^{M}\sum_{n=m+1}^{M}\iint \mathrm{d}b_{\boldsymbol{q},m}\mathrm{d}b_{\boldsymbol{q},n}\, p^{(m,n)}(b_{\boldsymbol{q},m}, b_{\boldsymbol{q},n})\ln p^{(m,n)}(b_{\boldsymbol{q},m}, b_{\boldsymbol{q},n})\right] - \Delta S_{cl}^{ah} \qquad (19)$$

leading to the corrected absolute single-molecule entropy estimate

$$S^{ctd} = S_{qm,o} + \Delta S_{cl}^{crc} = S_{qm,o} + \Delta S_{cl}^{ah} + \Delta S_{cl}^{pc} \qquad (20)$$

The relative magnitudes $f_{cl}^{crc}$, $f_{cl}^{ah}$, and $f_{cl}^{pc}$ of the correction terms $\Delta S_{cl}^{crc}$, $\Delta S_{cl}^{ah}$, and $\Delta S_{cl}^{pc}$ with respect to the QH entropy upper-bound $S_{qm,o}$ (expressed in percent), i.e.

$$f_{cl}^{crc} = \frac{100\Delta S_{cl}^{crc}}{S_{qm,o}}, f_{cl}^{ah} = \frac{100\Delta S_{cl}^{ah}}{S_{qm,o}}\ and\, f_{cl}^{pc} = \frac{100\Delta S_{cl}^{pc}}{S_{qm,o}} \qquad (21)$$

may then serve as a measure for the importance of the aforementioned corrections.

In practice, the 1D and 2D integrals involved in eqs 18 and 19 are evaluated numerically in the form of sums over corresponding histograms. It is reasonable to choose the bin width along a given QH mode in proportion to the width (first moment) of the probability distribution along this mode with proportionality factors $\kappa_1$ and $\kappa_2$ for 1D and 2D integrals, respectively. However, $\kappa_1$ and $\kappa_2$ values must be selected carefully in order to keep both finite-sampling and binning errors to a minimum, i.e., to ensure the independence of the results on these two parameters.[11] For this reason, we monitored the dependence of such numerical integrals on the width of histogram bins for increasing periods of time, as described in section 3.4.

Note, finally, that the absolute single-molecule entropies so far discussed exclude roto-translational contributions. In principle, a translational entropy contribution can be included using the quantum-mechanical expression of the Sackur−Tetrode equation for a specified standard state of the pressure (molecule in the gas phase) or of the concentration (molecule in solution). Similarly, the rotational entropy contribution could be included using the appropriate quantum-mechanical expression (e.g., rigid-rotor approximation, based on the average inertia tensor of the molecule[32−34]). However, these two contributions are likely to be highly coupled with each other and with $\tilde{S}^{ctd}$, i.e., they are not strictly additive, and their rigorous treatment is therefore still challenging. A recent study reported on relatively small effects of motional correlation on changes of reorientational entropy using selected QH modes from a 1.5 ns simulation of the ubiquitin protein.[35] In the present article, single-molecule configurational entropies refer to entropies excluding roto-translational effects.

**2.3. Computational Details.** A 1.1 $\mu$s long MD simulation of the cc$\beta$ peptide ($CH_3$-CO-S-I-R-E-L-E-A-R-I-R-E-L-E-L-R-I-$COO^-$) at 300 K was performed with the AMBER 9 software,[36] the AMBER 99SB parameter set,[37] and the compatible TIP3P water model.[38] The simulation was initialized from the α-helical configuration based on a X-ray model structure of the cc$\beta$ coiled coil (PDB ID 1s9z).[39] Trajectory snapshots were saved every 10 ps for analysis. The simulation setup and trajectory analyses are detailed elsewhere.[40] Backbone atom-positional root-mean-square deviations (RMSD) from the initial folded structure and radius of gyration (RGYR) were calculated using all $C^\alpha$ atoms.

Independent QH analyses were performed for 22 increasingly long segments of the simulation (differing in length by 50 ns) by calculation of the solute all-atom mass-weighted covariance matrix $\underline{\mathbf{D}}_{\boldsymbol{r}}$ (eq 9) in Cartesian coordinates after least-squares fit superposition[22] of all configurations onto the initial structure to eliminate overall translation and rotation and diagonalization (eq 4 with $\underline{\mathbf{A}}_{\boldsymbol{q}}^{1/2}\,\underline{\mathbf{C}}_{\boldsymbol{q}}\,\underline{\mathbf{A}}_{\boldsymbol{q}}^{1/2} = \underline{\mathbf{D}}_{\boldsymbol{r}}$). A total of 534 ($M = 3 \times 297 - 351 - 6$) modes associated with nonvanishing eigenvalues were considered. After determination of the QH modes (columns of the matrix $\underline{\mathbf{V}}_{\boldsymbol{r}}$ in eq 4; sorted in order of decreasing eigenvalues, i.e., increasing $\omega_m$ frequency in eq 5), the trajectory was projected in this basis set to obtain the time series of the corresponding QH coordinates $\boldsymbol{b}_{\boldsymbol{r}}$ (eq 7). This first part of the analysis was performed using the S_correction program as implemented in the gromos++ module of the GROMOS05 software[41] for biomolecular simulation.

The QH entropy upper bound, $S_{qm,o}$ (eq 10 with eq 12), the corrections for mode anharmonicity, $\Delta S_{cl}^{ah}$ (eq 18), and pairwise supralinear mode correlation, $\Delta S_{cl}^{pc}$ (eq 19), their sum, $\Delta S_{cl}^{crc}$ (eq 20), the improved absolute single-molecule entropy, $S^{ctd}$ (eq 20), the relative terms $f_{cl}^{crc}$, $f_{cl}^{ah}$, and $f_{cl}^{pc}$ (eq

21), and the sum of the eigenvalues, $Tr[\underline{\mathbf{F}}_r]$ (eq 6), were then calculated for each of the 22 trajectory segments.

Note that this analysis is computationally intensive because, as discussed in section 3.3, each of the 22 $\Delta S_{cl}^{ah}$ values requires the estimation of 534 1D integrals, while each of the 22 $\Delta S_{cl}^{pc}$ values requires the estimation of 142 311 2D integrals (eq 15). In addition, for each of these integrals the optimized proportionality factors $\kappa_1$ and $\kappa_2$ were determined based on multiple integral calculations for an accurate numerical integration (see section 3.4). As an indication of the actual computational cost, using an Intel Xeon X5450 3.0 GHz, dedicated software, and the above procedure, each 1D or 2D integral can be estimated with an average CPU time of 0.07 or 0.21 s, respectively, from 50 ns trajectory windows. Overall, the analyses presented in this work require a CPU time that sums up to ~6 months.

## 3. Results and Discussion

**3.1. Review of Previous Studies.** The key findings of previous studies concerning the uncorrected QH upper bound, $S_{qm,o}$ (eq 10 with eq 12), the improved absolute single-molecule entropy $S^{ctd}$ (eq 20), and the relative magnitude of the cumulative correction term $f_{cl}^{crc}$ (eq 21) are summarized graphically in Figure 1.

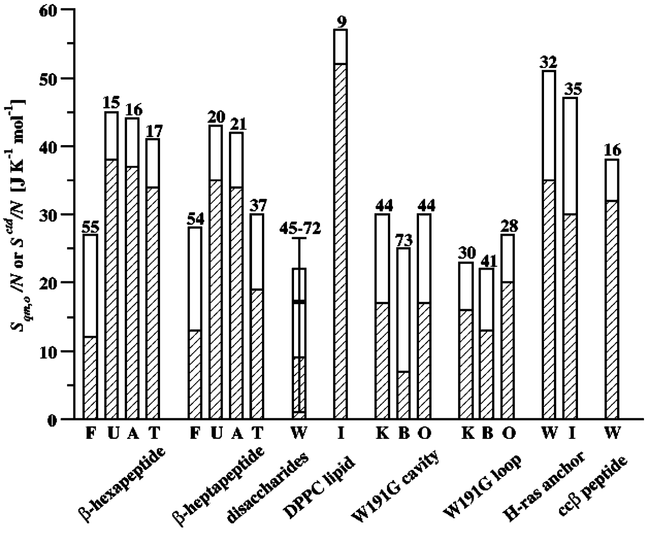


*Figure 1.* Summary of previous studies investigating the improved absolute single-molecule entropy. The QH entropy upper-bound $S_{qm,o}$ (eq 10 with eq 12; empty bars) and the improved entropy estimate $S^{ctd}$ (eq 20; hatched areas) are displayed scaled by the number $N$ of system particles. The relative (%) values of the cumulative correction term $f_{cl}^{crc}$ (eq 21; bar labels) are reported as a measure of the importance of the deviation from the QH approximation. From left to right: two $\beta$-peptides in methanol at high temperature (F, folded; U, unfolded; A, all; T, 298 instead of 340 K),[11] the 11 glucose-based disaccharides (W, free in water),[28] dipalmitoylphosphatidylcholine, DPPC (I, inserted in a hydrated bilayer),[24] the cavity and its gating loop of the W191G mutant of cytochrome $c$ peroxidase (K, bound to a K$^+$ ion with closed gating loop; B, bound to 2-amino-5-methylthiazole with closed gating loop; O, bound to a K$^+$ ion with open gating loop),[20] the H-ras lipopetide anchor,[27] and the cc$\beta$-peptide (this study). For the disaccharides,[28] corresponding mean entropy values are displayed (a vertical bar represents the range of values). See Supporting Information for details.

These results span systems with different chemical nature: 2 $\beta$-peptides in methanol,[11] the 11 disaccharides of gluscose in water,[28] the dipalmitoylphosphatidylcholine (DPPC) lipid in a hydrated bilayer,[24] the W191G mutant cavity and its gating loop within cytochrome $c$ peroxidase in water,[20] the H-ras lipopetide anchor in water or inserted into a model lipid membrane,[27] and the cc$\beta$-peptide in water (this study). In some cases, the QH analysis was also performed separately for different chemical environments or conformational states of the molecule, which permits estimating relative entropies, thereby quantifying the impact of the correction terms on the thermodynamic process of interest. These processes include reversible peptide folding (ref 11 and this study), conformational changes in carbohydrates[28] and lipids,[24] lipopeptide insertion in a model membrane bilayer,[27] ligand binding to a protein cavity,[20] and protein-loop gating.[20] The results presented in Figure 1 are scaled by the number, $N$, of atoms to allow for a comparison among molecules of varying size (the raw data is available as Supporting Information, Table S1).

Some clear qualitative trends are evident, although a direct comparison among these studies is not possible due to the different MD time scales and physicochemical conditions. In all systems the cumulative correction term $\Delta S_{cl}^{crc}$ (eq 20) is generally sizable, demonstrating an overall large deviation from a QH behavior as evaluated up to the pairwise supralinear level. The corresponding relative magnitudes, $f_{cl}^{crc}$, display values from 9% to 73% of the QH upper-bound value $S_{qm,o}$ (Figure 1). In detail, these important cumulative terms result from the sum of correction terms for mode anharmonicity ($\Delta S_{cl}^{ah}$; eq 18) that are always relatively small (up to 3% of the upper-bound value $S_{qm,o}$) and for pairwise supralinear mode correlation ($\Delta S_{cl}^{pc}$; eq 19) that are always dominant. The latter correction term has a magnitude that depends on the physical nature of the molecular motional correlation experienced by the molecular system in a given thermodynamic state.

The largest relative corrections, $f_{cl}^{crc}$, are expected and found for intrinsically more ordered systems (Figure 1). This can be explained by considering that restricted flexibility is typically promoted by inter- and/or intramolecular interactions, simultaneously inducing increased motional correlation. For example, the ligand-bound state of the W191G protein cavity[20] displays the largest $f_{cl}^{crc}$ value (73%), i.e., the thermodynamic ensemble involving the largest motional correlations and lowest entropy content among those studied. On the other end of the spectrum and in line with this qualitative picture, the smallest $f_{cl}^{crc}$ values were reported for the DPPC lipid in a bilayer (9%),[24] i.e., the ensemble characterized by the highest molecular flexibility and thus the lowest motional correlations. Interestingly, the 11 disaccharides of glucose in water[28] display high variability and always large $f_{cl}^{crc}$ values (45−72%). This behavior can be explained considering that these molecules involve a reduced number of degrees of freedom overall and the linkage between rather stiff glucose rings is the torsion defining major conformational changes.[25,26]

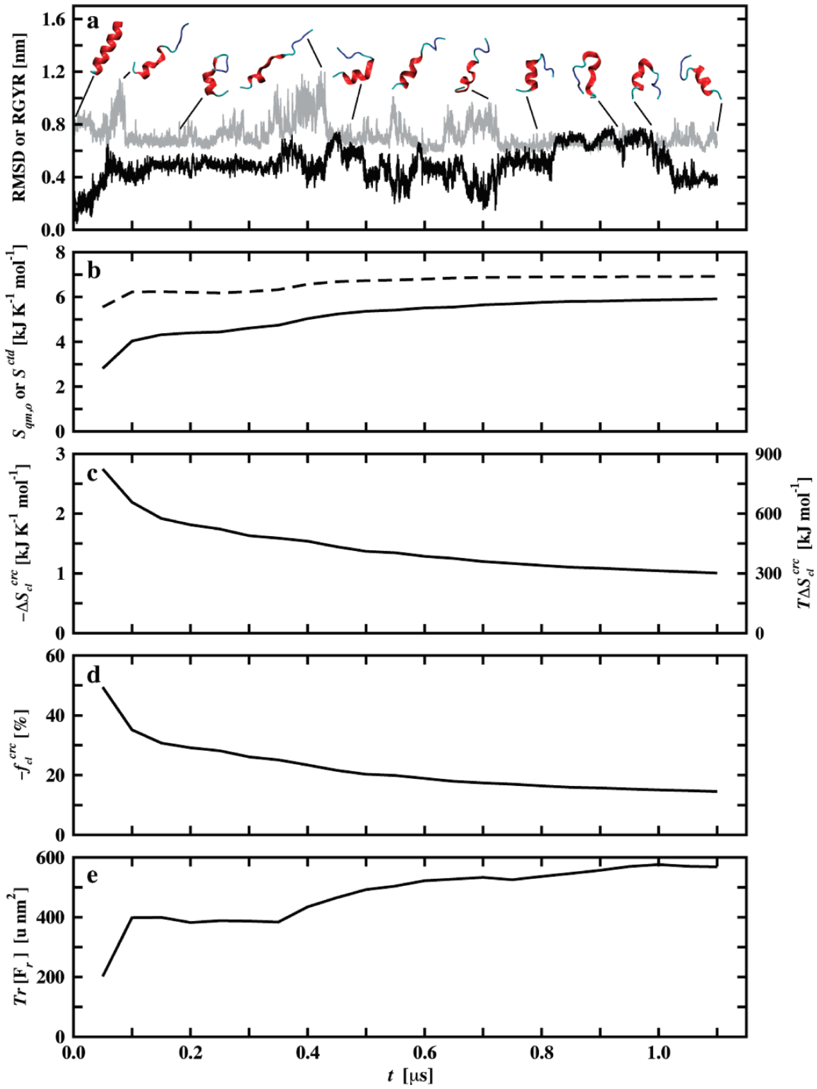


***Figure 2.*** cc$\beta$ peptide dynamics on the microseconds time scale and entropy convergence. (a) The backbone atom-positional root-mean-square deviation (rmsd; black) from the initial helical fold and of the backbone radius of gyration (RGYR; gray) are shown along the time, $t$. The cartoon representations highlight example configurations (oriented with the $CH_3$−CO terminus down). (b) Build-up curves of the QH entropy upper bound $S_{qm,o}$ (eqs 10 with 12; dashed line) and of the improved absolute single-molecule entropy $S^{ctd}$ (eq 20; solid line). Convergence of (c) the cumulative correction term $\Delta S_{cl}^{crc}$ (eq 20) and its contribution to the free energy $T\Delta S_{cl}^{crc}$, (d) its relative value $f_{cl}^{crc}$ (eq 21), and (e) the sum of the eigenvalues $Tr[\underline{\mathbf{F}}_r]$ (eq 6).

These qualitative trends are also in agreement with the observation that entropy is the measure of PS sampling for a molecular system. The QH upper bound, $S_{qm,o}$, and the improved absolute single-molecule entropy, $S^{ctd}$, are estimated based on the PS that has been *accessed* during a MD simulation of finite time scale, i.e., only a fraction of the PS *accessible* to the system. These time scales ranged from 50 ns (W191G mutant[20] showing the lowest entropy) to 25.6 $\mu$s (concatenated trajectory of the DPPC lipid,[24] showing the largest entropy). However, this hampers a quantitative comparison of $S_{qm,o}$, $S^{ctd}$, and $\Delta S_{cl}^{crc}$ values among previous studies.

Prompted by these observations, the dependence of these quantities on the extent of accessed PS was assessed on the microsecond time scale for a peptide test system.

**3.2. Convergence of the QH Analysis.** The cc$\beta$ peptide in water was chosen as a test system to investigate entropy convergence properties because of its small size and broad PS accessibility. Figure 2a shows the time series of the backbone atom-positional root-mean-square deviation (RMSD) from the folded structure and of the backbone radius of gyration (RGYR) along a 1.1 $\mu$s of MD simulation. The peptide undergoes several reversible folding/unfolding events and samples a variety of unfolded configurations and compact folds.[40]

The probability distributions $p^{(m)}(\boldsymbol{b}_{r,m})$ of the transformed QH coordinates $\boldsymbol{b}_r$ (eq 7) along selected QH modes ($m = 1$, 2, 6, 10, 50, and 500) are shown in Figure 3a. The reference Gaussian functions with identical variances and vanishing averages are also represented. The actual distributions become increasingly narrow and similar to the Gaussian functions for higher $m$ indices, i.e., the corresponding QH modes become increasingly stiff and harmonic. However, the distributions along the lowest frequency modes (e.g., Figure 3a, $m = 1$ or 2) differ significantly from Gaussian functions and evidently result from the superposition of

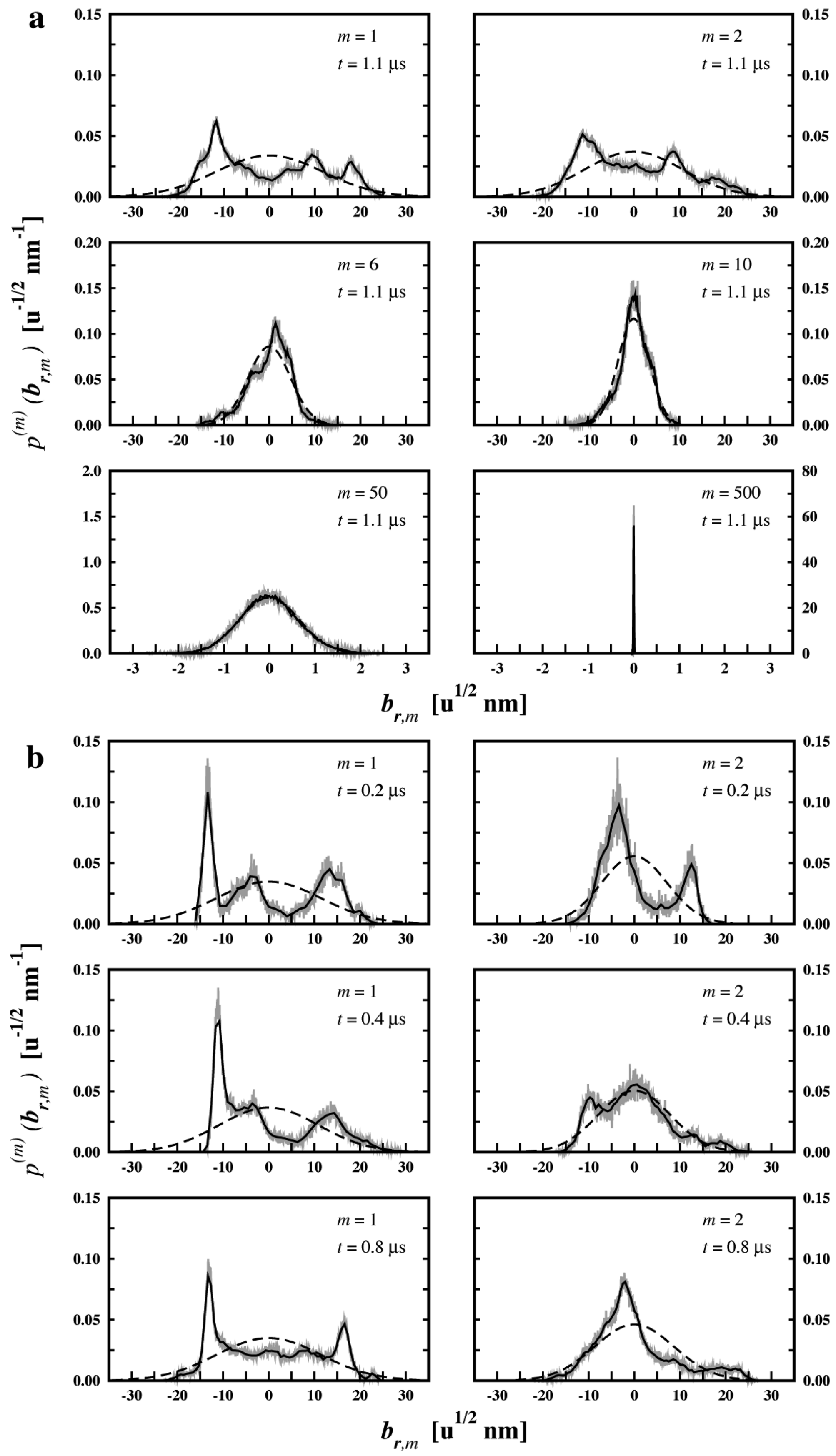


***Figure 3.*** Probability distributions along selected components of the QH coordinate $b$ for the cc$\beta$ peptide. The actual distributions (gray line) are displayed together with the corresponding Gaussians, i.e., $p'_{o,m}(b_{r,m}) = (2\pi F_{r,m})^{-1/2} e^{-1/2\, F_{r,m}^{-1}\, b_{r,m}^{2}}$ (dashed lines) for (a) increasing component indices, $m$, and (b) increasing periods of time, $t$. The distributions employed for optimal numerical integration of the actual distributions (eq 18) are also displayed (solid lines). All probability distributions are normalized. Note that different scaling may be employed for graphical purposes. The letter "u" stands for atomic mass unit.

multiple off-center Gaussian-like distributions. A similar observation was previously reported in the context of two $\beta$-peptides in methanol for which two main subensembles of folded and unfolded configurations could be disentangled based on the lowest frequency modes (Figures 3−5 in ref 11). In the present case, the most pronounced peaks for the cc$\beta$ peptide arise from folded configurations (see Figure 3a, $m$ = 1 and 2, leftmost peak).

The time dependence of these results was investigated as summarized in Figure 3b for the two coordinates $\boldsymbol{b}_r$ with lowest frequencies, i.e., those contributing the most to the total mean-square metric-tensor-weighted fluctuation of the system (eq 6). The corresponding distributions vary significantly with the extent of PS sampling, as revealed by averaging over the first 0.2, 0.4, and 0.8 $\mu$s periods, or over the entire 1.1 $\mu$s ensemble (cf. Figure 3b vs Figure 3a for $m$ = 1 and 2). Increasing the simulation time results into broader distributions due to the larger extent of PS sampled. The intensity of the leftmost peak, corresponding to the contribution of the folded configurations, clearly reduces along the simulation initialized from the cc$\beta$ helical fold and evolving through a broad range of heterogeneous configurations (Figure 1a). The data indicate that convergence of the probability distributions associated with the low-frequency QH coordinates in $\boldsymbol{b}_r$ requires sampling times longer than 1 $\mu$s, considering that such differences persist when comparing results from the first 800 ns period with the whole 1.1 $\mu$s simulation.

**3.3. Entropy Convergence.** The entropy convergence for the cc$\beta$ peptide in water as a function of sampling time is illustrated in Figure 2b. The upper-bound curve obtained by application of the uncorrected QH formula ($S_{qm,o}$; eq 10 with eq 12) is compared to the build-up curve of the improved absolute single-molecule entropy ($S^{ctd}$; eq 20). Both curves require periods of several hundred nanoseconds to reach a first plateau. Interestingly, the QH upper bound reaches convergence noticeably faster than the improved absolute single-molecule entropy. In detail, ~0.3 or ~0.7 $\mu$s is needed to sample 90% or 99% of the final $S_{qm,o}$ estimate of 6922 J K$^{-1}$ mol$^{-1}$, while larger sampling times of ~0.5 or ~1.0 $\mu$s are needed to sample 90% or 99% of the final $S^{ctd}$ estimate of 5916 J K$^{-1}$ mol$^{-1}$ (see also Supporting Information, Table S2). These results clearly demonstrate that the convergence of the $S_{qm,o}$ upper bound does not imply the convergence of the absolute single-molecule entropy, $S^{ctd}$. In fact, the first quantity relies on the convergence of linear motional correlations only (following from the definition of linearly independent QH modes in eq 7 with eq 8). Instead, as described below, the second quantity requires in addition the convergence of supralinear motional correlations. For this reason, $S^{ctd}$ seems to represent a better indicator of convergence (compared to $S_{qm,o}$) for the absolute single-molecule entropy.

The convergence behavior of the cumulative correction term, $\Delta S_{cl}^{crc}$ (eq 20), was also monitored (Figure 2c). This entropy term describes the overall deviation from the QH model due to mode anharmonicity ($\Delta S_{cl}^{ah}$, eq 18; 4 J K$^{-1}$ mol$^{-1}$ after 1.1 $\mu$s) and correlation ($\Delta S_{cl}^{pc}$, eq 19; 1002 J K$^{-1}$ mol$^{-1}$ after 1.1 $\mu$s) effects, associated with all unique combinations of modes $m$ and $n$. Its convergence behavior can be used as well as a measure of uncertainty on the entropy estimate. For comparison with previous studies (Figure 1), the time evolution of the corresponding relative contribution $f_{cl}^{crc}$ (eq 21) to the uncorrected $S_{qm,o}$ estimate (eq 10 with eq 12) is also displayed (Figure 2d). In all studies, the dominant part of this correction arises from QH pairwise (supralinear) mode correlations ($f_{cl}^{ah}$ values are <0.05%; see also Supporting Information, Tables S1 and S2).

Importantly, it is found that the magnitude of the correction term $\Delta S_{cl}^{crc}$ monotonically decreases from an initial value of −2745 J K$^{-1}$ mol$^{-1}$ (first 0.5 $\mu$s sampling) to a final value of −1006 J K$^{-1}$ mol$^{-1}$ (entire 1.1 $\mu$s sampling), showing an initial convergence behavior. This suggests that limited PS sampling results in both the underestimation of $S_{qm,o}$ and the overestimation of $\Delta S_{cl}^{crc}$ (predominantly through its $\Delta S_{cl}^{pc}$ component), both artifacts leading to an underestimation of the final absolute single-molecule entropy $S^{ctd}$.

This result can be explained considering that motional correlations are larger for a molecular system sampling a confined part of PS as opposed to sampling of a multiple-minima landscape. For the cc$\beta$ test system these results demonstrate that a limited PS sampling leads to the overestimation of corresponding motional correlations, thus of $\Delta S_{cl}^{pc}$ values with respect to that expected for a canonical ensemble of the same system at thermodynamic equilibrium.

The mass-weigthed root-mean-squared fluctuation, i.e., the sum of the eigenvalues of the mass-weighted covariance matrix, $Tr[\underline{\mathbf{F}}_r]$ (eq 9), was also monitored along time as an independent measure of convergence (Figure 2e). In terms of $Tr[\underline{\mathbf{F}}_r]$, we note that a first plateau region is reached after ~1.1 $\mu$s (Figure 2e), in line with the $S^{ctd}$ values (Figure 2b). This observation confirms as well that the cc$\beta$ peptide is not trapped in a few local minima. Instead, it explores new configurations even after several hundreds of nanoseconds (Figure 2a).

Three important general points are worth noting.

First, the magnitude of the cumulative correction term $\Delta S_{cl}^{crc}$ is large. This is evident when the term is expressed in the form of its contribution to the system free energy, $T\Delta S_{cl}^{crc}$ (Figure 2c, right axis). The resulting value (302 kJ mol$^{-1}$ based on 1.1 $\mu$s) is about an order of magnitude larger than the free energy changes of typical (bio)chemical processes. Thus, although partial cancellation of this term can be expected for entropy differences between two different molecular environments or conformational states, small differences will still lead to large free-energy contributions (of sign and magnitude difficult to be predicted a priori). The importance of this correction for reliable entropy calculations is therefore evident. In addition, this result suggests that time convergence of the entropy estimate should be taken into account as well when comparing the efficiency and accuracy of alternative computational approaches.

Second, we stress that all $M$ per-mode contributions need to be included for an accurate estimation of eqs 10 and 18 because modes with large $m$ indices (high frequencies) also contribute to $S_{qm,o}$ (data not shown; see Figure 8 in ref 11 for a similar analysis). This marks a difference with what is typically observed for the contribution of a reduced number of essential modes to the total system fluctuation.[23] Due to the similar mathematical formalism,[11] this argument can be easily demonstrated as well for the calculation of entropies from normal-mode analysis for systems sampling one local PS minimum.

Third, the analysis of the leading correction $\Delta S_{cl}^{pc}$ in terms of all $C(2,M) = [(M - 2)!2!]M!$ unique pair combinations reveals that not only modes with low indexes (high amplitudes, low frequencies) contribute substantially. Thus, all pairs of QH modes need to be considered in eqs 15 and 19. This requirement arises from the observation that high correlations can be present among modes with either large or small $m$ and $n$ indices (low or high frequencies; Figure 12 in ref 11). Interestingly, this behavior was observed for highly flexible systems (e.g., the cc$\beta$ peptide of this study or the reversibly folding $\beta$-peptides in ref 11) but not for more rigid systems confined to a local PS sampling (e.g., the W191G cavity in ref 20, unpublished results). Whether the latter result depends on a limited PS accessed or on the physical nature of QH-mode correlation remains to be addressed.

**3.4. Numerical Integration of the Correction Terms.** The analysis presented in this work relies on the numerical integration of the actual probability distributions $p^{(m)}(\boldsymbol{b}_{r,m})$ and $p^{(m,n)}(\boldsymbol{b}_{r,m},\boldsymbol{b}_{r,n})$ evaluated based on the MD trajectory (eqs 18 and 19). Two alternative procedures were described to estimate these 1D and 2D integrals with optimal (nonarbitrary) histogram bin widths, as detailed in Appendix C of ref 11. In this study, optimal parameters $\kappa_1{}^{o}$ and $\kappa_2{}^{o}$ were chosen as the midpoint between the intersections of a horizontal line with the limiting lines for too *fine* and too *coarse* integration at the optimal value of the 1D or 2D integrals in the graph showing these values as a function of ln $\kappa_1$ or ln $\kappa_2$, as summarized in Figure 4.

Figure 4a shows the values of the 1D integrals for a sample set of eigenvectors ($m$ = 1, 50, and 500), evaluated numerically using different values of $\kappa_1$. Both limiting lines are shown, together with the optimal $\kappa_1{}^{o}$ values. Values approaching these limiting curves are incorrect because they show a dependence of the evaluated integral on the bin size. However, for each curve, a clear plateau defines the range of $\kappa_1$ values for which the integration result is essentially independent of the bin size. Finite-sampling artifacts affect the integration with the smallest values of $\kappa_1$, while coarse-binning artifacts affect the integration with the largest values. Note that 1D integrals may be individually negative or positive.

Figure 4b shows the values of the 2D integrals for a sample group of eigenvector pairs ($m$,$n$ = 1,2; 1,100; 1,500), evaluated numerically with different values of $\kappa_2$, together with the corresponding limiting lines and the optimal $\kappa_2{}^{o}$ values. Here, the plateau regions are narrower and the value of $\kappa_2$ has to be chosen more carefully. Note that 2D integrals are always negative when estimated using the optimal $\kappa_2{}^{o}$ values, but incorrect positive values would be obtained based on too small $\kappa_2$ values.

The dependence of these curves on the simulation time was also monitored (Figure 4). All 1D and 2D curves show a coarse-integration limit that is essentially independent of the MD period considered. Yet, they also show that the fine-integration-limiting curves shift to lower $\kappa_1$ and $\kappa_2$ values upon increasing the simulation time, thus reducing the dependence on the integration bin size. This effect is more pronounced for the 2D integrals because they require more data points than 1D integrals.

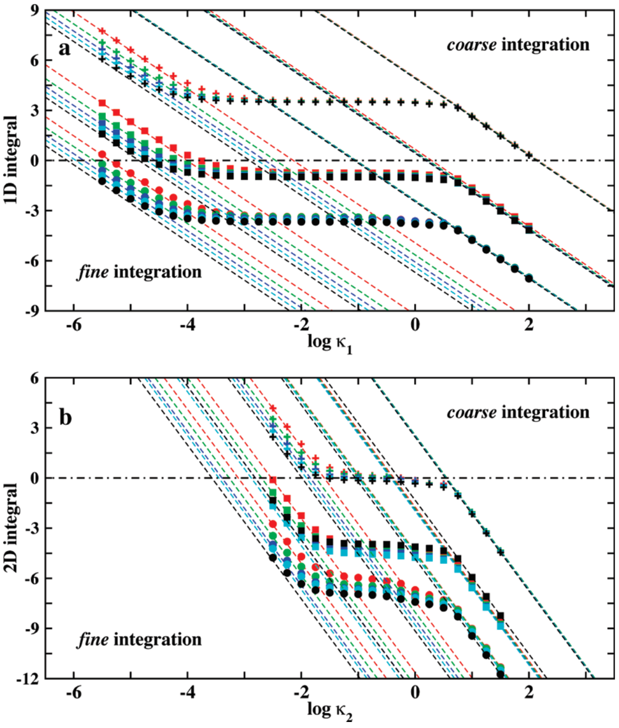


*Figure 4.* Dependence of the numerical integration of probability distributions on the width of histogram bins for increasing periods of time, $t$. (a) Integrals over the 1D distributions involved in eq 18 are shown for eigenvectors 1 (circles), 50 (squares), and 500 (crosses). (b) Integrals over the 2D distributions involved in eq 19 are shown for eigenvector pairs 1,2 (circles), 1,100 (squares), and 1,500 (crosses). The results are displayed for the whole ensemble (1.1 $\mu$s; black) or $t$ = 200 (red), 400 (green), 600 (blue), 800 (cyan) ns as a function of ln $\kappa_1$ (a) or ln $\kappa_2$ (b), where $\kappa$ is the ratio of the bin width along each dimension to the corresponding distribution width. The middle point between a pair of limiting lines for too fine (left side) and too coarse (right side) numerical integrations (dashed lines) defines optimal $\kappa_1{}^{o}$ or $\kappa_2{}^{o}$ values. A black dot−dashed reference line is drawn at zero.

Overall, these results demonstrate that the procedure employed in this work allows estimating both 1D and 2D integrals of eqs 18 and 19 in a nonarbitrary way as a function of the simulation time. The presence of plateau regions independent of the integration bin size for all MD periods considered shows that the observed change of integral values along the simulation time largely depends on the extent of PS sampled yet not on the numerical procedure employed. A similar analysis performed in the case of corresponding 3D integrals (triplewise combinations of QH modes) revealed that indeed no such behavior can be achieved, although using a 1.1 $\mu$s trajectory. In practice, eq 15 can be only estimated for the first two terms owing to finite sampling artifacts and data sparseness.

## 4. Conclusion

The theory and practical implementation of an approach recently proposed[11] to estimate improved configurational entropies from quasi-harmonic analysis of molecular dynamics simulations are briefly reviewed. It involves the calculation of correction terms of increasingly high order to account for deviations from the quasi-harmonic approximation in frustrated molecular systems. The convergence properties of the absolute single-molecule entropy are critically investigated using microsecond molecular dynamics simulation of the cc$\beta$ peptide in water. Prompted by the comparison of the results with previous studies addressing mode anharmonicity and correlation effects, the convergence behavior of individual quasi-harmonic modes, of the absolute single-molecule entropy, and of the correction terms for anharmonicity and pairwise (supralinear) correlations are analyzed. Our data provide a number of new insights to tackle the challenge of accurate entropy estimation by computer simulation.

In line with a previous study,[11] the probability distributions associated with components of the quasi-harmonic coordinates only deviate significantly from Gaussian functions for the first few components, resembling the behavior observed in the different context of a single-atom displacement for an α-helical peptide[42] and of essential modes for protein dynamics.[23] For these components, the probability distributions result from a superposition of clearly distinguishable contributions from the folded and unfolded ensembles. However, it is shown that the components of these eigenvectors converge slowly (>1 $\mu$s), consistent with the observation that the cc$\beta$ peptide steadily explores new configurations.

In line with previous studies,[11,20,25,26,28] the entropic contribution of anharmonicity is small while the pairwise (supralinear) correlation correction to the entropy is large. The deviation from the quasi-harmonic assumption affects more significantly conformational states dominated by high motional correlation. Using microsecond molecular dynamics simulation of a peptide test system we show that limited phase-space sampling results in an overestimation of correlation effects, and we discuss its implications for entropy estimation.

This study demonstrates that the convergence of the quasi-harmonic upper-bound entropy with simulation time does not imply the convergence of the system absolute single-molecule entropy. As a consequence, our study also suggests that the convergence of the absolute single-molecule entropy rather than that of the quasi-harmonic upper bound should be preferably monitored. Because the cumulative correction term accounting for both mode anharmonicity and pairwise (supralinear) correlation effects converges slowly and monotonically decreases, previous studies based on shorter time scales may have, in some cases, partly overestimated this correction term, thus leading to underestimated absolute entropy estimates.

Overall, the present study emphasizes the need of sufficient phase-space sampling to estimate entropic contributions from computer simulations. Ideally, only thermodynamic ensembles at equilibrium should be considered to this end, i.e., full phase-space sampling obtained from simulations on time scales of several microseconds. In practice, we suggest that enhanced sampling techniques[28,43] and/or concatenated copies of independent simulation trajectories[21,24] will be useful tools to alleviate these problems in the future if properly combined with the correction terms used herein.[28] This strategy will open the possibility to include as well correlation effects of higher order than the pairwise (supralinear) explicitly considered in this study. A bright future opens for the estimation of accurate thermodynamic properties for biomolecular systems using chemical theory and computation.

**Acknowledgment.** We thank the members of the McCammon group for useful discussions. This work was supported by the National Science Foundation grant PHY-0822283 and Center for Theoretical Biological Physics for the computing resources, the National Institutes of Health and National Science Foundation to R.B. and J.A.M., and the Howard Hughes Medical Institute and the National Biomedical Computing Resource to J.A.M.

**Supporting Information Available:** Summary of the single-molecule absolute entropy for systems of different chemical nature, Table S1; entropy and correction values along microsecond molecular dynamics of the cc$\beta$ peptide in water, Table S2; derivation of eq 15, Appendix. This material is available free of charge via the Internet at http://pubs.acs.org.

### References

(1) Gõ, N.; Scheraga, H. A. *J. Chem. Phys.* **1969**, *51*, 4751.

(2) Karplus, M.; Kushick, J. N. *Macromolecules* **1981**, *14*, 325.

(3) Di Nola, A.; Berendsen, H. J. C.; Edholm, O. *Macromolecules* **1984**, *17*, 2044.

(4) Edholm, O.; Berendsen, H. J. C. *Mol. Phys.* **1984**, *51*, 1011.

(5) Rojas, O. L.; Levy, R. M.; Szabo, A. *J. Chem. Phys.* **1986**, *85*, 1037.

(6) Schlitter, J. *Chem. Phys. Lett.* **1993**, *215*, 617.

(7) Schafer, H.; Mark, A. E.; van Gunsteren, W. F. *J. Chem. Phys.* **2000**, *113*, 7809.

(8) Andricioaei, I.; Karplus, M. *J. Chem. Phys.* **2001**, *115*, 6289.

(9) Reinhardt, W. P.; Miller, M. A.; Amon, L. M. *Acc. Chem. Res.* **2001**, *34*, 607.

(10) Chang, C. E.; Chen, W.; Gilson, M. K. *J. Chem. Theory Comput.* **2005**, *1*, 1017.

(11) Baron, R.; van Gunsteren, W. F.; Hünenberger, P. H. *Trends Phys. Chem.* **2006**, *11*, 87.

(12) Wang, J.; Bruschweiler, R. *J. Chem. Theory Comput.* **2006**, *2*, 18.

(13) Chang, C. E.; Chen, W.; Gilson, M. K. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1534.

(14) Killian, B. J.; Yundenfreund Kravitz, J.; Gilson, M. K. *J. Chem. Phys.* **2007**, *127*, 24107.

(15) Hnizdo, V.; Tan, J.; Killian, B. J.; Gilson, M. K. *J. Comput. Chem.* **2008**, *29*, 1605.

(16) Carlsson, J.; Åqvist, J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5385.

(17) van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glattli, A.; Hünenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F.; Yu, H. B.*Angew. Chem., Int. Ed.* **2006**, *45*, 4064.

(18) Meirovitch, H. *Curr. Opin. Struct. Biol.* **2007**, *17*, 181.

(19) Meirovitch, H.; Cheluvaraja, S.; White, R. P. *Curr. Protein Pept. Sci.* **2009**, *10*, 229.

(20) Baron, R.; McCammon, J. A. *ChemPhysChem* **2008**, *9*, 983.

(21) Baron, R.; de Vries, A. H.; Hünenberger, P. H.; van Gunsteren, W. F. *J. Phys. Chem. B* **2006**, *110*, 8464.

(22) McLachlan, A. D. *J. Mol. Biol.* **1979**, *128*, 49.

(23) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins* **1993**, *17*, 412.

(24) Baron, R.; de Vries, A. H.; Hünenberger, P. H.; van Gunsteren, W. F. *J. Phys. Chem. B* **2006**, *110*, 15602.

(25) Pereira, C. S.; Kony, D.; Baron, R.; Muller, M.; van Gunsteren, W. F.; Hünenberger, P. H. *Biophys. J.* **2006**, *90*, 4337.

(26) Pereira, C. S.; Kony, D.; Baron, R.; Muller, M.; van Gunsteren, W. F.; Hünenberger, P. H. *Biophys. J.* **2007**, *93*, 706.

(27) Gorfe, A. A.; Baron, R.; McCammon, J. A. *Biophys. J.* **2008**, *95*, 3269.

(28) Peric-Hassler, L.; Hansen, H. S.; Baron, R.; Hünenberger, P. H. Manuscript in preparation.

(29) Somani, S.; Killian, B. J.; Gilson, M. K. *J. Chem. Phys.* **2009**, *130*, 134102.

(30) Matsuda, H. *Phys. Rev. E* **2000**, *62*, 3096.

(31) Numata, J.; Wan, M.; Knapp, E. W. *Genome Inform.* **2007**, *18*, 192.

(32) Carlsson, J.; Åqvist, J. *J. Phys. Chem. B* **2005**, *109*, 6448.

(33) Darian, E.; Hnizdo, V.; Fedorowicz, A.; Singh, H.; Demchuk, E. *J. Comput. Chem.* **2005**, *26*, 651.

(34) Lu, B. Z.; Wong, C. F. *Biopolymers* **2005**, *79*, 277.

(35) Prompers, J. J.; Bruschweiler, R. *J. Phys. Chem. B* **2000**, *104*, 11416.

(36) Case, D. A.; Darden, T.; Cheatham, T. III; Simmerling, C.; Wang, J.; Duke, R.; Luo, R.; Merz, K.; Pearlman, D.; Crowley, M.; Walker, R.; Zhang, W.; Wang, B.; Hayik, A.; Roiberg, A.; Seabra, G.; Wong, K.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Morgan, J.; Hornak, V.; Cui, G.; Beroza, P.; Matthews, D.; Schfmeister, C.; Ross, W.; Kollman, P. *AMBER 9*; University of California: San Francisco 2006.

(37) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712.

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

(39) Kammerer, R. A.; Kostrewa, D.; Zurdo, J.; Detken, A.; Garcia-Echeverria, C.; Green, J. D.; Muller, S. A.; Meier, B. H.; Winkler, F. K.; Dobson, C. M.; Steinmetz, M. O. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4435.

(40) Baron, R.; de Oliveira, C. A. F.; McCammon, J. A. Manuscript in preparation.

(41) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Burgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Krautler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. *J. Comput. Chem.* **2005**, *26*, 1719.

(42) Perahia, D.; Levy, R. M.; Karplus, M. *Biopolymers* **1990**, *29*, 645.

(43) Minh, D. D.; Hamelberg, D.; McCammon, J. A. *J. Chem. Phys.* **2007**, *127*, 154105.

CT900373Z

JCTC Journal of Chemical Theory and Computation

# Performance of Kinetic Energy Functionals for Interaction Energies in a Subsystem Formulation of Density Functional Theory

Andreas W. Götz,*[†,‡] S. Maya Beyhan,[†] and Lucas Visscher*[†]

*VU University Amsterdam, Theoretical Chemistry, De Boelelaan 1083, 1081 HV Amsterdam, The Netherlands*

Received April 11, 2009

**Abstract:** We have tested the performance of a large set of kinetic energy density functionals of the local density approximation (LDA), the gradient expansion approximation (GEA), and the generalized gradient approximation (GGA) for the calculation of interaction energies within a subsystem approach to density functional theory. Our results have been obtained with a new implementation of interaction energies for frozen-density embedding into the Amsterdam Density Functional program. We present data for a representative sample of 39 intermolecular complexes and 15 transition metal coordination compounds with interaction energies spanning the range from −1 to −783 kcal/mol. This is the first time that kinetic energy functionals have been tested for such strong interaction energies as the ligand−metal bonds in the investigated coordination compounds. We confirm earlier work that GGA functionals offer an improvement over the LDA and are particularly well suited for weak interactions like hydrogen bonds. We do, however, not find a particular reason to prefer any of the GGA functionals over another. Functionals derived from the GEA in general perform worse for all of the weaker interactions and cannot be recommended. An unexpectedly good performance is found for the coordination compounds, in particular with the GEA-derived functionals. However, the presently available kinetic energy functionals cannot be applied in cases in which a density redistribution between the subsystems leads to strongly overlapping subsystem electron densities.

## 1. Introduction

The quantum chemical study of large molecular systems which are of importance in life sciences or nanotechnology requires the use of multilevel methods which can treat different parts of the total system using different approximations. The frozen-density-embedding (FDE) scheme[1,2] first developed by Wesolowski and Warshel is such a promising multilevel method. It has already been applied in a number of studies, for example, of solvent effects on absorption spectra,[3−5] electron spin resonance parameters,[6] and nuclear magnetic resonance chemical shifts.[7]

FDE is based on a subsystem formulation[8] of density functional theory (DFT),[9] in which a large system is assembled from an arbitrary number of subsystems which are coupled by an effective embedding potential. In this way, the ground-state electron density is obtained as a superposition of subsystem electron densities, and the ground-state energy is obtained as a sum of subsystem energies and an interaction energy. FDE can be regarded as an approximation to Kohn−Sham (KS) DFT in which part of the kinetic energy of the noninteracting reference system is calculated via an explicit density functional. Making the theory formally exact therefore requires knowledge of the exact functional ($T_s$) for this energy in addition to the functional for the exchange-

* Corresponding author e-mail: agoetz@sdsc.edu (A.W.G.); visscher@chem.vu.nl (L.V.).

[†] VU University Amsterdam.

[‡] Current address: San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive MC0505, La Jolla, California, 92037.

10.1021/ct9001784 CCC: $40.75 © 2009 American Chemical Society
Published on Web 11/02/2009

correlation (XC) energy that is already used in KS-DFT. In practice, one employs approximate functionals for both energy terms.

Although it is only a small fraction of the total kinetic energy which has to be approximated in FDE, the limited accuracy of available functionals puts severe restrictions on the possible partitioning of a system and limits the accuracy of numerical results which can be obtained with standard FDE calculations. In general, the error with respect to KS-DFT results becomes larger with increasing strength of the interactions between the subsystems. For example, the inclusion of covalent bonds between subsystems is at present only possible with an extension of the simple FDE scheme[10] in which capping groups are introduced to improve the description.

Most published applications of FDE employ either the Thomas−Fermi (TF) kinetic energy functional[11,12] or the generalized gradient approximation (GGA) functional PW91k,[13,14] which have been shown to deliver good accuracy for subsystems with weak interactions and hydrogen bonds.[14−21] Studies regarding the accuracy of kinetic energy functionals for FDE interaction energies, however, have either focused on one type of weak interactions for a small set of systems,[14,22] been limited to a restricted number of approximate kinetic energy functionals,[20,21] or both.[14−16,19,23,24]

The aim of this contribution is to bridge this gap and provide a systematic analysis of the performance of existing kinetic energy functionals in the context of FDE. To this end, we have extended a flexible and efficient implementation[25] of FDE in the Amsterdam Density Functional program package (ADF)[26−28] such that interaction energies can be computed. We have applied this implementation to a diverse range of data sets of molecular complexes covering not only weakly interacting and hydrogen-bonded systems but also transition metal complexes with coordination bonds.

In the next section, we sketch the aspects of the FDE formalism which are of importance for further discussion. Section 3 continues with a discussion of the form of the kinetic energy functionals which we have investigated in this study. In section 4, we present the data sets which we have used for assessing the performance of the kinetic energy functionals. Section 5 contains the computational details of our study. In section 6, we discuss the results which we have obtained with our implementation. We first analyze the convergence behavior of the interaction energies with respect to the optimization of the electron densities in the individual subsystems. Then, we assess the accuracy of FDE in combination with different kinetic energy functionals with respect to KS-DFT. Section 7 summarizes our findings.

## 2. Frozen-Density Embedding

The starting point for FDE[1,2,25] is a subsystem formulation of DFT,[8] in which the total system is partitioned into $N$ subsystems such that the total electron density $\rho^{\mathrm{tot}}(r)$ is represented as the sum of the subsystem electron densities $\rho^{(i)}(r)$:

$$\rho^{\mathrm{tot}}(r) = \sum_{i}^{N} \rho^{(\mathrm{i})}(r) \qquad (1)$$

with each subsystem containing a fixed integer number of electrons. The total electronic DFT energy for such a partitioning is most conveniently written as

$$E[\rho^{(1)}, ..., \rho^{(N)}] = \sum_{i}^{N} E^{(i)}[\rho^{(i)}] + E_{\mathrm{int}}[\rho^{(1)}, ..., \rho^{(N)}] \qquad (2)$$

where $E^{(i)}[\rho^{(i)}]$ is the standard KS-DFT total energy of subsystem $i$:

$$E^{(i)}[\rho^{(i)}] = T_{\mathrm{s}}[\rho^{(i)}] + \int v_{\mathrm{nuc}}^{(i)}(r)\, \rho^{(i)}(r)\, \mathrm{d}r + \frac{1}{2}\int \frac{\rho^{(i)}(r)\, \rho^{(i)}(r')}{|r - r'|}\, \mathrm{d}r\, \mathrm{d}r' + E_{\mathrm{xc}}[\rho^{(i)}] \qquad (3)$$

Here, $v_{\mathrm{nuc}}^{(i)}(r)$ is the electrostatic potential of the nuclei in subsystem $i$ and $E_{\mathrm{xc}}[\rho^{(i)}]$ is the XC energy of subsystem $i$. The interaction energy between the subsystems is then given as

$$E_{\mathrm{int}}[\rho^{(1)}, ..., \rho^{(N)}] = \sum_{i \neq j}^{N} \int v_{\mathrm{nuc}}^{(i)}(r)\, \rho^{(j)}(r)\, \mathrm{d}r + \sum_{i<j}^{N} \int \frac{\rho^{(i)}(r)\, \rho^{(j)}(r')}{|r - r'|}\, \mathrm{d}r\, \mathrm{d}r' + T_{\mathrm{S}}^{\mathrm{nad}}[\rho^{(1)}, ..., \rho^{(N)}] + E_{\mathrm{xc}}^{\mathrm{nad}}[\rho^{(1)}, ..., \rho^{(N)}] \qquad (4)$$

where the nonadditive kinetic energy $T_{\mathrm{s}}^{\mathrm{nad}}[\rho^{(1)}, ..., \rho^{(N)}]$ and the nonadditive XC energy $E_{\mathrm{xc}}^{\mathrm{nad}}[\rho^{(1)}, ..., \rho^{(N)}]$ are defined as

$$T_{\mathrm{s}}^{\mathrm{nad}}[\rho^{(1)}, ..., \rho^{(N)}] = T_{\mathrm{s}}[\rho^{\mathrm{tot}}] - \sum_{i}^{N} T_{\mathrm{s}}[\rho^{(i)}] \qquad (5)$$

and

$$E_{\mathrm{xc}}^{\mathrm{nad}}[\rho^{(1)}, ..., \rho^{(N)}] = E_{\mathrm{xc}}[\rho^{\mathrm{tot}}] - \sum_{i}^{N} E_{\mathrm{xc}}[\rho^{(i)}] \qquad (6)$$

In these expressions, $T_{\mathrm{s}}[\rho]$ is the kinetic energy of the reference system of noninteracting electrons, which can be calculated exactly if the KS orbitals are known for all densities. Within FDE, however, one would like to determine KS orbitals only for the individual subsystems. This goal can be realized by employing an approximate kinetic energy functional to evaluate $T_{\mathrm{s}}^{\mathrm{nad}}[\rho^{(1)}, ..., \rho^{(N)}]$.

Minimization of the total energy functional of eq 2 with respect to the electron density $\rho^{(i)}$ of a subsystem $i$ while keeping the electron density of all other subsystems frozen (fixed) leads to a set of coupled KS-like equations:

$$\left[-\frac{1}{2}\nabla^2 + v_{\mathrm{eff,KS}}^{(i)}[\rho^{(i)}](r) + v_{\mathrm{emb}}^{(i)}[\rho^{(1)}, ..., \rho^{(N)}](r)\right]\psi_k^{(i)}(r) = \varepsilon_k^{(i)}\psi_k^{(i)}(r) \qquad (7)$$

from which the KS orbitals $\{\psi_k^{(i)}\}$ and the associated electron density $\rho^{(i)}$ of subsystem $i$ can be determined. The KS effective potential $v_{\mathrm{eff,\,KS}}^{(i)}[\rho^{(i)}]$ contains the usual terms of the electrostatic potential of the nuclei, the Hartree potential, and the XC potential of subsystem $i$. The effective embedding potential $\nu^{(i)}_{\mathrm{emb}}[\rho^{(1)}, ..., \rho^{(N)}]$ describes the effect of all other subsystems on subsystem $i$ and reads

$$\nu^{(i)}_{\mathrm{emb}}[\rho^{(1)}, ..., \rho^{(N)}](r) = \sum_{j,j\neq i}^{N} \int \nu^{(j)}_{\mathrm{nuc}}(r) + \sum_{j,j\neq i}^{N} \int \frac{\rho^{(j)}(r')}{|r - r'|}\mathrm{d}r' + \left.\frac{\delta E_{\mathrm{xc}}[\rho]}{\delta \rho}\right|_{\rho=\rho^{\mathrm{tot}}(r)} - \left.\frac{\delta E_{\mathrm{xc}}[\rho]}{\delta \rho}\right|_{\rho=\rho^{(i)}(r)} + \left.\frac{\delta T_{\mathrm{s}}[\rho]}{\delta \rho}\right|_{\rho=\rho^{\mathrm{tot}}(r)} - \left.\frac{\delta T_{\mathrm{s}}[\rho]}{\delta \rho}\right|_{\rho=\rho^{(i)}(r)} \qquad (8)$$

In addition to the electrostatic potential of the nuclei and electrons of the environment, it contains contributions from the nonadditive XC functional and the nonadditive kinetic energy functional.

In order to be able to yield the exact ground-state electron density $\rho^{\mathrm{tot}}$, the subsystem electron density $\rho^{(i)} = \rho^{\mathrm{tot}} - \sum_{j,j\neq i}^{N}\rho^{(j)}$ has to be positive and noninteracting $\nu$ -representable. In practice, it will be difficult to initialize the frozen subsystem densities such that the positivity condition for the active density is fulfilled. In such cases, the subsystem densities have to be determined in an iterative fashion.[8] This can be achieved by so-called "freeze-and-thaw" cycles,[29] in which the electron density of one active subsystem is updated and then frozen again while the electron density of the next subsystem is determined. This procedure can be repeated in a self-consistent fashion until the electron densities of all subsystems are converged.

If the exact density functional for the KS kinetic energy $T_{\mathrm{s}}[\rho]$ were known, the freeze-and-thaw procedure of FDE would represent an alternative way to determine the density of the total KS system. This means that convergence of this procedure should, irrespective of the employed XC functional, lead to the same ground state electron density and electronic ground state energy as with conventional KS-DFT (for the chosen XC functional). In an exact theory, it would thereby be possible to partition the density in different ways over the subsystems, the different choices resulting in differences in the individual subsystem energies that are compensated by differences in the interaction energy to yield a consistent total energy. [One could also think of a flexible setup in which the number of electrons on each subsystem is allowed to vary dynamically.] With the approximate kinetic energy functionals used in practical calculations, not all partitionings are meaningful, however. As a rule of thumb, one should avoid situations in which the subsystem densities overlap strongly and yield a large value of $T_{\mathrm{s}}^{\mathrm{nad}}$.[30] While this is easy to avoid for the starting densities, errors in the embedding potential may lead to an overestimation of charge transfer effects in a freeze-and-thaw optimization, bringing the partitioning outside the realm of applicability of the approximate functional.

## 3. Kinetic Energy Functionals

Several approximate types of kinetic energy functionals are available nowadays, and a comprehensive review is available in reference 31. An overview on the use of such functionals in FDE can be found, for example, in references 2 and 22. We therefore will restrict ourselves to a short overview on the functionals which we have investigated in this work. It should therefore be noted that among others the TF functional was originally constructed to describe the total kinetic energy $T$ rather than the KS kinetic energy $T_{\mathrm{s}}$ of the noninteracting reference system that is applicable when the correlation-kinetic energy $T_{\mathrm{c}}$ is implicitly included in the XC energy $E_{\mathrm{xc}}$.[31]

We would, furthermore, like to point out that there does not seem to exist a correlation between the accuracy of approximate kinetic energy functionals for the total KS kinetic energy $T_{\mathrm{s}}$ and kinetic energy differences such as the nonadditive kinetic energy $T_{\mathrm{s}}^{\mathrm{nad}}$.[2,30−32] For instance, some of the functionals investigated in this work fail to give binding of some simple molecules if used in orbital-free DFT with the correct KS density (for a given XC functional) as input.[33] However, this does not mean that these functionals will perform equally as bad if used to approximate only $T_{\mathrm{s}}^{\mathrm{nad}}$.

We have considered kinetic energy functionals which, for the general case of spin-polarized systems with spin density $\rho_\sigma$ ($\sigma = \alpha, \beta$), can be written in the following form:

$$T_{\mathrm{s}}^{\mathrm{approx}}[\rho] = 2^{2/3}C_{\mathrm{F}} \sum_{\sigma=\alpha,\beta} \int \rho_\sigma^{5/3}(r)\, F_{\mathrm{T}}(s_\sigma(r))\, \mathrm{d}r \qquad (9)$$

Here, $C_{\mathrm{F}} = (3/10)(3\pi^2)^{2/3} \approx 2.871$ is the TF constant. $F_{\mathrm{T}}(s_\sigma)$ denotes the enhancement factor, which is a function of the reduced density gradient

$$s_\sigma(r) = \frac{|\nabla \rho_\sigma(r)|}{2k_{\mathrm{F},\sigma}(r)\, \rho_\sigma(r)} \qquad (10)$$

with the Fermi vector

$$k_{\mathrm{F},\sigma}(r) = [6\pi^2 \rho_\sigma(r)]^{1/3} \qquad (11)$$

The analytic form of $F_{\mathrm{T}}(s_\sigma)$ determines the gradient dependence of the approximate kinetic energy functional.

**3.1. Local Density Approximation (LDA).** Just as for the total kinetic energy, the dominant part of the nonadditive kinetic energy $T_{\mathrm{s}}^{\mathrm{nad}}$ can be derived from the LDA, that is, TF theory.[14,22] The TF enhancement factor is given as

$$F_{\mathrm{TF}}(s_\sigma) = 1 \qquad (12)$$

**3.2. Gradient Expansion Approximation (GEA).** For slowly varying electron densities, the kinetic energy can be represented by a gradient expansion[34,35] in which TF theory represents the zeroth-order term. Truncation to second order yields the TF functional with von Weizsäcker correction, for which the enhancement factor takes the form

$$F_{\mathrm{TF9W}}(s_\sigma) = 1 + \frac{5}{27}s_\sigma^2 \qquad (13)$$

*E00.* A kinetic energy functional which represents the GEA up to fourth order and for which the enhancement factor takes the form

$$F_{\mathrm{E00}}(s_\sigma) = \frac{135 + 28s_\sigma^2 + 5s_\sigma^4}{135 + 3s_\sigma^2} \qquad (14)$$

has been proposed by Ernzerhof.[32]

*P92.* Perdew proposed a functional which reproduces the GEA up to sixth order.[36] The enhancement factor

$$F_{\mathrm{P92}}(s_\sigma) = \frac{1 + 88.3960 s_\sigma^2 + 16.3683 s_\sigma^4}{1 + 88.2108 s_\sigma^2} \qquad (15)$$

has the same functional form as the E00 enhancement factor and differs only in its parameters.

*OL1 and OL2.* On the basis of scaling properties of $T_s[\rho]$, Ou-Yang and Levy[37] proposed two kinetic energy functionals which replace fourth- and higher-order GEA terms with approximate, simple expressions. The corresponding enhancement factors take the following forms:

$$F_{\mathrm{OL1}}(s_\sigma) = 1 + \frac{5}{27}s_\sigma^2 + \frac{C_4^{(1)}}{C_\mathrm{F}} b s_\sigma \qquad (16)$$

and

$$F_{\mathrm{OL2}}(s_\sigma) = 1 + \frac{5}{27}s_\sigma^2 + \frac{C_4^{(2)}}{C_\mathrm{F}} \frac{b s_\sigma}{(1 + 4 b s_\sigma)} \qquad (17)$$

where $b = 2(3\pi^2)^{1/3}$, $C_4^{(1)} = 6.77 \times 10^{-3}$ and $C_4^{(2)} = 8.87 \times 10^{-2}$.

**3.3. Generalized Gradient Approximation (GGA).** The GGA seems at present to be the most promising and successful route to approximate kinetic energy functionals. According to a conjecture of Lee, Lee, and Parr (LLP), the kinetic energy and the exchange energy can be considered "conjoint" such that kinetic energy functionals may be constructed using the same analytical function for the enhancement factor as used in GGA exchange energy functionals.[38] It should be kept in mind that this conjointness conjecture is not strictly correct.[33]

*LLP91.* LLP suggested the use of the analytical form of the Becke (B88) exchange functional,[39] but reparametrized for the kinetic energy[38]

$$F_{\mathrm{LLP91}}(s_\sigma) = 1 + \alpha \frac{(2^{1/3} b s_\sigma)^2}{1 + \gamma (2^{1/3} b s_\sigma) \sinh^{-1}(2^{1/3} b s_\sigma)} \qquad (18)$$

with $\alpha = 4.4188 \times 10^{-3}$ and $\gamma = 2.53 \times 10^{-2}$.

*PW86.* The Perdew–Wang (PW86) exchange functional[40] with enhancement factor

$$F_{\mathrm{PW86}}(s_\sigma) = (1 + 1.296 s_\sigma^2 + 14 s_\sigma^4 + 0.2 s_\sigma^6)^{1/15} \qquad (19)$$

has been tested for the kinetic energy by Lacks and Gordon.[41]

*PW91k.* The functional which has been applied most widely in applications of FDE has been dubbed PW91k. It uses the analytic form of the enhancement factor of the Perdew–Wang (PW91) exchange functional[42] with parameters optimized for the kinetic energy by Lembarki and Chermette (LC94).[13] The application of this kinetic energy functional in the context of FDE was investigated by Wesolowski and co-workers,[14,22] who analyzed both the original PW91 as well as the LC94 enhancement factor and found that both improved upon earlier functionals due to the better behavior of these functions at large values of $s_\sigma$.

$$F_{\mathrm{PW91k}}(s_\sigma) = 1 + \frac{[A_2 - A_3 \exp(-A_4 s_\sigma^2)] s_\sigma^2 - B_1 s_\sigma^4}{1 + A_1 s_\sigma \sinh^{-1}(A s_\sigma) + B_1 s_\sigma^4} \qquad (20)$$

with $A_1 = 0.093907$, $A_2 = 0.26608$, $A_3 = 0.0809615$, $A_4 = 100.0$, $A = 76.320$, and $B_1 = 0.57767 \times 10^{-4}$.

*TW02.* The analytic form of the enhancement factor suggested by Becke[43] and used in the Perdew–Burke–Ernzerhof (PBE) functional[44,45] has been reparametrized by Tran and Wesolowski to reproduce the kinetic energy of He and Xe atoms.[46] It is given as

$$F_{\mathrm{TW02}}(s_\sigma) = 1 + \frac{\mu s_\sigma^2}{1 + (\mu/\kappa) s_\sigma^2} \qquad (21)$$

with $\kappa = 0.8438$ and $\mu = 0.2319$.

*T92.* The kinetic energy functional of Thakkar[47]

$$F_{\mathrm{T92}}(s_\sigma) = 1 + \frac{\beta (2^{1/3} b s_\sigma)^2}{1 + \gamma (2^{1/3} b s_\sigma) \sinh^{-1}(2^{1/3} b s_\sigma)} - \frac{C_4 (2^{1/3} b s_\sigma)}{1 + 4 b s_\sigma} \qquad (22)$$

with $\beta = 0.0055$ and $C_4 = 0.072$ uses a linear combination of the enhancement factor of the B88 exchange and OL2 kinetic energy functionals and has been fitted to the kinetic energy of 77 molecules.

*PBE2, PBE3, and PBE4.* A different route was taken by Karasiev et al.[33] They reparametrized kinetic energy functionals to better reproduce forces, that is, the shape of the KS potential energy surface instead of focusing on the accuracy of total energies. Although it cannot be expected that these functionals perform well for absolute energies, one might hope to obtain good relative energies (like the bond energies studied in this work). To investigate the feasibility of this ansatz, we have included these functionals in our study. The enhancement factors considered are of the two-parameter PBE form[43−45] and the three- and four-parameter series expansions of the PBE form,[48] called PBE*n* ($n$ = 2, 3, 4) in the following. They can be written as

$$F_{\mathrm{PBE}n}(s_\sigma) = 1 + \sum_{i=1}^{n-1} C_i^{(n)} \left[ \frac{s_\sigma^2}{1 + a^{(n)} s_\sigma^2} \right]^i \qquad (23)$$

where $a^{(1)} = 0.2942$, $C_1^{(1)} = 2.0309$, $a^{(2)} = 4.1355$, $C_1^{(2)} = -3.7425$, $C_2^{(2)} = 50.258$, $a^{(3)} = 1.7107$, $C_1^{(3)} = -7.2333$, $C_2^{(3)} = 61.645$, and $C_3^{(3)} = -93.683$. Although the parameters have been determined for a very limited set of molecules, some transferability of the functionals was found when applied to CO, a molecule which was not included in the training set.[33]

## 4. Data Sets

Truhlar and co-workers have recently suggested different sets of molecular complexes for testing XC functionals in KS-DFT. Some of these test sets have already been employed by Dulak and Wesolowski to assess the accuracy of FDE interaction energies for two combinations of XC and kinetic energy functionals in FDE.[21] They have confirmed older studies by Wesolowski and co-workers[15,16,19,21,23,24] and shown that FDE can outperform KS-DFT for interaction energies if an appropriate combination of functionals for the XC energy and the nonadditive kinetic energy is chosen. It is clear, however, that these encouraging findings are the result of an error cancellation between the applied XC and kinetic energy functionals. The accuracy of FDE for a given combination of functionals with respect to highly accurate interaction energies derived from wave function-based methods is system-dependent. Comparison of the combinations (i) LDA for both XC and nonadditive kinetic energy and (ii) PW91[42] as a XC functional and PW91k for the nonadditive kinetic energy has shown[21] that, for hydrogen-, dipolar- and weakly bonded dimers combination (i) performs well, while combination (ii) performs worse or even qualitatively wrong. For $\pi$-stacking complexes and complexes with charge-transfer character, however, the situation is reversed, and combination (ii) yields results which are in better agreement with the reference data.

These results[21] are important for practical purposes since they show for which type of systems we can expect maximum error cancellation with the combination of functionals under investigation. They do not, however, directly assess the performance of a given kinetic energy functional for FDE. With the exact kinetic energy functional, FDE should reproduce the KS-DFT results and not perform better. We therefore assess the accuracy of kinetic energy functionals for FDE by comparison to the corresponding KS-DFT results obtained with the same XC functional as in the FDE calculations.

We use Truhlar's NC31/05 data set of noncovalently bound molecular complexes with equilibrium geometries obtained from benchmark wave-function-based calculations.[49,50] The strength of interactions in this data set covers the range up to 16.15 kcal/mol at the benchmark level of theory. We use the fixed geometries of the benchmark equilibrium structures of the monomers and dimers from this data set to analyze the accuracy of FDE interaction energies obtained with different kinetic energy functionals in comparison to the corresponding KS-DFT results. While KS-DFT with the presently available GGA XC functionals does not adequately describe interactions between the molecular fragments in this data set, which are dominated by dispersion forces, a comparison of FDE to supermolecular KS-DFT still allows us to assess the accuracy of the kinetic energy functionals. In addition, if the KS-DFT results are recovered by FDE, existing empirical corrections for dispersion forces like the very successful DFT-D correction[51] can be added. The complexes in the NC31/05 data set are further subdivided into five groups:[49,50]

- HB6/04 (hydrogen bond). $(NH_3)_2$, $(HF)_2$, $(H_2O)_2$, $NH_3\cdots H_2O$, $(HCONH_2)_2$, and $(HCOOH)_2$ (benchmark interaction energies ranging from $-3.15$ to $-16.15$ kcal/mol)
- CT7/04 (charge transfer). $C_2H_4\cdots F_2$, $NH_3\cdots F_2$, $C_2H_2\cdots ClF$, $HCN\cdots ClF$, $NH_3\cdots Cl_2$, $H_2O\cdots ClF$, and $NH_3\cdots ClF$ (benchmark interaction energies ranging from $-1.06$ to $-10.62$ kcal/mol)
- DI6/04 (dipole interaction). $(H_2S)_2$, $(HCl)_2$, $HCl\cdots H_2S$, $CH_3Cl\cdots HCl$, $HCN\cdots CH_3SH$, and $CH_3SH\cdots HCl$ (benchmark interaction energies ranging from $-1.66$ to $-4.16$ kcal/mol)
- WI7/05 (weak interaction). HeNe, HeAr, $(Ne)_2$, NeAr, $CH_4\cdots Ne$, $C_6H_6\cdots Ne$, and $(CH_4)_2$ (benchmark interaction energies ranging from $-0.04$ to $-0.51$ kcal/mol)
- PPS5/05 ($\pi-\pi$ stacking). $(C_2H_2)_2$, $(C_2H_4)_2$, sandwich (S) $(C_6H_6)_2$, T-shaped (T) $(C_6H_6)_2$, and parallel-displaced (PD) $(C_6H_6)_2$ (benchmark interaction energies ranging from $-1.34$ to $-2.78$ kcal/mol)

In addition, we use the BP8/05 data set[52] of stacking interactions in six nucleic acid base complexes and hydrogen-bonding interactions in two Watson−Crick (WC) type base pairs.

- BP8/05 (nucleobase pairs). Adenine−thymine (A$\cdots$T), guanine−cytosine (G$\cdots$C), antiparallel cytosine dimer (anti C$\cdots$C), displaced cytosine dimer (displ C$\cdots$C), parallel cytosine dimer (par C$\cdots$C), uracil dimer (U$\cdots$U), WC adenine−thymine (WC A$\cdots$T), and WC guanine−cytosine (WC G$\cdots$C) (benchmark interaction energies ranging from $-2.45$ to $-28.80$ kcal/mol)

In order to check the performance of the kinetic energy functionals for stronger interactions than contained in the databases listed above, we added data sets containing coordination compounds to the list of systems to be tested. For this purpose, we employ a data set consisting of five complexes of $Zn^{2+}$ with simple inorganic ligands, which we will refer to as Zn5/08. The structures have been obtained from a data set which has recently been compiled by Amin and Truhlar for the purpose of testing the performance of density functionals.[53] As did those authors, we studied the dissociation of the complexes into the $Zn^{2+}$ ion and the ligand and, in the case of $Zn(OH)_2$, into $Zn(OH)^+$ and $OH^-$.

- Zn5/08 ($Zn^{2+}$ coordination complexes). $Zn(NH_3)^{2+}$, $Zn(H_2O)^{2+}$, $ZnOH^+$, $Zn(OH)_2$, and $Zn(SCH_3)^+$ (benchmark interaction energies ranging from $-96.83$ to $-428.18$ kcal/mol)

Finally, we have put together a data set of transition metal coordination complexes with ligands of the spectrochemical series. In order to prevent problems with high-spin/low-spin splittings in transition metal complexes, which are in general difficult to treat accurately with DFT in any case, we focus on octahedral Cr(III) complexes, which in general have a quadruplet ground state due to the three unpaired electrons in the metal $t_{2g}$ orbitals. The structures have been obtained as described in section 5, and we studied the interaction energies between the ligand $L^{n-}$ and the corresponding $[Cr(OH_2)_5]^{3+}$ fragment, that is, the ligand bonding energies. The accuracy of these energies is indicative, for example, of the applicability to the study of ligand exchange reactions.

- Cr10/09. $[Cr(OH_2)_5L]^{(3-n)+}$, where $n$ is the charge of the ligand $L^{n-}$ and the ligands are (in order of increasing ligand field splitting parameter) $I^-$, $Br^-$, $S^{2-}$, $Cl^-$, $F^-$, $OH^-$, $H_2O$, $NH_3$, $NO_2^-$, and CO (interaction energies ranging from $-43.83$ to $-782.78$ kcal/mol at the PBE/TZ2P level of theory)

## 5. Computational Details

All calculations were performed with a development version of the ADF program[26-28] using the gradient-corrected PBE[44,45] XC functional. We used the TZ2P basis set of the ADF basis set library, which is a triple-$\zeta$ valence/double-$\zeta$ core all-electron Slater basis augmented with two sets of polarization functions. The ADF default settings for the numerical integration grid and the self-consistent field (SCF) procedure were used. Geometries of the Cr(III) complexes were considered as converged if the maximum element of the gradient was below $10^{-4}$ au/Å.

For the FDE calculations, we used the approximate kinetic energy functionals described in section 3. We tested both a monomolecular expansion basis, denoted as FDE(m), in which only basis functions located on the active subsystem are used for the expansion of its KS orbitals, and a supermolecular (global) expansion basis, denoted as FDE(s), in which the basis functions of all subsystems are used for the expansion of the KS orbitals of the active subsystem.[54] In FDE(m) calculations, the numerical integration grid was centered on the active subsystem with only a few grid points added to deal with the singularities of the Coulomb potential at point nuclei of the frozen subsystem in close proximity to the active subsystem.[4]

Unless otherwise noted, all reported FDE energies have been converged to $10^{-5}$ au in the iterative solution of the FDE eqs (eq 7). Both KS-DFT and FDE(s) energies were corrected for the basis set superposition error (BSSE) with the counterpoise technique.[55]

The setup and execution of all calculations and the subsequent data extraction and analysis were automatized with the help of PyADF,[56] which is a scripting framework for quantum chemistry realized in the Python[57] programming language. PyADF was also employed to handle the freeze-and-thaw iterations in the FDE calculations of the Cr(III) complexes since these can at present not be handled within ADF for open-shell systems.

**5.1. Implementation Notes.** The FDE energy evaluation according to eq 2 relies on the implementation of total KS-DFT energies in ADF. The FDE interaction energies [see eq 4] were obtained from a new implementation into ADF which builds on an earlier implementation by Wesolowski and Dulak[58] for the nonadditive kinetic energy $T_s^{\mathrm{nad}}$ and the nonadditive XC energy $E_{\mathrm{xc}}^{\mathrm{nad}}$. In the current implementation, the $T_s^{\mathrm{nad}}$ and $E_{\mathrm{xc}}^{\mathrm{nad}}$ contributions to the interaction energy and the embedding potential are obtained from exact (as opposed to fitted) densities to increase the numerical accuracy. New is the calculation of the remaining terms, out of which the electrostatic interaction energy between the electron densities (Coulomb energy) of the subsystems merits some comments. We use the shorthand notation

$$J[\rho^{(i)}, \rho^{(j)}] = \int \frac{\rho^{(i)}(r)\,\rho^{(j)}(r')}{|r - r'|}\mathrm{d}r\,\mathrm{d}r' \qquad (24)$$

With Slater-type basis functions, these integrals cannot be calculated analytically. Therefore, in ADF, the electron densities are expanded into an auxiliary basis set of Slater functions.[27,59] The electrostatic potential of these fitted electron densities

$$\tilde{\rho}^{(i)}(r) = \rho^{(i)}(r) - \delta\rho^{(i)}(r) \qquad (25)$$

can be evaluated on the numerical integration grid so that we can compute the approximate Coulomb energy by numerical quadrature as

$$\tilde{J}[\rho^{(i)}, \rho^{(j)}] = J[\rho^{(i)}, \tilde{\rho}^{(j)}] + J[\delta\rho^{(j)}, \tilde{\rho}^{(i)}] = J[\rho^{(j)}, \tilde{\rho}^{(i)}] + J[\delta\rho^{(i)}, \tilde{\rho}^{(j)}] \qquad (26)$$

$\tilde{J}[\rho^{(i)}, \rho^{(j)}]$ has an error of $\mathcal{O}(\delta\rho^{(i)}\delta\rho^{(j)})$.[60] An integration grid in the region of subsystems $i$ and $j$ is required in order to evaluate the integrals of eq 26. In FDE(m) calculations, however, efficiency reasons make it advantageous to center the numerical integration grid on the active subsystem. We therefore proceed as follows. Let us assume that we commence the $n$th freeze-and-thaw cycle and we optimize electron density $\rho^{(i)}(n)$ in the presence of all other subsystem electron densities $\rho^{(j)}(n - 1)$ of the previous freeze-and-thaw iteration. We now compute $J[\rho^{(i)}(n), \tilde{\rho}^{(j)}(n - 1)]$ and $J[\delta\rho^{(i)}(n), \tilde{\rho}^{(j)}(n - 1)]$ using the grid of subsystem $i$ and take $J[\rho^{(j)}(n - 1), \tilde{\rho}^{(i)}(n - 1)]$ and $J[\delta\rho^{(j)}(n -1), \tilde{\rho}^{(i)}(n - 1)]$ from the $(n - 1)$th iteration (computed on a grid of subsystem $j$) to obtain an approximate Coulomb energy:

$$\tilde{J}[\rho^{(i)}, \rho^{(j)}](n) = \frac{1}{2}\{\tilde{J}'[\rho^{(i)}, \rho^{(j)}](n) + \tilde{J}''[\rho^{(i)}, \rho^{(j)}](n)\} \qquad (27)$$

where

$$\tilde{J}'[\rho^{(i)}, \rho^{(j)}](n) = J[\rho^{(i)}(n), \tilde{\rho}^{(j)}(n - 1)] + J[\delta\rho^{(j)}(n - 1), \tilde{\rho}^{(i)}(n - 1)] \qquad (28)$$

and

$$\tilde{J}''[\rho^{(i)}, \rho^{(j)}](n) = J[\rho^{(j)}(n - 1), \tilde{\rho}^{(i)}(n - 1)] + J[\delta\rho^{(i)}(n), \tilde{\rho}^{(j)}(n - 1)] \qquad (29)$$

Upon convergence of the freeze-and-thaw cycles, eqs 28 and 29 become equivalent, and the difference

$$\Delta\tilde{J}[\rho^{(i)}, \rho^{(j)}](n) = |\tilde{J}'[\rho^{(i)}, \rho^{(j)}](n) - \tilde{J}''[\rho^{(i)}, \rho^{(j)}](n)| \qquad (30)$$

has to vanish.

## 6. Results and Discussion

**6.1. Convergence Behavior of Energy Contributions.** In Table 1, we show typical examples of the convergence behavior of the FDE energy (eq 2) during the iterative solution of the FDE eqs (eq 7). The error $\Delta\tilde{J}$ in the Coulomb energy [see eq 30] drops quickly to $10^{-6}$ au after the third to fourth freeze-and-thaw iteration, and at the same time, the energy converges to $10^{-5}$ au. It takes one additional iteration to converge FDE(s) calculations to the same accuracy as FDE(m) calculations.

We have observed a similar convergence behavior of FDE(m) calculations for all of the molecular complexes investigated, independent of the applied kinetic energy functional. Exceptions are the PBE$n$ functionals which caused SCF convergence problems in FDE(m) calculations for some of the transition metal complexes.

***Table 1.*** Convergence Behavior of FDE Energy Terms (in au without BSSE correction) with the Number, *n*, of Freeze-and-Thaw Iterations for Two Representative Examples[a]

| | FDE(m) | | | | FDE(s) | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | $\tilde{J}(n)$ | $\Delta\tilde{J}(n)$ | $E_{int}(n)$ | $E(n)$ | $\tilde{J}(n)$ | $\Delta\tilde{J}(n)$ | $E_{int}(n)$ | $E(n)$ |
| | | | | $NH_3\cdots H_2O$ | | | | |
| 1 | 17.639472 | 0.006771 | −0.009632 | −132.904592 | 17.646163 | 0.011736 | −0.009104 | −132.903440 |
| 2 | 17.633830 | 0.000169 | −0.012977 | −132.907889 | 17.643158 | 0.000663 | −0.014824 | −132.908967 |
| 3 | 17.633686 | 0.000003 | −0.013064 | −132.907978 | 17.642992 | 0.000032 | −0.015151 | −132.909281 |
| 4 | 17.633680 | 0.000007 | −0.013066 | −132.907977 | 17.642982 | 0.000005 | −0.015169 | −132.909302 |
| 5 | | | | | 17.642982 | 0.000007 | −0.015171 | −132.909300 |
| | | | | $(HCOOH)_2$ | | | | |
| 1 | 93.989199 | 0.003520 | −0.039358 | −379.324115 | 94.034378 | 0.001900 | −0.048987 | −379.329414 |
| 2 | 93.985858 | 0.000203 | −0.042262 | −379.325771 | 94.034852 | 0.000137 | −0.052773 | −379.330506 |
| 3 | 93.985663 | 0.000015 | −0.042440 | −379.325864 | 94.035125 | 0.000029 | −0.053048 | −379.330448 |
| 4 | 93.985649 | 0.000001 | −0.042452 | −379.325870 | 94.035163 | 0.000004 | −0.053075 | −379.330436 |
| 5 | 93.985648 | 0.000000 | −0.042453 | −379.325869 | 94.035169 | 0.000002 | −0.053078 | −379.330429 |
| 6 | | | | | 94.035169 | 0.000000 | −0.053078 | −379.330430 |

[a] The PW91k functional was used for the nonadditive kinetic energy.

The convergence behavior of FDE(s) calculations, however, depends strongly on the type of intermolecular interactions and the kinetic energy functional employed. In general, FDE(s) calculations with LDA and GGA functionals (except PBE*n*) converge smoothly, with some exceptions. Around 25 freeze-and-thaw iterations were required for a converged FDE(s) energy of the charge-transfer complex $NH_3\cdots ClF$, and with the LLP91 GGA functional no convergence could be obtained at all. Almost none of the Cr(III) complexes could be converged with FDE(s), irrespectively of the applied functional. The $H_2O$ complex could be converged with the TF, PW91k, and TW02 functionals and the CO complex with the TF and PW91k functionals.

We encountered SCF convergence problems (either immediately or after some freeze-and-thaw iterations) with FDE(s) calculations employing kinetic energy functionals derived from the GEA (TF9W, E00, P92, OL1, OL2, T92) or the PBE*n* kinetic energy functionals for some of the complexes of the CT7/04, DI6/04, PPS/05, BP8/05, and Zn5/08 data sets and all complexes of the Cr10/09 data set (see also next section). This holds in particular for the E00, PBE2, and PBE4 functionals. Whenever good SCF convergence was achieved, however, the FDE(s) energy converged well within a few freeze-and-thaw iterations also for these functionals.

In the cases in which no convergence could be achieved, the functional derivative of the kinetic energy functional which enters the embedding potential [see eq 8] obviously leads to a qualitatively wrong potential. This does not necessarily lead to a problem in FDE(m) calculations. The availability of the full basis set for FDE(s), however, allows the electron density to probe regions where the embedding potential is too attractive.[61] As a consequence, the electron density may redistribute, and the initially chosen subsystem partitioning can get lost such that the system enters the strong overlap regime and the approximate functional in use no longer is able to describe the situation.

***Table 2.*** Mulliken Charges Associated with the Fluorine and Chromium Atoms in the Two Subsystems of $[Cr(OH_2)_5F]^{2+}$ during the Course of Freeze-and-Thaw Iterations[a]

| | $[Cr(OH_2)_5]^{3+}$ subsystem | | $F^-$ subsystem | |
|---|---|---|---|---|
| freeze-and-thaw cycle | Cr | F | Cr | F |
| 0 | 1.5 | | | −1.0 |
| 1 | 1.8 | 0.0 | −0.2 | −0.8 |
| 2 | 2.0 | 0.0 | −0.3 | −0.7 |
| 3 | 2.0 | 0.0 | −0.5 | −0.5 |
| 4 | 2.1 | 0.0 | −1.2 | 0.3 |
| 5 | 2.2 | 0.0 | −1.8 | 0.9 |

[a] The TF functional was used as nonadditive kinetic energy functional.

Let us illustrate this with the example of an FDE(s) calculation with the TF functional for the $[Cr(OH_2)_5F]^{2+}$ complex, which is partitioned into $[Cr(OH_2)_5]^{3+}$ and $F^-$. Table 2 shows the Mulliken charges[62] associated with the chromium and fluorine atoms in the two subsystems during the course of freeze-and-thaw iterations. We start out with a chromium atom which carries a charge of 1.5. The remaining charge of 1.5 is buffered by the water ligands. The fluoride atom has of course a charge of −1. What we then observe is a gradual transfer of density from the fluoride to the metal center during the course of the freeze-and-thaw iterations. Some charge transfer, accompanied by some back-bonding from the metal to the ligand, is to be expected during the formation of the coordination bond. However, in this case, too much density moves from the ligand to the metal center. As a consequence, after a few freeze-and-thaw iterations, a subsystem partitioning with strongly overlapping subsystem electron densities is reached, and results become meaningless.

**6.2. Accuracy of Interaction Energies.** We will discuss the accuracy of FDE interaction energies obtained with the different kinetic energy functionals for each data set, which represents a particular type of interaction, separately. In general, however, we found distinct performance for the LDA, for the group of functionals which are derived from the GEA (TF9W, E00, P92, OL1, OL2), standard GGA functionals (LLP91, PW86, PW91k, TW02), and the PBE*n* functionals. As could be expected from its functional form, the accuracy of the T92 functional is usually in between the accuracy of the GEA-derived functionals and that of the GGA functionals.

Figures 1 and 2 summarize the accuracy of the different kinetic energy functionals in FDE(m) and FDE(s) calcula-

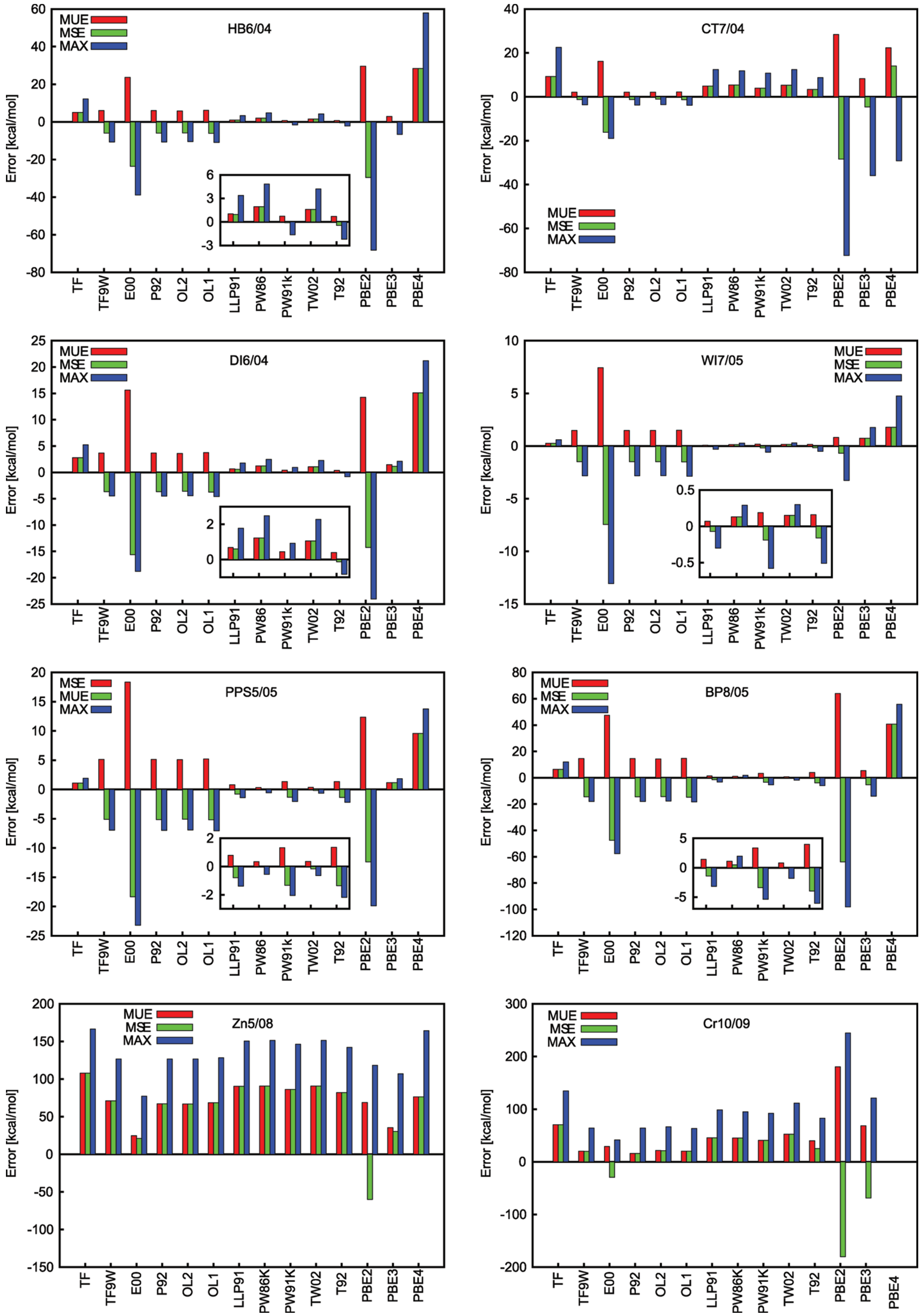


*Figure 1.* Performance of kinetic energy functionals for FDE(m) interaction energies for all investigated data sets. Mean unsigned error (MUE), mean signed error (MSE), and maximum error (MAX) for BSSE-corrected interaction energies.

tions, respectively, for all data sets investigated. No data are presented for the Cr10/09 data set with FDE(s) since essentially none of the calculations could be converged (see also section 6.1). Tables containing all interaction energies for supermolecular KS-DFT, FDE(m), and FDE(s) can be found in the Supporting Information.

Figures 3 and 4 show a comparison of KS and FDE(m) interaction energies and KS and FDE(s) interaction energies,

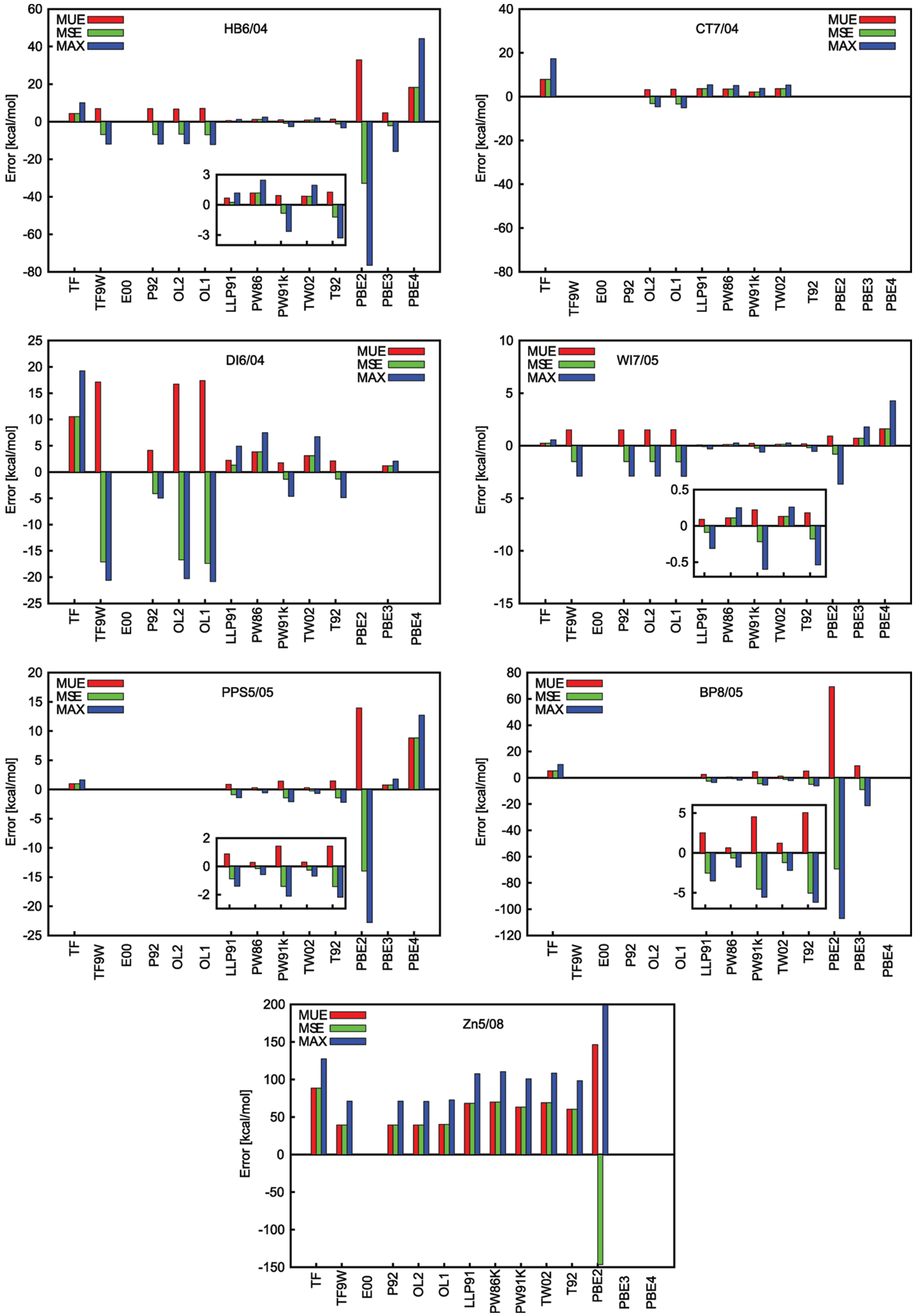


*Figure 2.* Performance of kinetic energy functionals for FDE(s) interaction energies for all investigated data sets. Mean unsigned error (MUE), mean signed error (MSE), and maximum error (MAX) of BSSE-corrected interaction energies.

respectively, for the three kinetic energy functionals which perform best for a given data set.

When discussing the performance of the kinetic energy functionals for the FDE interaction energies, one has to keep in mind that it is not only the form of the approximate functional for the kinetic energy which determines the quality of the results. It is the functional derivative of the kinetic energy functional which enters the expression for the embedding potential [see eq 8] and as such determines the quality of the electron density obtained. It may well be

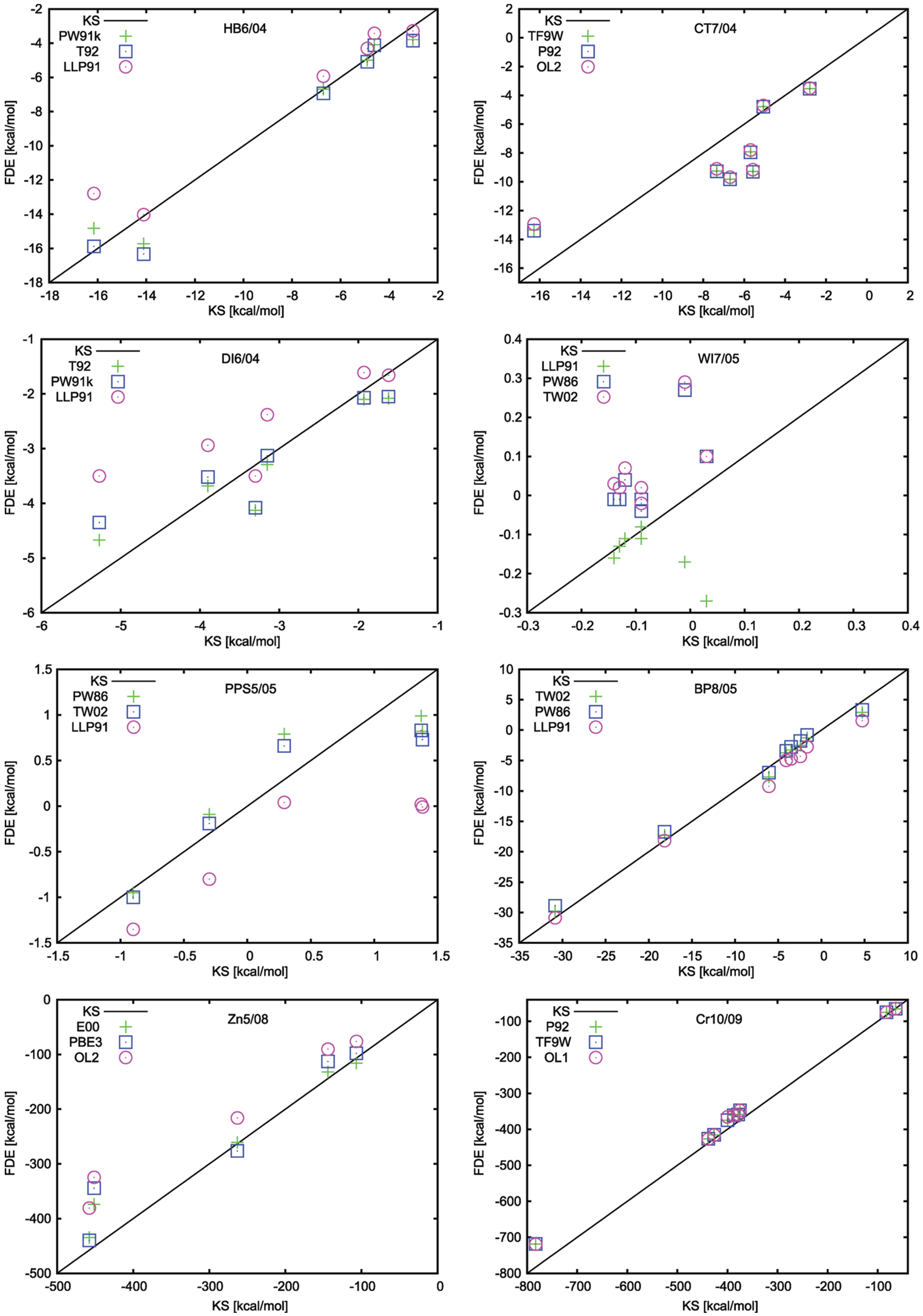


***Figure 3.*** Comparison of KS interaction energies with those of FDE(m) for the kinetic energy functionals which have the smallest mean unsigned error (MUE) for a given data set.

the case that error cancellations happen in the sense that a good interaction energy is obtained, although the underlying electron density obtained from the preceding FDE calculation is not good at all.

From the point of view of computational efficiency, FDE(m) is clearly superior to FDE(s), the latter even having a higher computational cost than a corresponding supermolecular KS-DFT calculation. However, while

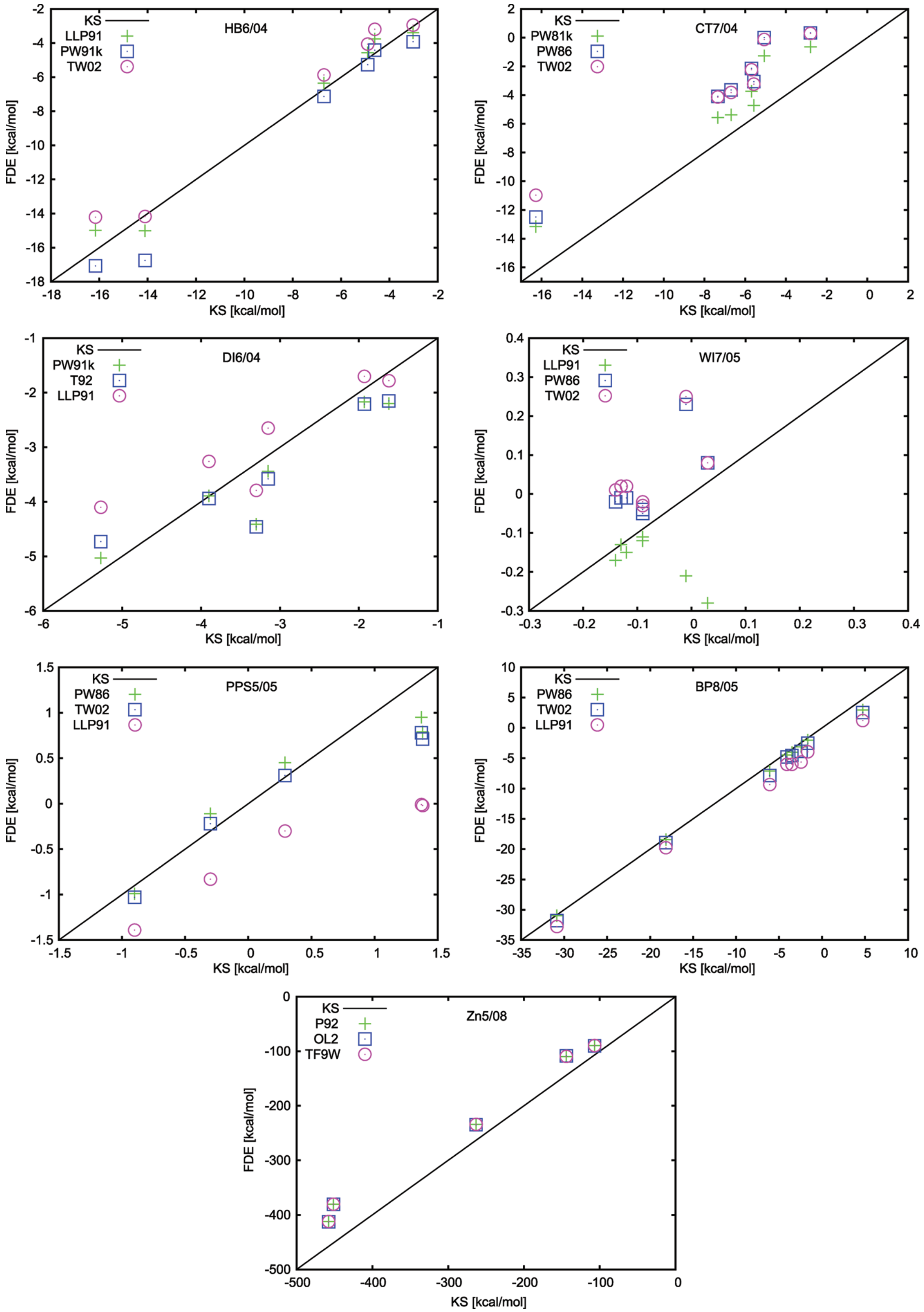


***Figure 4.*** Comparison of KS interaction energies with those of FDE(s) for the kinetic energy functionals which have the smallest mean unsigned error (MUE) for a given data set.

FDE(m) allows for polarization of the subsystem electron densities, no charge transfer between the subsystems can be described with FDE(m). We would like to stress that only FDE(s) can exactly recover the results of a supermolecular KS-DFT calculation (with the unknown, exact kinetic energy functional for the nonadditive kinetic energy $T_s^{nad}$). Thus, the proper reference for judging the accuracy of kinetic energy functionals for the nonadditive kinetic energy $T_s^{nad}$ within the FDE formalism must be the results from FDE(s) calculations. However, in general, whenever there is no net charge transfer between the subsystems, FDE(m) is an excellent approximation to FDE(s). This holds in particular if a sufficiently good basis set is employed which allows for a proper polarization of the subsystem electron densities. This certainly holds for the large TZ2P basis set employed in our studies.

Very good results are obtained for the interaction energies of the hydrogen-bonded dimers in the HB6/04 data set with the standard GGA functionals and the T92 functional, both with FDE(m) and FDE(s). The PW91k functional performs best for FDE(m) calculations, while the LLP91 functional takes the lead for FDE(s) calculations. Overall, however, the difference between these functionals is not big, and there is no clear reason to prefer one over another. The LDA performs much worse. It consistently underestimates the magnitude of the binding energies and cannot be recommended for energetics in hydrogen-bonded systems. The GEA-derived functionals offer no improvement over the LDA and consistently overshoot the binding energy. The PBE*n* functionals perform particularly badly, with the exception of the PBE3 functional which, however, shows an unacceptably large maximum error in the case of FDE(s).

For the CT7/04 data set, the LDA performs very badly and does not predict any binding both for FDE(m) and for FDE(s). The GEA-derived functionals apart from E00 perform rather well for FDE(m); however, this must be due to an error cancellation because FDE(m) calculations are not able to model a real charge transfer. For FDE(s) calculations, in which such charge transfer is possible, similar performance is found, except for the TF9W functional for which the SCF could not be converged for some dimers. This is due to a too strong charge transfer due to errors in the embedding potential as discussed above. Acceptable results are also obtained with the standard GGA functionals for FDE(s), PW91k taking the lead. The PBE*n* functionals perform very badly, again.

For the DI6/04 data set, very good results are obtained again with the standard GGA functionals, in particular, PW91k, but T92 is as good. Also, the PBE3 functional is able to deliver good results. Given the relatively small interaction energies, the LDA and GEA-derived functionals do not perform well, the errors being as large as the bonding energies. All functionals perform worse with FDE(s), indicating that some errors in the embedding potential are not probed by the smaller FDE(m) basis.

The intermolecular interactions in the WI7/05 data set are very weak, and most of the functionals perform similarly, both for FDE(m) and for FDE(s) calculations. Again, the standard GGA functionals and T92 give the best agreement with supermolecular KS-DFT calculations. Results obtained with the GEA-derived functionals and the PBE*n* functionals in general have an error which is larger than the interaction energies themselves.

The interaction energies of the PPS5/05 data set are also rather weak, and it should be mentioned that KS-DFT does not predict any binding of the benzene dimers. The picture is similar to that for the other data sets, the standard GGA functionals performing best. Out of these, PW86 and TW02 are clearly superior, and PW91k has the largest errors. Also, the LDA and the PBE3 functional give interaction energies relatively close to the KS-DFT results. The GEA-derived functionals and PBE2 and PBE4, however, yield useless results or, in the case of FDE(s), do not even converge.

For the BP8/05 data set, a similar picture as before arises. The GGA functionals and T92 perform rather well. Among these, PW91k shows the largest errors in the interaction energies. PW86 and TW02 are the most reliable functionals for these types of interactions. The LDA and the PBE3 functional still yield somewhat reasonable results. All other functionals either do not converge or show errors which are too large to make them useful for any practical interaction energy calculations.

At first sight, rather poor results are obtained for the Zn5/08 data set with very large maximum and mean errors. Given the strong interactions between the subsystems, however, the relative errors are much smaller than one might have expected. Due to the charge on the $Zn^{2+}$ ion and the dipole moment or charge on the ligands, the interaction energy is dominated by the electrostatic contribution [nuclear−nuclear repulsion and first two terms of eq 4]. It is important to point out that the nonadditive kinetic energy contribution to the interaction energy is on the same order of magnitude as the electrostatic contribution. For instance, in an FDE(m) calculation on $[Zn(OH)]^+$ with the E00 functional, the electrostatic contribution to the interaction energy of eq 4 between $Zn^{2+}$ and $OH^-$ is −623.73 kcal/mol, while the contribution due to the nonadditive kinetic and XC energy is +145.50 and −44.67 kcal/mol, respectively. This emphasizes the importance of a proper treatment of the nonadditive kinetic energy. Except for PBE2, all functionals underestimate the binding energies. For FDE(m), E00 performs best, and PBE3 is able to deliver better results than any other GGA functional. For FDE(s), however, one hits the problem described in section 6.1, and the SCF does not converge for any of the systems both with E00 and PBE3 as well as PBE4. It is particularly disappointing that GGA functionals do not bring a significant improvement over the LDA. As to be expected, for the cases in which the SCF converges, the FDE(s) results are in better agreement with the supermolecular KS-DFT calculations. Here, all GEA-derived functionals apart from E00 prove to be superior to other functionals. Unlike the other functionals, PBE2 overshoots the interaction energy for FDE(s) and exhibits particularly large errors. Nevertheless, it is encouraging to see (Figures 3 and 4) that the relative strength of the interaction energies is correctly reproduced by the best FDE calculations and, as mentioned above, the relative errors are rather small.

As already discussed in section 6.1, almost none of the complexes from the Cr10/09 data set could be converged with FDE(s). In the cases in which convergence was achieved, however, FDE(s) leads to an improvement over FDE(m) (see Supporting Information) with the exception of the CO complex for which the FDE(m) interaction energy with the PW91k functional is in better agreement with the supermolecular KS-DFT result. The FDE(m) results for the Cr10/09 data set are surprisingly good for most of the functionals, with the exception of the PBE*n* series, which, again, performs rather poorly. In particular with PBE4, SCF convergence could be achieved only for the fluoride complex. All functionals apart from E00 and PBE*n* underestimate the binding energies. The best results are obtained with the GEA-derived functionals, out of which E00 is the worst. But also the GGA functionals apart from PBE*n* still yield satisfactory interaction energies. At this point, it is important to note that, as for the $Zn^{2+}$ coordination compounds, the electrostatic and nonadditive kinetic energy contributions to the FDE interaction energy are on the same order of magnitude and one order of magnitude larger than the nonadditive XC energy contribution. Thus, also here, the good correlation between KS and FDE interaction energies (Figure 3) is not only due to electrostatic interactions between the charged $[Cr(H_2O)_5]^{3+}$ complex and ligand fragments. The nonadditive kinetic-energy functional plays a very important role.

## 7. Conclusions

We have implemented the calculation of interaction energies in the framework of FDE into the ADF program package. Using this implementation, we have performed a systematic analysis of the performance of a large set of LDA, GEA-derived, and GGA kinetic energy density functionals for the nonadditive kinetic energy contribution to FDE interaction energies. We have studied a representative data set with interaction energies ranging from −1 kcal/mol in weakly interacting dimers to −783 kcal/mol in coordination bonds of transition metal complexes.

We have shown that the freeze-and-thaw cycle for a self-consistent update of the subsystem electron densities and the accompanying FDE energy converges within a few iterations whenever the SCF of an FDE calculation converges smoothly. This is in general the case both for FDE(m) and for FDE(s) calculations with LDA and GGA kinetic-energy functionals, except for the PBE*n* functionals. FDE(m) calculations with the GEA-derived and PBE*n*−GGA functionals also converge properly. In FDE(s) calculations, however, convergence problems are encountered frequently with these functionals and, in the case of the Cr(III) complexes, for all functionals we have tested. Particularly bad in this respect are the E00, PBE2, and PBE4 functionals. Qualitatively wrong embedding potentials are obtained in these cases, which lead to strongly overlapping subsystem electron density partitionings which lie outside the domain of applicability of these approximate functionals.

Reasonable interaction energies can already be obtained with the LDA. However, GGA kinetic-energy functionals (apart from PBE*n*) in general significantly improve upon the LDA. Exceptions are the very weak interactions of the WI7/05 data set and the coordination bonds in the Zn5/08 data set for which the errors with the LDA and GGA functionals are practically of equal magnitude. In most cases, the LLP91, TW02, and PW91k functionals work best, but there is no strong indication to prefer one GGA over another if interaction energies are the target of interest. Also, the PBE3 functional yields very good interaction energies in many cases, but it is not reliable, in particular for FDE(s) calculations for which no SCF convergence can be achieved in some cases.

It is particularly encouraging that reasonable results can be obtained for the bond energies in coordination compounds. In contrast to the weak intermolecular interactions, for these systems, the GEA-derived functionals perform better than the GGA functionals. This indicates that FDE using these functionals may be useful for studies of ligand exchange reactions or catalytic reactions in which transition metal centers are involved.

We expect our results to be transferable to FDE calculations carried out in conjunction with XC functionals other than PBE which we have employed throughout this work.

**Acknowledgment.** This work has been supported by NWO in form of a VICI grant for L.V.

**Supporting Information Available:** Tables containing interaction energies for supermolecular KS-DFT, FDE(m), and FDE(s) calculations for all investigated data sets. This information is available free of charge via the Internet at http://pubs.acs.org/.

### References

(1) Wesolowski, T. A.; Warshel, A. *J. Phys. Chem.* **1993**, *97*, 8050–8053.

(2) Wesolowksi, T. A. One-Electron Equations for Embedded Electron Density: Challenge for Theory and Practical Payoffs in Multi-Level Modelling of Complex Polyatomic Systems. In *Computational Chemistry: Reviews of Current Trends*; Leszczynski, J., Ed.; World Scientific: River Edge, NJ, 2006; Vol. 10, Chapter 1, pp 1−82.

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

(4) Neugebauer, J.; Jacob, C. R.; Wesolowski, T. A.; Baerends, E. J. *J. Phys. Chem. A* **2005**, *109*, 7805–7814.

(5) Jacob, C. R.; Neugebauer, J.; Jensen, L.; Visscher, L. *Phys. Chem. Chem. Phys.* **2006**, *8*, 2349–2359.

(6) Neugebauer, J.; Louwerse, M. J.; Belanzoni, P.; Wesolowski, T. A.; Baerends, E. J. *J. Chem. Phys.* **2005**, *123*, 114101.

(7) Jacob, C. R.; Visscher, L. *J. Chem. Phys.* **2006**, *125*, 194104.

(8) Cortona, P. *Phys. Rev. B* **1991**, *44*, 8454–8458.

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

(10) Jacob, C. R.; Visscher, L. *J. Chem. Phys.* **2008**, *128*, 155102.

(11) Thomas, L. H. *Proc. Camb. Phil. Soc.* **1927**, *23*, 542.

(12) Fermi, E. *Rend. Accad. Lincei* **1927**, *6*, 602.

(13) Lembarki, A.; Chermette, H. *Phys. Rev. A* **1994**, *50*, 5328–5331.

(14) Wesolowski, T. A. *J. Chem. Phys.* **1997**, *106*, 8516–8526.

(15) Wesolowski, T. A.; Ellinger, Y.; Weber, J. *J. Chem. Phys.* **1998**, *108*, 6078–6083.

(16) Wesolowski, T. A.; Tran, F. *J. Chem. Phys.* **2003**, *118*, 20072–2080.

(17) Wesolowski, T. A. *J. Am. Chem. Soc.* **2004**, *126*, 11444–11445.

(18) Jacob, C. R.; Wesolowski, T. A.; Visscher, L. *J. Chem. Phys.* **2005**, *123*, 174104.

(19) Kevorkyants, R.; Dulak, M.; Wesolowski, T. A. *J. Chem. Phys.* **2007**, *124*, 024104.

(20) Dulak, M.; Kaminski, J. W.; Wesolowski, T. A. *J. Chem. Theory Comput.* **2007**, *3*, 735–745.

(21) Dulak, M.; Wesolowski, T. A. *J. Mol. Model.* **2007**, *13*, 631–642.

(22) Wesolowski, T. A.; Chermette, H.; Weber, J. *J. Chem. Phys.* **1996**, *105*, 9182–9190.

(23) Tran, F.; Weber, J.; Wesolowski, T. A. *Helv. Chim. Acta* **2001**, *84*, 1489–1503.

(24) Wesolowski, T. A.; Morgantini, P.-Y.; Weber, J. *J. Chem. Phys.* **2002**, *116*, 6411–6421.

(25) Jacob, C. R.; Neugebauer, J.; Visscher, L. *J. Comput. Chem.* **2007**, *29*, 1011–1018.

(26) *ADF2008.01*; SCM, Theoretical Chemistry, Vrije Universiteit: Amsterdam, The Netherlands. http://www.scm.com (accessed May 26th, 2009).

(27) Guerra, C. F.; Snijders, J.; te Velde, G.; Baerends, E. J. *Theor. Chem. Acc.* **1998**, *99*, 391–403.

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

(29) Wesolowski, T. A.; Weber, J. *Chem. Phys. Lett.* **1996**, *248*, 71–76.

(30) Bernard, Y. A.; Dulak, M.; Kaminski, J. W.; Wesolowski, T. A. *J. Phys. A: Math. Theor.* **2008**, *41*, 055302.

(31) Ludeña, E. V.; Karasiev, V. V. Kinetic energy functionals: History, challenges and prospects. In *Reviews of Modern Quantum Chemistry*; Sen, K. D., Ed.; World Scientific Publishing: Singapore, 2002; Vol. 1.

(32) Ernzerhof, M. *J. Mol. Struct.: THEOCHEM* **2000**, *501*, 59–64.

(33) Karasiev, V. V.; Trickey, S. B.; Harris, F. E. *J. Comput.-Aided Mater. Des.* **2006**, *13*, 111–129.

(34) Kirzhnits, D. A. *Sov. Phys. JETP* **1957**, *5*, 64.

(35) Hodges, C. H. *Can. J. Phys.* **1973**, *51*, 1428–1437.

(36) Perdew, J. P. *Phys. Lett. A* **1992**, *165*, 79–82.

(37) Ou-Yang, H.; Levy, M. *Int. J. Quantum Chem.* **1991**, *40*, 379–388.

(38) Lee, H.; Lee, C.; Parr, R. G. *Phys. Rev. A* **1991**, *44*, 768–771.

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

(40) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1986**, *33*, 8800–8802.

(41) Lacks, D. J.; Gordon, R. G. *J. Chem. Phys.* **1994**, *100*, 4446–4452.

(42) Perdew, J. P.; Chevary, J.; Vosko, S.; Jackson, K. A.; Pederson, M. R.; Singh, D.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*, 6671–6687.

(43) Becke, A. D. *J. Chem. Phys.* **1986**, *84*, 4524–4529.

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

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

(46) Tran, F.; Wesolowski, T. A. *Int. J. Quantum Chem.* **2002**, *89*, 441–446.

(47) Thakkar, A. J. *Phys. Rev. A* **1992**, *46*, 6920–6924.

(48) Adamo, C.; Barone, V. *J. Chem. Phys.* **2002**, *116*, 5933–5940.

(49) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 415–432.

(50) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656–5667.

(51) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.

(52) Zhao, Y.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2701–2705.

(53) Amin, E. A.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 75–85.

(54) Wesolowski, T. A.; Weber, J. *Int. J. Quantum Chem.* **1997**, *61*, 303–311.

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

(56) Jacob, C. R.; Götz, A. W.; Bulo, R. E.; Beyhan, S. M.; Visscher, L. *PyADF*; VU University Amsterdam and ETH Zurich, 2007−2009.

(57) Python. http://www.python.org (accessed May 26th, 2009).

(58) Dulak, M.; Wesolowski, T. A. *Int. J. Quantum Chem.* **2005**, *101*, 543–549.

(59) Baerends, E. J.; Ellis, D.; Roos, P. *Chem. Phys.* **1973**, *2*, 41–51.

(60) Dunlap, B. I.; Connoly, J. W. D.; Sabin, J. R. *J. Chem. Phys.* **1979**, *71*, 3396–3402.

(61) Jacob, C. R.; Beyhan, S. M.; Visscher, L. *J. Chem. Phys.* **2007**, *126*, 234116.

(62) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833–1840.

CT9001784

JCTC Journal of Chemical Theory and Computation

# Using Molecular Similarity to Develop Reliable Models of Chemical Reactions in Complex Environments

Volkan Ediz, Anthony C. Monda, Robert P. Brown, and David J. Yaron*

*Department of Chemistry, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, Pennsylvania 15213*

Received August 10, 2009

**Abstract:** The use of molecular similarity to develop reliable low-cost quantum mechanical models for use in quantum mechanical/molecular mechanical simulations of chemical reactions is explored, using the $H + HF \rightarrow H_2 + F$ collinear reaction as a test case. The approach first generates detailed quantum chemical data for the reaction center in geometries and electrostatic environments that span those expected to arise during the molecular dynamics simulations. For each geometry and environment, both high- and low-level ab initio calculations are performed. A model is then developed to predict the high-level results using only inputs generated from the low-level theory. The inputs used here are based on principal component analysis of the low-level distributed multipoles, and the model is a simple linear regression. The distributed multipoles are monopoles, dipoles, and quadrupoles at each atomic center, and they summarize the electronic distribution in a manner that is comparable across basis set. The error in the model is dominated by extrapolation from small to large basis sets, with extrapolation from uncorrelated to correlated methods contributing much less error. A single regression can be used to make predictions for a range of reaction-center geometries and environments. For the trial collinear reaction, separate regressions were developed for the transition region and the entrance and exit channels. These models can predict the results of CCSD(T)/cc-pVTZ computations from HF/3-21G distributed multipoles, with an average error for the reaction energy profile of 0.69 kcal/mol.

## 1. Introduction

Quantum chemistry has made great strides in developing highly accurate methods for computing the properties of small molecules, but application to large molecules remains challenging due to the rapid increase in computational cost with system size. One means of reducing cost is to restrict the full quantum description to a small locus, the reaction center, of a larger system. This approach is used in quantum mechanics/molecular mechanics (QM/MM) methods, where QM is used to describe a handful of atoms in the reaction center, while MM is used for the thousands of atoms in the remainder of the system (e.g., protein and solvent).[1−4] In such simulations, the QM algorithm is typically called millions of times to generate the energy, forces, and charge distribution of the reaction center in the presence of electrostatic interactions with the MM environment.[4] Despite the use of QM methods only in the reaction center, the QM computations are often the bottleneck in such simulations. It is this high cost of QM that this work seeks to substantially reduce, thus further expanding the reach of QM/MM models to the large and complex systems of relevance in biological and materials applications.[4−9] The computational savings of QM/MM stem from the local nature of chemistry, such that QM can be used on only a small locus of the system. Here, we explore the use of molecular similarity to further reduce the computational cost. We begin by performing high-level quantum computations on the reaction center in a range of environments that span those expected to arise in the QM/MM simulation. This data is then mined for a low cost model that can describe the reaction center in similar environments.

The QM portion of a QM/MM simulation is defined by the boundary with the MM region and by the level of electron

* Corresponding author. E-mail: yaron@cmu.edu.

10.1021/ct9004195 CCC: $40.75 © 2009 American Chemical Society
Published on Web 11/19/2009

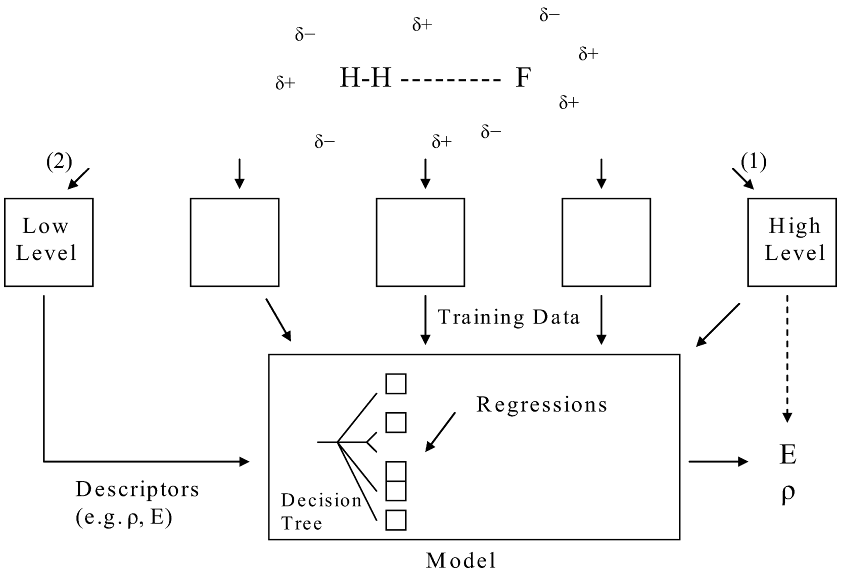


***Figure 1.*** Scheme for the development and the use of a model of the chemical reaction H + HF → $H_2$ + F. The middle boxes represent various levels of quantum chemical calculations with accuracies and costs that increase from left to right. The bottom box represents a model that maps from the energy (*E*) and the charge distribution (*ρ*) of a low-level model to that of the high-level model. The model may use different regressions, selected according to region along the reaction coordinate or other criteria.

structure theory used in the reaction center. A wide variety of quantum chemical methods (e.g., semiempirical and ab initio calculations) are now available with differing reliabilities and with CPU times ranging from seconds to days.[10,11] The level of computation needed to achieve a certain target accuracy varies widely, both with system type and with position along the reaction coordinate. A challenging aspect of chemical reactions is the need to obtain an accurate description of the energy and the configuration of the transition-state (TS), since the multireference character of the TS region often requires QM methods with the highest computational cost.[12] The formal scaling of computational effort for ab initio calculations on a N-electron system ranges from O ($N^3$) for Hartree−Fock theory to O ($N^5$) for MP2 and O ($e^N$) for the exact full-configuration interaction (full-CI) solution, making the most accurate and reliable methods difficult to apply in QM/MM simulations where the QM calculations must be called at each time step of a molecular dynamics (MD) trajectory.[13]

Many approaches have been developed to reduce the computational expense of ab initio calculations.[14−24] Generally, these attempts take advantage of two common features of molecular systems: nearsightedness and molecular similarity.

Nearsightedness relates to the local character of the interactions present in a molecule, such that interactions occurring on long length scales can be simplified to interactions between electrostatic multipoles and van der Waals forces.[25] Linear-scaling methods aim to take advantage of this local character wherever possible in a quantum chemical calculation.[26] For instance, divide-and-conquer methods reduce the computational cost of self-consistent field calculations[14,27] or correlated calculations,[28] while fast multipole methods accelerate the computation of Coulomb interactions.[16,29]

Molecular similarity relates to the tendency of molecular fragments, such as functional groups, to behave similarly in different molecules and environments. The assumption of molecular similarity underlies the use of atom- or functional-group-specific parameters in semiempirical quantum chemistry and molecular mechanics.[21,22,30] For instance, in semiempirical methods, the ab initio Hamiltonian is replaced with a simpler model Hamiltonian, which is parametrized either to experimental[21] or ab initio data.[31−34] Similarly, force fields in MM are parametrized using both experimental and theoretical data regarding the structure and the interactions between functional groups. Both methodologies lead to substantially lower computational costs than ab initio calculations, however, with a loss in accuracy that may limit their applicability.[4,35,36]

Here, we explore the use of molecular similarity to develop models for use in QM/MM calculations of chemical reactions that have substantially lower cost than ab initio calculations and that have controllable accuracy and reliability. Our approach first generates detailed quantum chemical data of the reaction center in configurations and electrostatic environments that span those expected to arise during the MD trajectory. For each configuration and environment, both high- and low-level ab initio calculations are performed. These data are then analyzed to develop a low-cost model that can, given only the output of a low-level ab initio calculation, predict the output of the high-level calculation. Development of the model also yields information on the reliability of the mapping from low- to high-level results, including both the expected error of the prediction and the range of configurations and environments over which the assumption of molecular similarity can be expected to hold. Such a model can then be used to perform QM/MM calculations at the cost of the low-level quantum theory, while generating results that approach the accuracy of the high-level theory.

Figure 1 shows a schematic representation of the approach applied to the collinear reaction of H + HF → $H_2$ + F, the trial reaction considered in this work. This reaction is sufficiently small that high-level computations can be done quickly, but sufficiently complex that it provides a realistic test of the approach. The transition-state has a substantial multireference character, and its position and energy are sensitive both to the level of ab initio calculation[37−40] and to the environment. Pathway 1 of Figure 1 shows the ideal, but computationally unfeasible, approach of using a high-level quantum chemical method to generate the information (energy of the system, $E$, and charge distribution of the reaction center, $\rho$) needed at each time step in an MD simulation. This work explores an alternative approach, pathway 2 of Figure 1, which generates this information from a model that takes as input information obtained from a low-level quantum chemical method. This model is trained on data generated for a set of molecular configurations that vary both the geometry of the reaction center and the electrostatic environment. Figure 1 shows that levels of computation between that of the low- and high-level theories may be useful either as inputs for pathway 2 or as additional information for the model development.

Section 2 describes the development of the model that maps low- to high-level quantum chemical results. The data used to extract the model is discussed in Section 2.1. One important aspect of the model development is determining a set of descriptors that can serve as useful inputs to a regression that maps low- to high-level results. The descriptors used here are based on a principal component analysis of Stone's distributed multipoles,[41] as discussed in Section 2.2. The general form of the regression model is introduced in Section 2.3 with the details of the model development being discussed in Section 3.1. Sections 3.2 and 3.3 present the results, and Section 4 gives a brief summary and future directions.

## 2. Methods

**2.1. Data Generation.** To test our model, we chose to study the collinear $H + HF \rightarrow H_2 + F$ reaction. This collinear reaction has only two degrees of freedom and is sufficiently small that calculations may be carried out quickly. Nevertheless, the reaction involves breaking a H−F bond and forming a H−H bond, which is sufficiently complex that accurate prediction of the reaction surface requires both large basis sets and high levels of electron correlation.[37−40] The sensitivity to basis set and correlation have been attributed to the multireference character of the wave function in the vicinity of the transition-state. The best agreement with experiment is obtained using either multireference−configuration interaction (MRCI) calculations with the Davidson correction[38] or coupled-cluster calculations.[37] The coupled-cluster calculations estimate the barrier height of the linear and the bent transition-states as 2.16 and 1.63 kcal/mol, respectively, with an uncertainty of 0.1 kcal/mol. In addition to this sensitivity to the quantum chemical approach, the results below show that the reaction surface is sensitive to the environment, and so this system provides a reasonable test of the ability of the approach to describe reactions in complex environments.

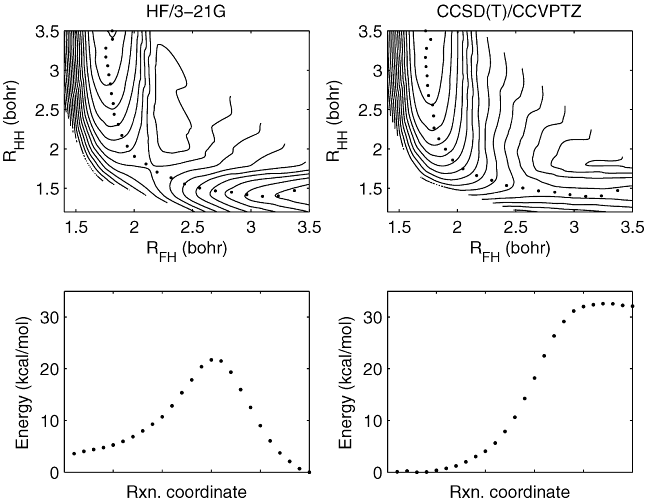


***Figure 2.*** HF/3-21G (top left) versus CCSD (T)/CCVPTZ (top right) surfaces of the gas-phase collinear $H + HF \rightarrow H_2 + F$ trial reaction. The contours are from 0 to 33 kcal/mol in steps of 3 kcal/mol. The dotted lines show reaction paths (top panels) and reaction energy profiles (lower panels).

***Table 1.*** Relative CPU Times for Some Typical QM Methods on $C_5H_{12}$[10,a]

| | 3-21G ($N = 69$) | 6-31G* ($N = 99$) | 6-31+G* ($N = 119$) | 6-311++G** ($N = 194$) |
|---|---|---|---|---|
| HF [$N^{2.7}$] | 1 | 3.8 | 5 | 23.1 |
| B3YLP [$N^{\sim 3}$] | 2.5 | 5 | 7 | 31 |
| MP2 [$O N^4$] | 1.4 | 7.6 | 10.2 | 60.8 |
| MP4 [$O^3V^4$] | 29.9 | 131.5 | 296.7 | 4066.2 |
| QCSID(T) [$O^3V^4$] | 63.3 | 220.9 | 558.3 | 8900.3 |

[a] Columns correspond to basis sets and rows to levels of electron correlation. Asymptotic scalings are in brackets ($N$ = total number of basis functions, O = number of occupied orbitals, V = number of unoccupied orbitals).

Figure 2 shows the substantially different reaction surfaces obtained from a low-level (HF/3-21G) versus a high-level (CCSD (T)/cc-pVTZ) computation. Our goal is to predict the high-level surface from outputs generated by the low-level method. We note that DFT methods may provide low-level models with higher accuracy and, for semilocal DFT, lower computational costs. HF theory is used here as the low-level model to allow a clear test of the ability of the map to include the correlation energy. Table 1 illustrates the computational savings possible from such an approach. This table lists relative CPU times as a function of both the size of the basis set and the level of electron correlation, along with the asymptotic scalings of the correlated methods. The times are listed for a somewhat larger system, $C_5H_{12}$, than that studied here to better illustrate the potential computational savings from use of this approach.

Data spanning the regions of interest were generated using an automated system that varies both the geometry of the reaction center and the electrostatic environment. The 46 geometries shown as symbols in Figure 3 were chosen to span the relevant region of the potential energy surface (PES) of the collinear reaction. The different symbols in Figure 3 indicate grouping of the geometries into three regions: entrance channel, transition-state, and exit channel. This classification of the PES into three regions is used to test

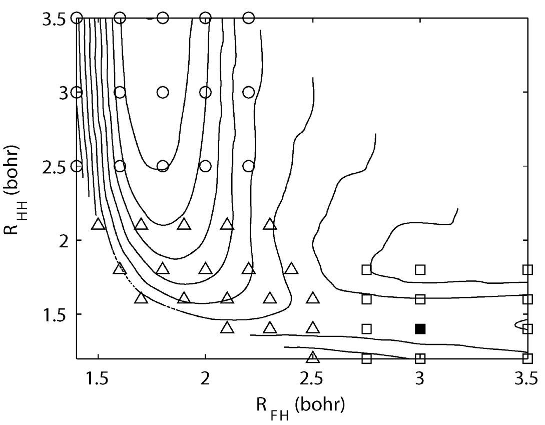


***Figure 3.*** Contour plot of PES of H + HF → $H_2$ + F reaction in a typical environment. Symbols are geometries where calculations were performed for the transition region (tri) and for the entrance (circ) and exit (square) channel regions. The filled square is the geometry[38] used in the analyses of Tables 2 and 3.

***Table 2.*** Mean Absolute Errors in kcal/mol for Predicting Correlated (QCISD) Self-Energies[a]

| | HF | MP2 | QCISD |
|---|---|---|---|
| STO-3G | 0.12 | 0.05 | 0.00 |
| 3-21G | 0.05 | 0.05 | 0.00 |
| 6-31G* | 0.01 | 0.00 | 0.00 |
| 6-31++G** | 0.02 | 0.01 | 0.00 |
| cc-pVTZ | 0.03 | 0.01 | 0.00 |

[a] From the output of lower-level computations via eq 2 with $N_{pca}$ = 10, for the geometry shown as a filled square in Figure 3.

***Table 3.*** Mean Absolute Error in kcal/mol for Predicting Large Basis (cc-pVTZ) Self-Energies[a]

| | STO-3G | 3-21G | 6-31G* | 6-31++G** | cc-pVTZ |
|---|---|---|---|---|---|
| HF | 0.52 | 0.18 | 0.21 | 0.10 | 0.00 |
| MP2 | 0.54 | 0.36 | 0.23 | 0.11 | 0.00 |
| QCISD | 0.57 | 0.53 | 0.24 | 0.12 | 0.00 |

[a] From the output of computations performed at the same level of correlation but with smaller basis sets. The model is that of eq 2 with $N_{pca}$ = 10. The reaction-center geometry is that of the filled square in Figure 3. For instance, a model mapping descriptors generated at the HF/6-31G* level to energies of a HF/cc-pVTZ has a RMS error of 0.21 kcal/mol.

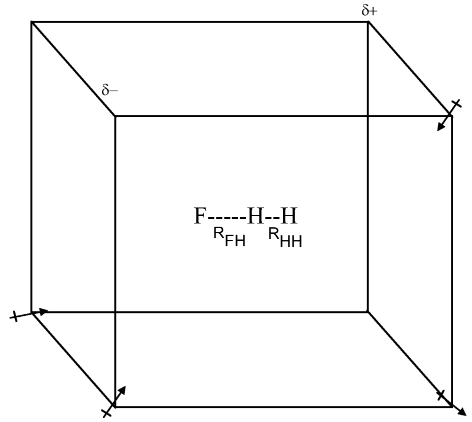


***Figure 4.*** Schematic of electrostatic environments used for the H + HF → $H_2$ + F collinear reaction. $R_{FH}$ and $R_{HH}$ are the bond distances between the F−H and H−H atoms, respectively.

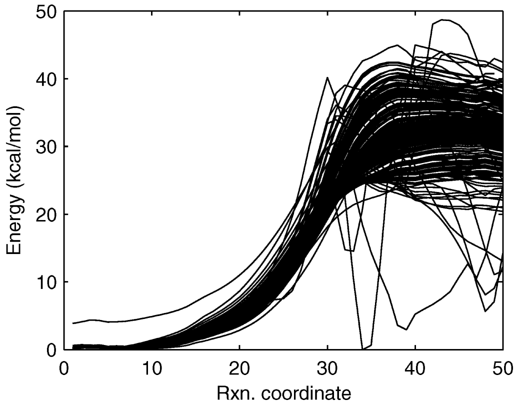


***Figure 5.*** Reaction energy profiles from CCSD(T)/CCVPTZ calculations for the 250 electrostatic environments included in this study.

the ability of a single model to make predictions over a fairly broad region of the configuration space (Section 3.3). For the point shown as a filled square in Figure 3, data were generated using the 15 levels of theory obtained from combining 3 correlation methods (HF, MP2, and QCISD) with 5 basis sets (STO-3G, 3-21G, 6-31G*, 6-31++G**, and cc-pVTZ). These data are used to explore various aspects of mapping low- to high-level theories. For the remainder of the geometries, data were generated using one low-level (HF/3-21G) and one high-level method (CCSD(T)/cc-pVTZ). The data consists of single-point energy calculations carried out using a tight convergence threshold of $1.5 \times 10^{-5}$ Hartree/Bohr with the GAUSSIAN03 software package.[42] The reaction surfaces are obtained from the 46 points of Figure 3 by triangle-based cubic interpolation. The reaction path and the energy profiles are then obtained from these interpolated surfaces.

For each of the reaction-center geometries of Figure 3, we generate a set of plausible electrostatic environments. In QM/MM calculations, the reaction center experiences the environment only through electrostatic interactions,[13] and so each environment consists of a set of external charges. For a biological reaction, MD trajectories would be run, and snapshots along this trajectory would be selected to yield a set of environments that span those likely to arise in the free energy computations. For our study, we generated a set of 250 random electrostatic environments by placing either a randomly oriented dipole (probability, $p = 0.8$), a single charge ($p = 0.1$), or void ($p = 0.1$) at each of the 8 corners of a 12 Bohr cube surrounding the reaction center, as shown in Figure 4. The magnitude for the single charges is selected randomly from a uniform distribution ranging from $-19.2 \times 10^{-19}$ to $+19.2 \times 10^{-19}$ Coulomb and for the dipoles from 0 to 4.72 D. These environments were chosen to induce substantial perturbations in the reaction surface. The barrier heights range from 0 to 24 kcal/mol with an average and standard deviation of 1.71 ± 2.92 kcal/mol at CCSD(T)/cc-pVTZ level (see Figure 5). The combination of 250 environments and 46 reaction-center geometries yields a data set with 11 500 entries, where each entry contains the results of both the low-level (HF/3-21G) and high-level (CCSD(T)/cc-pVTZ) computations. Of the 11 500 combinations, 12 failed to converge at the high-level theory within the time feasible given the computational resources, and so were not included in the data set.

Our goal is to create a model that can map from low- to high-level methods by predicting the output of the high-level method, using only inputs generated by the low-level method. Since mapping from low- to high-level methods requires mapping across basis sets, the approach used to generate descriptors must yield results that are comparable across basis sets. For this reason, we use Stone's distributed multipoles.[41] The distributed multipoles are multipoles (monopole, dipole, quadrupole, etc.) centered on each atom and, optionally, on any chosen additional site, such as bond centers. The distributed multipoles reproduce the electrostatic potential surrounding the molecule to chemical accuracy, and so provide an essentially complete representation of the electronic distribution of the molecule.[43] The advantage of this description is that the same number and type of distributed multipoles can be generated for any basis set, as opposed to an object such as the one-electron density matrix, whose size and meaning changes with basis set and cannot be easily compared across basis sets.[44] An additional benefit is that distributed multipoles integrate well in QM/MM since the distributed multipoles are sufficient to compute the electrostatic interaction between the reaction center and the environment.[45,46]

Distributed multipoles are calculated using the distributed multipole Analysis of Gaussian98 wave functions (GDMA) software package.[47] Distributed multipoles up to the sixth order are generated first at just the atom centers, for a total of 108 distributed multipoles, and then at both atom and bond centers, for a total of 180 distributed multipoles.

The mapping from low- to high-level theories is meant to predict the properties of the reaction center in configurations and environments that are similar to those in the original data set. The quantity to be predicted by the model is, therefore, the self-energy of the reaction, which is extracted from the total energy, $E_{\mathrm{TOT}}$, produced by the ab initio calculation as follows:

$$E = E_{\mathrm{TOT}} - E_{\mathrm{CHARGES}} - E_{\mathrm{INT}} \qquad (1)$$

where $E$ is the self-energy of the reaction center, $E_{\mathrm{CHARGES}}$ is the self-energy of the environment (i.e., the interaction energy between the fixed charges within the environment), and $E_{\mathrm{INT}}$ is the interaction energy between the reaction center and the fixed charges of the environment. $E_{\mathrm{TOT}}$ and $E_{\mathrm{CHARGES}}$ are generated with GAUSSIAN03 by default. $E_{\mathrm{INT}}$ is obtained using the ORIENT[48] software package to predict the energy of interaction between the distributed multipoles of the reaction center and the fixed charges of the environment.

### 2.2. Feature Extraction.

Feature extraction methods are useful to discover a minimal set of variables that may be used to describe the electronic structure of the reaction center. In ab initio calculations, a large set of variables is needed to obtain a form for the electronic wave function that is sufficiently flexible that accurate solutions of the Schroedinger equation may be obtained. For mapping low- to high-level quantum chemical results, the variables need only capture the variation in the electronic structure across situations that lie within the target range of validity for the model. The number of variables needed to describe differences among similar molecular structures is likely to be much smaller than the number of variables needed to obtain accurate solutions of the ab initio Hamiltonian.[49,50] Here, feature extraction is used to discover a reduced set of variables that describe variations in the electronic structure across the data set of Section 2.1.

The feature extraction method used here is principal component analysis (PCA), which is a simple and widely used approach.[51] PCA produces an orthogonal linear transformation of the feature space in such a way that the first linear combination of the original features, the first principal component, explains the greatest variance in data, and the second principal component is orthogonal to the first one and explains the greatest remaining variance and so on. Thus, each principal component identifies and ranks the most important features needed to capture the variability in the data. Principal components extracted in this manner define a new feature space that contains the same information as the original feature space but along dimensions that are ranked according to importance. Typically, only a few principal components are sufficient to capture the variability in data.

Here, PCA is applied to the distributed multipoles obtained from the high-level method. This yields principal components that are linear combinations of distributed multipoles. Since the units of the distributed multipoles vary with order (dipole, quadrupole, etc), some scaling approach is needed to make the various orders of the distributed multipoles comparable. A common approach in PCA analysis is the standardization of data by dividing each distributed multipole by the standard deviation of that particular distributed multipole in the input data. This gives each distributed multipole unit variance and so gives equal weight to all distributed multipoles, even those whose variance in the input data is quite small. Here, we instead weight the distributed multipoles according to their interaction with the fixed charges of the environment. This is done by first computing the average interaction energy between the DM with unit magnitude (e.g., the $x$ component of the dipole on the F atom) and the fixed charges of the environment. The distributed multipoles are then divided by this average interaction energy. Figure 6 shows the result of PCA on the distributed multipoles generated at atomic centers from 11 488 calculations (the 46 reaction-center geometries of Figure 3 in the 250 environments of Figure 4). Figure 6 indicates that the number of degrees of freedom needed to capture the variation in the electronic structure of the reaction center is about five for both low- and high-level QM computations. The distributed multipoles from the low-level computations are then projected onto the high-level PCA vectors to give scores, $S_i^{\mathrm{LL}}$, where $i$ labels the principal components in order of importance.

### 2.3. Model Fitting.

Above, we considered the choice of descriptors to be used as input to a model that maps from low- to high-level QM algorithms. The form of the model used here is a simple linear regression:

$$E^{\mathrm{HL}} = p^{\mathrm{const}} + p^{\mathrm{ener}} E^{\mathrm{LL}} + \sum_{i=1}^{N_{\mathrm{pca}}^{\mathrm{lin}}} p_i^{\mathrm{lin}} S_i^{\mathrm{LL}} + \sum_{i=1}^{N_{\mathrm{pca}}^{\mathrm{quad}}} p_i^{\mathrm{quad}} (S_i^{\mathrm{LL}})^2 \tag{2}$$

where $p$ is the model parameter, $E$ is the self-energy of the reaction center (the energy with electrostatic interactions between the reaction center and the environment removed via eq 1), and $S_i^{\mathrm{LL}}$ is the projection of the low-level distributed multipoles onto the $i^{th}$ principle component. $N_{\mathrm{pca}}^{\mathrm{lin}}$ and $N_{\mathrm{pca}}^{\mathrm{quad}}$ are the number of principal component descriptors included in the model for the linear and quadratic terms, respectively. When $N_{\mathrm{pca}}^{\mathrm{lin}}$ and $N_{\mathrm{pca}}^{\mathrm{quad}}$ are equal, they are quoted below as simply $N_{\mathrm{pca}}$.

All results presented below use five-fold cross validation, such that the model is trained on a randomly selected subset of 80% of the data and tested on the remaining 20%. The data is divided randomly into five equally sized subsets, and predictions for each subset are obtained from a model trained to the other four subsets.

## 3. Results

**3.1. Form of the Canonical Model.** The model involves choices regarding the number and type of distributed multipoles, the form of the regression in eq 2 (linear versus quadratic terms and inclusion of self-energy from the low-level model), and the number of PCA descriptors included in the regression. The canonical model used in the bulk of this paper includes distributed multipoles up to quadrupoles on each atom center, and includes the low-level energy in the regression along with both linear and quadratic terms for the 10 most important PCA vectors ($N_{\mathrm{pca}} = 10$ in eq 2). We next examine the sensitivity of the model predictions to these choices.

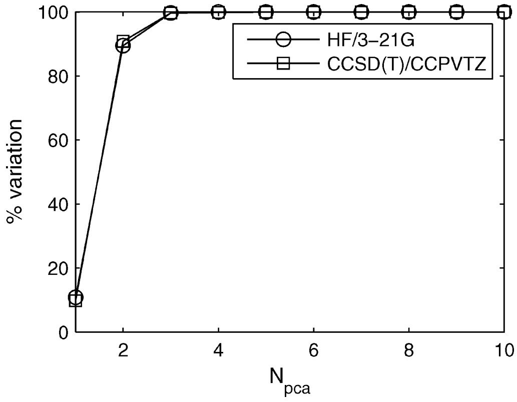


*Figure 6.* Percent variance of the electronic structure of the reaction center (i.e., the distributed multipoles) explained as a function of the number of principal components retained in the description.

We will initially examine some general aspects of the model fitting, holding the reaction center at the geometry shown as a filled triangle in Figure 3. This point is in the transition-state region of the isolated reaction center, where QM computations are expected to be most challenging.

The sensitivity of the model to the choice of distributed multipoles is shown in Figure 7. The average errors are to be compared to the standard deviation of 1.7 kcal/mol for the self-energy of the QCISD/cc-pVTZ calculations. Figure 7 shows the mean absolute error as a function of the number of principal components included in the regression, $N_{\mathrm{pca}}$ of eq 2, for various levels of multipoles and with (right) and without (left) inclusion of distributed multipoles at bond centers. The results in Figure 7 show that inclusion of distributed multipoles at bond centers does little to improve the performance of the model, and so the canonical model includes distributed multipoles only at atomic centers. Figure 7 also shows that the model performance increases significantly with addition of high-order distributed multipoles up to quadrupoles, but inclusion of higher ranks do not significantly decrease the error. Therefore, the canonical model includes only distributed multipoles up to quadrupoles on atomic centers, yielding a set of 27 raw descriptors. Figure 7 also shows that the error in the fit drops rapidly for the first seven principal components and then levels off. (This is a slower convergence than that seen in Figure 6, suggesting that an alternative approach to selecting input variables for the model could be beneficial.) Figure 7 suggests that $N_{\mathrm{pca}}$ should be greater than 7, but our final choice of $N_{\mathrm{pca}}$ for the canonical model will be based on fits to all reaction-center geometries discussed below.

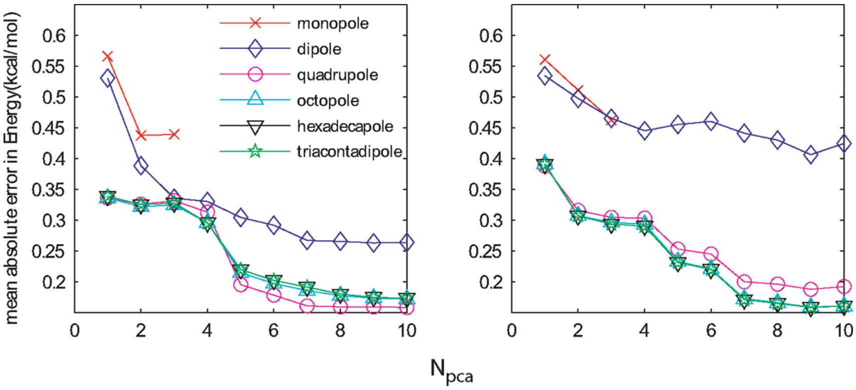


*Figure 7.* Mean absolute errors in kcal/mol from models predicting the QCISD/cc-pVTZ self-energy from output of HF/3-21G computations at the reaction center geometry, shown as a filled square in Figure 3. The lines show the average error versus the number of principal components included in both the linear quadratic terms of eq 2, when distributed multipoles up to the indicated level are included in the analysis. Distributed multipoles are included only on atoms (left) or on both atoms and bond centers (right).

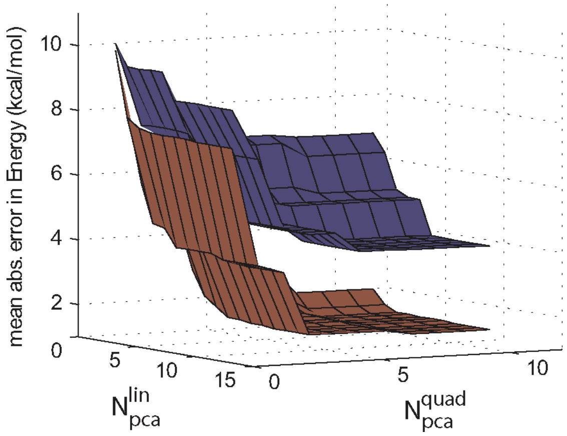


***Figure 8.*** Surface plot of mean absolute error as a function of the number of linear and quadratic terms included in eq 2 for a map from HF/3-21G to CCSD (T)/cc-pVTZ theory, using a single regression for all 46 geometries of Figure 3. Results are shown with (red) and without (blue) including the HF/3-21G self-energy in the model of eq 2.

Next, we examine the model performance when multiple reaction-center geometries are included in a single regression. Figure 8 shows results for the global fit as a function of $N_{pca}$ for the linear and quadratic terms of eq 2. Results are also shown both with and without inclusion of the self-energy from the low-level theory, $E_{LL}$, in the model of eq 2. Inclusion of $E_{LL}$ substantially improves the quality of the fit, since this allows the model to focus on corrections to the energy arising from use of larger basis sets and from inclusion of electron correlation, as opposed to fitting the energy itself. The error drops smoothly with the addition of parameters to the fit and the addition of linear and quadratic terms leads to roughly equivalent improvements. The canonical model used in the remainder of this paper includes up to the first 10 principal components in both the linear and quadratic terms, at which point the errors are 0.6 kcal/mol for the point-by-point fit and 1.1 kcal/mol for the regional fit (see Supporting Information). Addition of cubic terms was also explored but found not to significantly outperform the model of eq 2 (data not shown).

**3.2. Extrapolation Across Basis Sets and Electron Correlation Methods.** This section examines the ability of the canonical model to extrapolate along the two dimensions that establish the level of the quantum chemical computation: correlation method and basis set. The analysis is done at the geometry, shown as a filled square in Figure 3.

Table 2 shows the ability to extrapolate across electron-correlation methods for a variety of basis sets. The results suggest that accurate maps can be developed from low to high levels of correlation. This success is consistent with the assumption of density functional theory (DFT), that the correlation energy is a functional of the one-electron density. The distributed multipoles used as inputs to the model capture the electronic distribution and so contain much of the information present in the one-electron density. In previous work, Janesko et al. used feature extraction algorithms to develop models for correlation energy based explicitly on density matrices.[28] That work developed a model to predict the two-electron density matrix, $\rho^{(2)}{}_{i,j,k,l}$ from the one electron density matrix, $\rho^{(1)}{}_{i,j}$, thereby predicting the correlation energy from the one-electron density, as in DFT methods. However, the use of density matrices has the disadvantage of making it difficult to develop models that connect across basis sets, and it is for this reason that we have developed the DM approach described here.

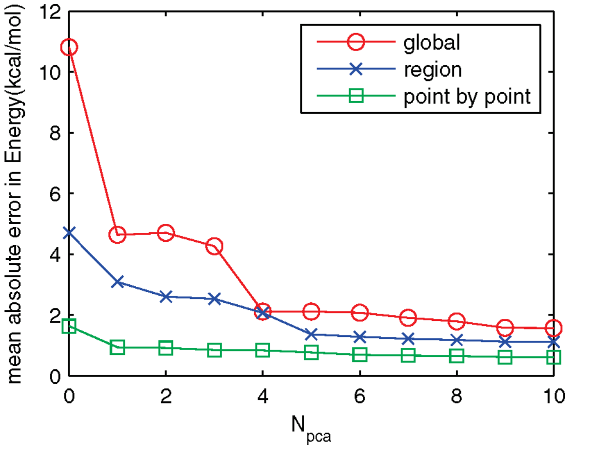


***Figure 9.*** Error versus number of principle components in the linear regression of eq 2 for a model mapping HF/3-21G to CCSD(T)/cc-pVTZ self-energies. Results are shown for global, regional, and point-by-point fits. The total number of fitting parameters in the model of eq 2 is 92 $N_{pca}$ + 1, 6 $N_{pca}$ + 1, and 2 $N_{pca}$ + 1, for point-by-point, regional and global fits respectively.

Table 3 shows the ability to extrapolate across basis sets, for a variety of levels of correlation. Comparison with Table 2 reveals that most of the error in the model predictions arises from mapping across basis sets. For the HF calculations, substantial improvement is attained by using 3-21G as the low-level theory as opposed to that of STO-3G. This result is consistent with Janesko's work on functional group basis sets derived from PCA of natural orbitals, which found that the intrinsic dimensionality of a functional group is larger than the number of degrees of freedom in a minimal (STO-3G) basis and is roughly equivalent to that of a 3-21G basis set.[50] Note, however, that the accuracy of the basis set extrapolation depends on correlation method, with higher levels of correlation requiring larger basis sets as the low-level input to the model.

**3.3. Mapping from Low- To High-Level Potential Energy Surfaces.** This section examines the degree to which a single regression can be used to make predictions for different reaction center geometries. Figure 9 shows that using a single global regression for all the reaction-center geometries (points labeled with symbols in Figure 3) yields an error that is substantially larger than that obtained from a point-by-point fit in which a separate regression is performed at each reaction-center geometry. This is expected, since the number of fitting parameters is substantially larger for a point-by-point fit. Also, the success of the regression is related to molecular similarity, and the reaction center changes its character as the reaction progresses. Regression of the three regions (transition-state and entrance and exit channel geometries of Figure 3) yields better results than a global fit, while still using a single regression to make predictions for a range of geometries. Figures 10 and 11 show

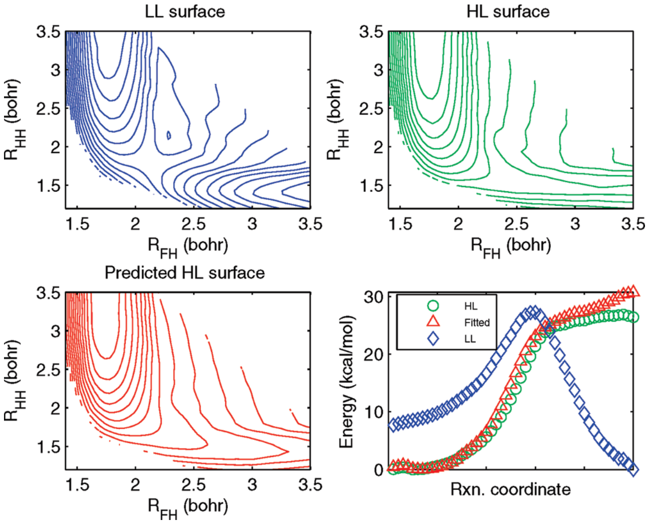


***Figure 10.*** Self-energy obtained from low-level HF/3-21G (left top) and high-level CCSD(T)/cc-pVTZ calculations (right top) and from a global fit from low- to high-level self-energies (left bottom) for the collinear H + HF → $H_2$ + HF trial reaction in a typical environment. The contours are from 0 to 30 kcal/mol in steps of 3 kcal/mol. The reaction energy profiles are compared in the lower right panel.

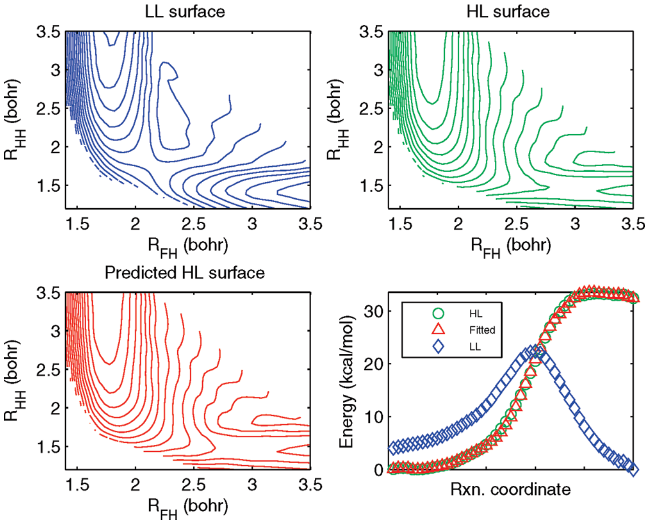


***Figure 11.*** Results of a regional fit from HF/3-21G to CCSD(T)/cc-pVTZ methods for the collinear H + HF → $H_2$ + HF trial reaction in a typical electrostatic environment. The notation is as in Figure 10.

both a representative fitted reaction surface and path. The regional fits substantially outperform global fits and yield smooth potential energy surfaces. Fits of the remaining 249 environments are summarized in Figure 12, which shows the error in the reaction energy profiles (the lower panels of Figures 10 and 11) averaged over all environments. We note that there are a few environments that induce very large changes in the reaction profiles (see Figure 5), and removal of these extreme environments would further reduce the average error.

The results for the regional fits in Figure 12 indicate that it is possible to develop a single regression that can handle a range of reaction-center geometries. The success of such fits relies on grouping of geometries into sets where molecular similarity may be expected to apply. The division into regions, shown in Figure 3, is ad hoc; however, a more systematic approach can be envisioned in which cluster analysis is used to group electronic structures (as opposed to geometries) into similar regions. Such an approach will lead to a model of the type shown in Figure 1, where a decision tree is first used to classify the electronic structure from the low-level theory into a region, and then a region-specific regression is used to map from low-to high-level results. Since the MD algorithm requires forces, smoothing the results from different regions may be necessary, in which case functions that smoothly switch between regressions may be used for cases that lie near boundary regions between clusters. However, no smoothing was used at the boundaries in the current regional and point-by-point fits, and the resulting reaction surfaces and energy profiles are smooth and will lead to smoothly varying forces.

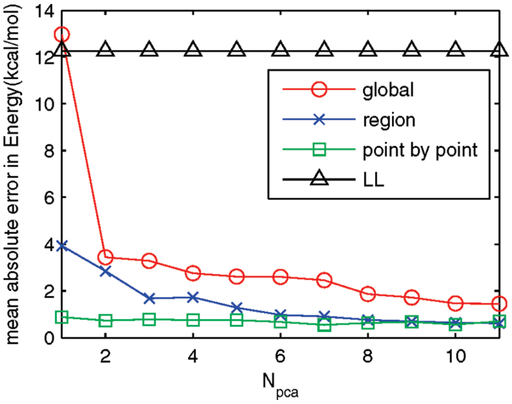


***Figure 12.*** Mean absolute error in the reaction profile energies for a map from HF/3-21G to CCSD(T)/cc-pVTZ theories, averaged over all 250 environments. Results are shown versus number of principle components included in both the linear and quadratic terms of eq 2 for global (green square), regional (blue star) and point-by-point (red circle) fits. The average error between the low- and high-level calculated reaction paths is shown as black triangles for reference.

The results presented above consider only models for the self-energy of the reaction center (eq 2). To compute the interaction energy between the reaction center and the environment, the charge distribution of the reaction center is also needed. This interaction can be well computed from the distributed multipoles of the reaction center,[46,46] and so a model that predicts the high-level distributed multipoles from the low-level distributed multipoles would be sufficient for this purpose. A preliminary investigation revealed that for the entire data set $r^2$ is 0.93 between low-level (HF/3-21G) and high-level (QCISD/6-31++G**) distributed multipoles, as opposed to 0.34 for the correlation between low- and high-level self-energies of the reaction center. This suggests that the prediction of distributed multipoles is a relatively easy task when compared to the prediction of the self-energy, and so predictions of distributed multipoles are not explicitly addressed in this paper.

## 4. Conclusion

Here, we explore the use of molecular similarity to develop models for use in QM/MM simulations that can, at the cost of a low-level ab initio calculation, produce results that approach the accuracy of high-level ab initio calculations. Our approach first generates detailed quantum chemical data on the reaction center in geometric configurations and electrostatic environments that span those expected to arise during the MD trajectory. This data includes results obtained from both low- and high-level ab initio methods. This information is then used to develop a low-cost model that can reproduce the output of the high-level theory using only inputs generated from the low-level theory. This approach was tested on the $H + HF \rightarrow H_2 + F$ collinear reaction. This reaction center is sufficiently small that high-level calculations can be performed quickly. Yet despite the small size, the reaction still involves breaking and forming of bonds in a manner that is sensitive to the environment and provides a realistic test of the approach.

The ability to predict high-level results using only descriptors generated from low-level calculations was tested along the two dimensions that define a quantum chemical method: the level of electron correlation and the size of the basis set. The models predict the results of the high-level theory using, as input, distributed multipoles obtained from the low-level method. The distributed multipoles are monopoles, dipoles, and quadrupoles placed on each atomic center, and they summarize the electronic distribution in a manner that is independent of basis set. Including electron correlation, by predicting QCISD results from HF inputs, leads to an average error of less than 0.05 kcal/mol for split-valence basis sets. This relatively low error is consistent with the assumption of DFT, that the correlation energy is a functional of the electronic density. Extrapolating across basis sets is the primary source of error in the models, with the extrapolation from 6-31G* to cc-pVTZ basis sets giving an error of 0.21 kcal/mol within HF theory and 0.24 kcal/mol within QCISD theory. The models used here were parametrized to about $10^4$ high-level computations and so will lead to substantial savings for situations, such as MD simulations, where the quantum algorithm is called $10^6$ or more times.

An important criterion regarding the applicability of this approach to more complex reaction centers is the extent to which a single model can handle a range of reaction-center geometries. The current study showed that reasonable accuracy can be obtained when configurations of relevance to the collinear trial reaction are broken into three regions: the transition region and the entrance and exit channels. This suggests that regressions can be developed that span fairly large regions of configuration and environment space. The ability of a single regression to describe a range of configurations benefits from the use of the model only to build basis set and correlation corrections onto the energy generated by a low-level method. The energy from the low-level method already contains reasonable estimates to the interactions energies and contains how these vary with geometry.

The use of machine learning to group input configurations into regions, i.e., to develop the optimal decision tree for selecting regressions in Figure 1, should yield even better results than the ad hoc selection of regions used here. This may become especially important for larger reaction centers. Consider, for instance, the important class of biological reactions that involve transfer of a hydrogen atom, a phosphate group or other small molecular fragment between two groups. The dimensions corresponding to fragment transfer will have the greatest total spread. Fluctuations in the orientation of the groups between which the transfer occurs must also be included but with smaller amplitudes, since such motions are often constrained by covalent attachment to the protein backbone.

Inclusion of additional information from the QM methods may also lead to better performing models. In particular, many QM methods have analytical derivative methods that generate forces and higher energy derivatives at little additional cost.[52] These derivatives provide additional information that may aid in the development of the model mapping low- to high-level energies. Analytical gradient information may also allow for a direct and, thus, a more efficient prediction of forces.

The success of the models presented here is encouraging, especially given the simplicity of the methods, i.e., PCA and linear regression, used to discover the models. The use of PCA for feature extraction can be extended both to nonlinear feature extraction methods and to methods that select latent variables based on the importance to the model, as opposed to the variability in the input data. Likewise, the linear regression used here is among the simplest possible means to map between low- and high-level theories, and a wide array of alternative methods from statistics and machine learning can be envisioned.[53]

**Acknowledgment.** This work was funded by the National Science Foundation (CHE-0719350).

**Supporting Information Available:** Maps from (HF/3-21G) to another (QCISD/6-31++G**) theory to show that similar results are obtained for maps to different high-level theories. The cost of and the complexity of the high-level calculation can, therefore, be based on the degree of accuracy appropriate for the given application. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Lin, H.; Truhlar, D. *Theor. Chim. Acta.* **2007**, *117* (2), 185–199.

(2) Friesner, R.; Guallar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389–427.

(3) Senn, H.; Thiel, W. *Top. Curr. Chem.* **2007**, *268*, 173–290.

(4) Senn, H.; Thiel, W. *Angew. Chem. Int.* **2009**, *48* (7), 1198−1229.

(5) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; Konig, P.; Li, G.; Xu, D.; Guo, H. *J. Phys. Chem. B* **2006**, *110* (13), 6458–6469.

(6) Xie, W.; Song, L.; Truhlar, D.; Gao, J. *J. Phys. Chem. B* **2008**, *112* (45), 14124–14131.

(7) Vreven, T.; Byun, K.; Komaromi, I.; Dapprich, S.; Montgomery Jr, J.; Morokuma, K.; Frisch, M. *J. Chem. Theor. Comp.* **2006**, *2* (3), 815–826.

(8) Freindorf, M.; Shao, Y.; Furlani, T.; Kong, J. *J. Comput. Chem.* **2005**, *26* (12), 1270−1278.

(9) Crespo, A.; Scherlis, D.; Martí, M.; Ordejon, P.; Roitberg, A.; Estrin, D. *J. Phys. Chem. B* **2003**, *107* (49), 13728–13736.

(10) Foresman, J. B.; Frisch, A. *Exploring Chemistry With Electronic Structure Methods*; Gaussian Inc.: Pittsburgh, PA, 1996; 123−125.

(11) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry Introduction to Advanced Electronic Structure Theory*; Dover Publications: Mineola, N.Y, 1996; 32−45.

(12) Noodleman, L.; Lovell, T.; Han, W.; Li, J.; Himos, F. *Chem. Rev.* **2004**, *104* (2), 459–508.

(13) Gordon, M.; Mullin, J.; Pruitt, S.; Roskop, L.; Slipchenko, L.; Boatz, J. *J. Phys. Chem. B* **2009**, *113* (29), 9646–9663.

(14) Yang, W. *Phys. Rev. Lett.* **1991**, *66* (11), 1438–1441.

(15) Steele, R.; DiStasio Jr, R.; Shao, Y.; Kong, J.; Head-Gordon, M. *J. Chem. Phys.* **2006**, *125*, 074108.

(16) White, C.; Johnson, B.; Gill, P.; Head-Gordon, M. *Chem. Phys. Lett.* **1996**, *253* (3−4), 268–278.

(17) Wolinski, K.; Pulay, P. *J. Chem. Phys.* **2003**, *118*, 9497.

(18) Lee, M.; Head-Gordon, M. *J. Chem. Phys.* **1997**, *107*, 9085.

(19) Berghold, G.; Parrinello, M.; Hutter, J. *J. Chem. Phys.* **2002**, *116*, 1800.

(20) Lu, W.; Wang, C.; Schmidt, M.; Bytautas, L.; Ho, K.; Ruedenberg, K. *J. Chem. Phys.* **2004**, *120*, 2629.

(21) Dewar, M.; Zoebisch, E.; Healy, E.; Stewart, J. *J. Am. Chem. Soc.* **1985**, *107* (13), 3902–3909.

(22) Ridley, J.; Zerner, M. *Theor. Chim. Acta.* **1973**, *32* (2), 111–134.

(23) Scuseria, G. *J. Phys. Chem. A* **1999**, *103* (25), 4782–4790.

(24) Schutz, M.; Werner, H. *J. Chem. Phys.* **2001**, *114*, 661.

(25) Kohn, W. *Rev. Mod. Phys.* **1999**, *71* (5), 1253–1266.

(26) Goedecker, S. *Rev. Mod. Phys.* **1999**, *71* (4), 1085–1123.

(27) Van der Vaart, A.; Gogonea, V.; Dixon, S.; Merz JR., K. *J. Comput. Chem.* **2000**, *21* (16), 1494−1504.

(28) Janesko, B.; Yaron, D. *J. Chem. Phys.* **2003**, *119*, 1320−1328.

(29) Greengard, L.; Rokhlin, V. *J. Comput. Phys.* **1987**, *73* (2), 325–348.

(30) Machida, K. *Principles of Molecular Mechanics*; John Wiley & Sons Inc: New York, 1999, 15−19.

(31) Tangney, P.; Scandolo, S. *J. Chem. Phys.* **2002**, *117*, 8898.

(32) Ercolessi, F.; Adams, J. *Europhys. Lett.* **1994**, *26* (8), 583–588.

(33) Mehl, M.; Papaconstantopoulos, D. *Phys. Rev. B: Condens. Matter* **1996**, *54* (7), 4519–4530.

(34) Tabacchi, G.; Mundy, C.; Hutter, J.; Parrinello, M. *J. Chem. Phys.* **2002**, *117*, 1416.

(35) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B.* **2001**, *105* (2), 569–585.

(36) Sastry, K.; Johnson, D.; Thompson, A.; Goldberg, D.; Martinez, T.; Leiding, J.; Owens, J. *Mater. Manuf. Processes* **2007**, *22* (5), 553–561.

(37) Werner, H.; Kallay, M.; Gauss, J. *J. Chem. Phys.* **2008**, *128*, 034305.

(38) Stark, K.; Werner, H.-J. *J. Chem. Phys.* **1996**, *104* (17), 6515–6530.

(39) Gonzalez-Luque, R.; Merchan, M.; Roos, B. O. *Chem. Phys.* **1993**, *171* (1−2), 107–118.

(40) Cardoen, W.; Gdanitz, R.; Simons, J. *J. Phys. Chem. A* **2006**, *110* (2), 564–571.

(41) Stone, A.; Alderton, M. *Mol. Phys.* **2002**, *100* (1), 221–233.

(42) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A. *Gaussian 03*, Revision 6.0; Gaussian, Inc.: Wallingford CT, 2004.

(43) Stone, A. *The Theory of Intermolecular Forces*; Clarendon Press: Oxford, U.K., 1997, 105−119.

(44) Stone, A. *J. Chem. Theor. Comp.* **2005**, *1* (6), 1128–1132.

(45) Gordon, M.; Freitag, M.; Bandyopadhyay, P.; Jensen, J.; Kairys, V.; Stevens, W. *J. Phys. Chem. A.* **2001**, *105* (2), 293–307.

(46) Gordon, M.; Slipchenko, L.; Li, H.; Jensen, J. *Ann. Rep. Comp. Chem.* **2007**, 177.

(47) Stone, A. *Distributed Multipole Analysis of Gaussian98 Wavefunctions, Revision 2.1*; University of Cambridge: UK 1999.

(48) Stone, A.; Dullweber, A.; Hodges, M.; Popelier, P.; Wales, D. *ORIENT*, Version 4.6; University of Cambridge: Cambridge, U.K., 1995.

(49) Janesko, B.; Yaron, D. *J. Chem. Phys.* **2004**, *121* (5635).

(50) Janesko, B.; Yaron, D. *J. Chem. Theory Comp.* **2005**, *1* (2), 267–278.

(51) Duda, R.; Hart, P.; Stork, D. *Pattern Classification*; Wiley: New York: 2001, 114−117.

(52) Pulay, P. *Adv. Chem. Phys.* **1987**, *69*, 241–286.

(53) Mitchell, T.; *Machine Learning*; WCB/McGraw Hill: Hightstown, NJ, 1997, 5−14.

CT9004195

JCTC Journal of Chemical Theory and Computation

# Oxidation of the Benzyl Radical: Mechanism, Thermochemistry, and Kinetics for the Reactions of Benzyl Hydroperoxide

Gabriel da Silva,*,† M. Rafiq Hamdan,† and Joseph W. Bozzelli*,‡

*Department of Chemical and Biomolecular Engineering, The University of Melbourne, Victoria 3010, Australia, and Department of Chemistry and Environmental Science, New Jersey Institute of Technology, Newark, New Jersey 07102*

Received July 12, 2009

**Abstract:** Oxidation of the benzyl radical plays a key role in the autoignition, combustion, and atmospheric degradation of toluene and other alkylated aromatic hydrocarbons. Under relevant autoignition conditions of moderate temperature and high pressure, and in the atmosphere, benzyl reacts with $O_2$ to form the benzylperoxy radical, and the further oxidation reactions of this radical are not yet fully characterized. In this contribution, we further develop the reaction chemistry, thermodynamics, and kinetics of benzyl radical oxidation, highlighting the important role of benzyl hydroperoxide and the benzoxyl (benzyloxyl) radical. The benzylperoxy + H reaction mechanism is studied using computational chemistry and statistical reaction rate theory. High-pressure limit rate constants in the barrierless benzylperoxy + H association are obtained from variational transition state theory calculations, with internal rotor contributions. The benzylperoxy + H reaction is seen to produce an activated benzyl hydroperoxide adduct that has 87 kcal mol$^{-1}$ excess energy over the ground state. We show that this activated adduct proceeds almost exclusively to the benzoxyl radical + OH across a wide range of temperature and pressure conditions. Minor reaction paths include benzyl + $HO_2$, α-hydroxylbenzyl + OH, and benzaldehyde + $H_2O$, each constituting around 1% of the total reaction rate at higher temperatures. Thermal decomposition of benzyl hydroperoxide, formed by hydrogen abstraction reactions in the benzylperoxy radical and at low temperatures in the benzylperoxy + H and benzyl + $HO_2$ reactions, is also investigated. Decomposition to benzoxyl + OH is fast at temperatures of 900 K and above. The contribution of benzyl hydroperoxide chemistry to the ignition and oxidation of alkylated aromatics is discussed. Benzyl radical oxidation chemistry achieves the conversion of toluene to benzaldehyde, aiding autoignition via processes that either release large amounts of energy or form reactive free radicals through chain-branching.

## Introduction

Alkylated aromatic hydrocarbons are a major and growing component of liquid transportation fuels, including gasoline and jet fuel. It is important that we understand the ignition and oxidation chemistry of these fuel components across the range of temperature and pressure conditions encountered in spark ignition and jet engines. Particular uncertainties remain with regards to the chemistry taking place during the autoignition of these alkylated aromatics, where high pressures (tens of atmospheres) and moderate temperatures (ca. 800−1200 K) are encountered. Modeling autoignition behavior is important, for example, in understanding engine knock and $NO_x$ formation and in designing advanced homogeneous charge compression ignition (HCCI) engines. Understanding the low-temperature oxidation chemistry of

* To whom correspondence should be addressed. E-mail: gdasilva@unimelb.edu.au (G.d.S.), bozzelli@njit.edu (J.W.B.).

† The University of Melbourne.

‡ New Jersey Institute of Technology.

10.1021/ct900352f CCC: $40.75 © 2009 American Chemical Society
Published on Web 11/04/2009

alkylated aromatics in the atmosphere is also of significance, as these compounds are a major component of air pollution in urban environments.

As the parent alkylated aromatic, much attention has been paid to the oxidation of toluene (methylbenzene). It is well-known that the initial stages of toluene oxidation predominantly result in the formation of the benzyl radical.[1] The benzyl radical is thermally stable,[2] and at low to moderate temperatures it is removed from combustion systems by oxidation reactions with species such as $O_2$,[3] OH,[4] O,[5] and $HO_2$.[6] Benzyl associates with $O_2$ to form the benzylperoxy radical in a mildly exothermic reaction (ca. 20 kcal $mol^{-1}$),[3] and the benzylperoxy adduct has little excess energy to go into forward reactions to new, dissociated products. At higher temperatures, the activated benzylperoxy adduct does form some phenol + OH, but it predominantly dissociates back to benzyl + $O_2$. At lower temperatures and higher pressures, the activated benzylperoxy adduct is stabilized by bath gas collisions and is available to participate in further bimolecular reactions. The benzylperoxy radical is known to undergo a self-reaction to products including two benzoxyl radicals + $O_2$ or to react with $HO_2$ to form benzyl hydroperoxide + $O_2$ (an important process in the atmospheric oxidation of toluene).[7] Benzylperoxy can also abstract a H atom from surrounding hydrocarbons or react with free H atoms to produce the benzyl hydroperoxide molecule. Benzyl hydroperoxide decomposes to the benzoxyl radical + OH with a relatively low barrier,[6] and the benzoxyl radical then undergoes chain-propagating decomposition reactions, mainly to benzaldehyde + H.[8] In competition with bimolecular reactions, the benzylperoxy radical will decompose to benzyl + $O_2$ with an activation (dissociation) energy of around 20 kcal $mol^{-1}$. Bimolecular reactions are expected to dominate at low temperatures, where benzylperoxy radical lifetimes are large with respect to thermal decomposition, but they may also be important at higher temperatures, where an equilibrium concentration of benzylperoxy should be established.

In this study, we investigate the kinetics and products of the benzylperoxy + H reaction, using theoretical thermochemical kinetic techniques. The reaction of benzylperoxy with free H atoms should be of significance to fuel-rich flames, where H atoms are found at relatively high concentrations. Kinetics of the benzylperoxy + H association reaction are treated using variational transition state theory. Further reaction of the activated benzyl hydroperoxide adduct is studied as a function of temperature and pressure in master equation simulations, with RRKM theory for $k(E)$, providing branching ratios and apparent rate constants for input to kinetic models. The role of benzyl hydroperoxide chemistry in the oxidation and autoignition of alkylated aromatic hydrocarbons is discussed.

## Computational Methods

The G3B3 composite theoretical method is used to study all species.[9] The G3B3 method uses B3LYP/6-31G(d)-optimized structures and frequencies, with higher-level corrections for accurate energies. All electronic structure calculations are performed using Gaussian 03.[10] G3B3 results for benzyl hydroperoxide and its decomposition products are taken from a previous study,[6] while the benzylperoxy radical and transition state structures for H addition are newly studied here. The G3B3 calculations represent a compromise between accuracy and computational efficiency for the relatively large species studied here (nine heavy atoms). Using the G2/97 test set the G3B3 method reproduces a range of thermochemical properties with a root-mean-square error of ±1.0 kcal $mol^{-1}$.[9] Our reported reaction enthalpies and barrier heights are thought to be accurate to ±2 kcal $mol^{-1}$ (around two standard deviations). Optimized structures and vibrational frequencies are provided in the Supporting Information for benzylperoxy, benzyl hydroperoxide, and the transition state structures **TS1−TS4**.

The standard enthalpy of formation of the benzylperoxy radical is calculated from an atomization work reaction. Here, the 0 K reaction energy for formation of the atomic constituents in their ground state is first determined, and then the 0 K benzyl hydroperoxide heat of formation is obtained using atom reference enthalpies of 69.977, 51.634, 58.984 kcal $mol^{-1}$ for the C, O, and H,[11] as recently recommended.[12] The 0 K heat of formation is adjusted to 298 K using enthalpy corrections ($H_{298} - H_0$) of 0.251, 1.037, and 1.010 kcal $mol^{-1}$, for C, O, and H. Entropy and heat capacity values are obtained using statistical mechanics techniques, with the rigid rotor−harmonic oscillator (RRHO) approximation. Vibrational frequencies for rotation about $C_6H_5$—$CH_2OO$ and $C_6H_5CH_2$—OO bonds in the benzylperoxy radical were removed from the RRHO analysis and treated as hindered internal rotors using B3LYP/6-31G(d) rotor potentials.

The kinetics of H addition to the peroxy radical site in benzylperoxy is evaluated using variational transition state theory.[13] The minimum energy potential (MEP) for H addition is calculated at the B3LYP/6-31G(d) level of theory and scaled by the G3B3 reaction enthalpy. Rate constants are calculated as a function of temperature for structures at 0.1 Å intervals along the MEP according to canonical transition state theory, in the program ChemRate.[14] The C−C and C−O internal rotors in the transition state structures are modeled using rotor potentials from the benzylperoxy radical, while the O−O rotor is treated as being similar to that in benzyl hydroperoxide (a single-fold rotor with 6.3 kcal $mol^{-1}$ barrier).[6] Rate constants are minimized as a function of position along the MEP to obtain the canonical variational rate constant at each temperature. All structures on the MEP possess a single imaginary frequency, with the mode of vibration corresponding to motion along the bond-breaking coordinate. The use of canonical transition state theory in the variational analysis neglects conservation of angular momentum and is thus expected to provide an upper limit to the true addition rate constant.

Apparent rate constants in the activated benzylperoxy + H reaction mechanism are obtained from master equation simulations, with RRKM theory for $k(E)$, in the ChemRate program. Simulations are performed for pressures between 0.001 and 1000 atm and temperatures between 300 and 2000 K. Collisional energy transfer is described using an exponential-down model, with $\Delta E_{down} = 500$ $cm^{-1}$, and the bath gas is $N_2$. Lennard-Jones parameters used for the $C_7H_8O_2$

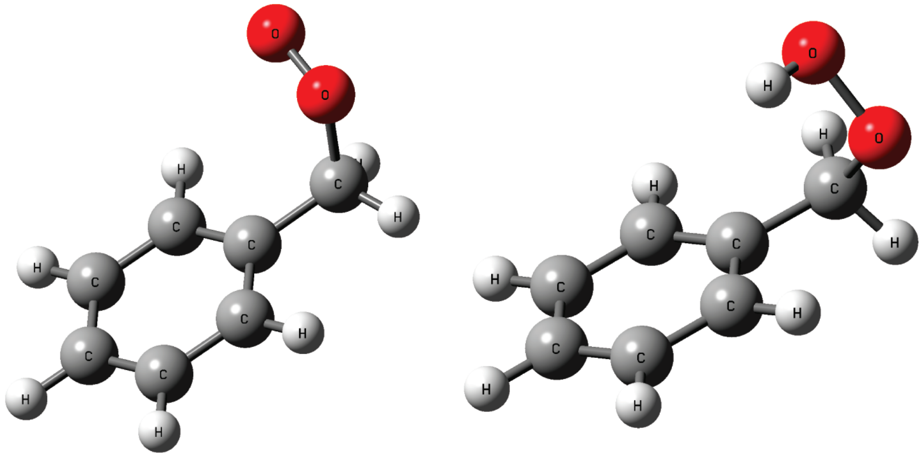


***Figure 1.*** Optimized structures for the benzylperoxy radical and benzyl hydroperoxide (B3LYP/6-31G(d)).

***Table 1.*** Vibrational Frequencies ($cm^{-1}$) of the Benzylperoxy Radical [B3LYP/6-31G(d)][a]

| | | |
|---|---|---|
| **38.33** | **89.66** | 142.48 |
| 296.97 | 319.25 | 416.74 |
| 473.09 | 512.91 | 611.95 |
| 635.67 | 712.10 | 764.54 |
| 829.92 | 855.97 | 859.73 |
| 935.16 | 969.48 | 999.90 |
| 1005.61 | 1019.60 | 1058.54 |
| 1121.95 | 1171.11 | 1194.77 |
| 1215.62 | 1237.28 | 1278.88 |
| 1361.38 | 1369.22 | 1379.85 |
| 1503.17 | 1505.45 | 1547.22 |
| 1645.97 | 1665.05 | 3095.08 |
| 3156.55 | 3180.64 | 3188.42 |
| 3198.85 | 3208.27 | 3215.64 |

[a] Values listed in bold correspond to internal rotational modes.

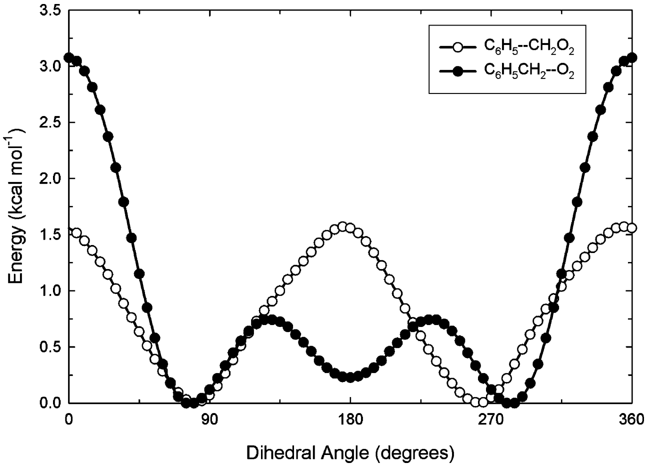


***Figure 2.*** Internal rotor potentials in the benzylperoxy radical [B3LYP/6-31G(d)].

species are $\sigma = 7$ Å and $\varepsilon/k_b = 500$ K. All rate constants quoted in this study are in $s^{-1}$ or $cm^3\ mol^{-1}\ s^{-1}$ units, with activation energies in kcal $mol^{-1}$ and temperatures in K.

## Results and Discussion

**Properties of the Benzylperoxy Radical.** The optimized benzylperoxy radical structure is illustrated in Figure 1, compared to that for benzyl hydroperoxide. Vibrational frequencies for benzylperoxy are listed in Table 1. The 38.33 $cm^{-1}$ vibration is attributed to internal rotation about the $C_6H_5$—$CH_2O_2$ bond, while the 89.66 $cm^{-1}$ vibration corresponds to rotation around $C_6H_5CH_2$—$O_2$, although both internal rotational modes appear significantly coupled. Rotor profiles obtained from relaxed B3LYP/6-31G(d) scans about these C—C and C—O bonds are presented in Figure 2.

Thermochemical properties [$\Delta_f H°_{298}$, $S°_{298}$, $C_p(T)$] are reported in Table 2 for the benzylperoxy radical, as well all other $C_7$ and $C_6$ species in the benzylperoxy + H mechanism (from ref [6]). Smaller decomposition fragments are modeled using literature thermochemistry.[15−18] The G3B3 benzylperoxy heat of formation is calculated as 29.6 kcal $mol^{-1}$, while Fenter et al.[3b] report a relatively similar value of 28.0 ± 1.4 kcal $mol^{-1}$. According to our reported thermochemistry, the benzylperoxy + H reaction is 86.7 kcal $mol^{-1}$ exothermic. This high exothermicity means that the benzyl hydroperoxide adduct produced in this association process will be highly activated, with more than enough energy to proceed on to new decomposition products.

**Variational Analysis.** Rate constants for the barrierless benzylperoxy + H addition reaction and the reverse dissociation process have been calculated according to canonical variational transition state theory. The general procedure employed here has been successfully used to calculate rate constants in barrierless H association[2d,19] and other[13,20] reactions.

A minimum energy profile for H addition to benzylperoxy, at the G3B3//UB3LYP/6-31G(d) level of theory, is depicted in Figure 3. The dissociation reaction has a loose transition state structure, with energies within 1 kcal $mol^{-1}$ of the dissociated products at O—H bond lengths of 2.8 Å and greater. A very loose structure is not unexpected for this radical recombination reaction, given the large enthalpy change. Rate constants calculated at each contributing transition state structure are listed in the Supporting Information. The association reaction is found to be controlled by a very loose 3.0 Å structure at 300 K, tightening to the 2.0 Å structure at 2000 K. At 2.0 Å, the transition state energy is 6.1 kcal $mol^{-1}$ below that of the dissociated products, while at 3.0 Å it is only 0.7 kcal $mol^{-1}$ below. Fitting the minimum rate constants to a three-parameter modified Arrhenius equation using a least-squares procedure, we obtain the rate

***Table 2.*** Enthalpies of Formation ($\Delta_f H^\circ_{298}$, kcal mol$^{-1}$), Entropies ($S^\circ_{298}$, cal mol$^{-1}$ K$^{-1}$), and Heat Capacities [$C_p(T)$, $T$ = 300−2000 K, cal mol$^{-1}$ K$^{-1}$] for Selected Species in the Benzylperoxy + H Reaction Mechanism

| | $\Delta_f H^\circ_{298}$ | $S^\circ_{298}$ | $C_p(300)$ | $C_p(400)$ | $C_p(500)$ | $C_p(600)$ | $C_p(800)$ | $C_p(1000)$ | $C_p(1500)$ | $C_p(2000)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| benzylperoxy | 29.6 | 93.218 | 30.993 | 39.746 | 47.207 | 53.241 | 62.103 | 68.231 | 77.257 | 81.759 |
| benzyl hydroperoxide | −5.0 | 93.275 | 32.699 | 42.050 | 50.181 | 56.802 | 66.439 | 72.940 | 82.225 | 86.756 |
| benzaldehyde | −8.3 | 79.021 | 25.984 | 33.933 | 40.804 | 46.441 | 54.806 | 60.557 | 68.717 | 72.543 |
| benzyne | 111.1 | 69.075 | 18.851 | 24.536 | 29.302 | 33.128 | 38.715 | 42.553 | 48.161 | 50.937 |
| benzoxyl | 31.1 | 84.277 | 27.994 | 36.292 | 43.419 | 49.233 | 57.856 | 63.856 | 72.685 | 77.062 |
| α-hydroxybenzyl | 11.1 | 82.855 | 29.443 | 37.990 | 45.069 | 50.696 | 58.870 | 64.536 | 72.997 | 77.185 |

expression $k$ [cm$^3$ mol$^{-1}$ s$^{-1}$] = $1.81 \times 10^{12} T^{0.48} \exp(-0.21/T)$ [the reverse rate expression is $k$ [s$^{-1}$] = $6.97 \times 10^{14} T^{0.01} \exp(-43.73/T)$].

The variational rate constant for benzylperoxy + H is plotted in Figure 4, as a function of temperature. In addition to the hindered rotor (HR) treatment of low-frequency vibrations used here, this rate constant has also been calculated using free rotor (FR) and RRHO treatments, with the results included in Figure 4. The FR rate constants are obtained by treating the $C_6H_5CH_2O-OH$ internal rotor in benzyl hydroperoxide and in the transition state structures as a free rotation, with all remaining modes treated as vibrational frequencies. Because the C−O and C−C rotations are conserved in the reactants, transition states, and products, their treatment should have only a minor effect on the rate constant. In Figure 4 we show that the RRHO treatment provides rate constants that are consistently around $10^{13}$ cm$^3$ mol$^{-1}$ s$^{-1}$ from 300 to 2000 K, with the FR rate constants being higher by around a factor of 5. Rate constants obtained with the HR treatment are intermediate between the other results, being close to the RRHO rate constants at low temperatures (around $2 \times 10^{13}$ cm$^3$ mol$^{-1}$ s$^{-1}$), increasing to be similar to the FR rate constants at higher temperatures (around $5 \times 10^{13}$ cm$^3$ mol$^{-1}$ s$^{-1}$). The HR results are used further in our RRKM modeling of the benzylperoxy + H reaction. We note that, in the temperature range of interest (ca. 1000 K and above), both free and hindered rotor treatment of the internal rotational modes provide rate constants of similar magnitude.

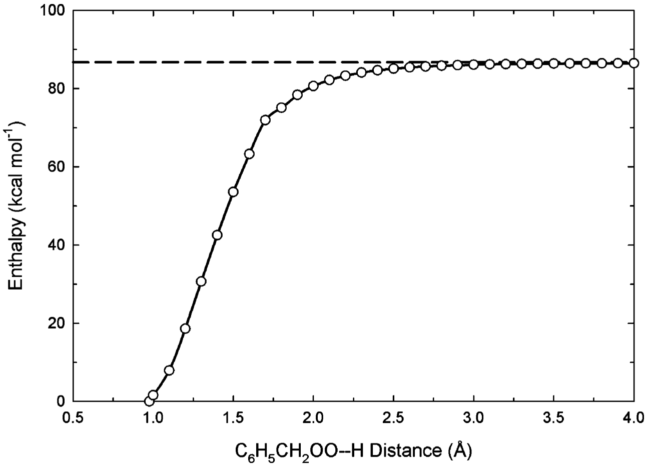


***Figure 3.*** Minimum energy potential for O−H bond dissociation in benzyl hydroperoxide at the G3B3//UB3LYP/6-31G(d) level. The dashed line indicates the energy of infinitely separated products.

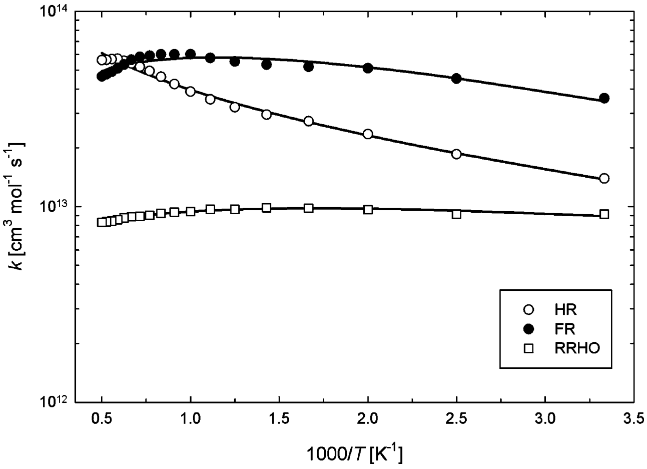


***Figure 4.*** Variational rate constants for the barrierless benzylperoxy + H reaction, calculated using hindered rotor (HR), free rotor (FR), and rigid rotor−harmonic oscillator (RRHO) treatments. Solid lines represent three-parameter Arrhenius fits.

**Benzylperoxy + H Kinetics.** The benzylperoxy + H reaction process is evaluated using the energy surface depicted in Figure 5 (transition state structures are shown in Figure 6). The activated benzyl hydroperoxide adduct is seen to have sufficient energy to proceed to a range of products, which have been discussed in detail elsewhere.[6] The lowest-energy pathway available is for the concerted formation of benzoxyl + OH, where the barrier height is 41.1 kcal mol$^{-1}$ below the entrance channel. Because this is a simple dissociation, without any intrinsic activation barrier, the reaction is entropically favored with a loose transition state structure (i.e., large pre-exponential factor). A second higher-energy pathway to benzoxyl + OH is also depicted (**TS1** in Figure 6), which proceeds in a stepwise mechanism via the 3-methoxy-4-hydroxy-2,5-cyclohexadien-1-yl radical, although this reaction channel is not expected to contribute significantly to benzoxyl formation. The next most energetically favored pathway, behind concerted formation of benzoxyl + OH, is $H_2O$ elimination to benzaldehyde (**TS2**). This reaction requires a barrier that is 40.6 kcal mol$^{-1}$ below the entrance channel, although the tight transition state structure results in a small pre-exponential factor. Other product sets considered include α-hydroxybenzyl ($C_6H_5C\bullet HOH$) + OH (**TS3**), benzyne ($C_6H_4$) + HCHO + $H_2O$ (**TS4**, which proceeds to the unstable $C_6H_4CH_2O$ intermediate), and benzyl + $HO_2$. The reaction to benzyl + $HO_2$ requires the largest barrier of any of the forward reaction paths (26.3 kcal mol$^{-1}$ below the entrance channel); however,

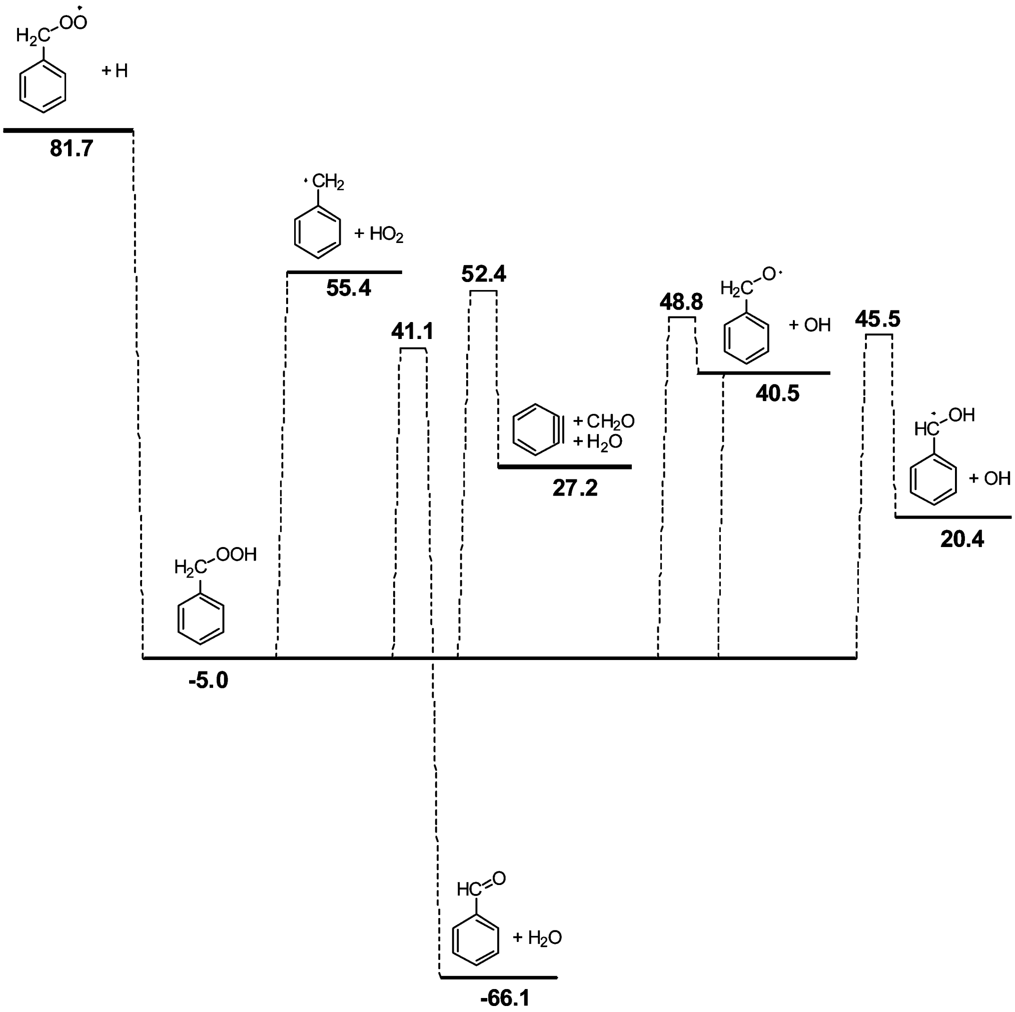


***Figure 5.*** Energy diagram for the benzylperoxy + H reaction mechanism (G3B3 298 K enthalpies).

the loose (barrierless) transition state structure for C−OOH dissociation results in a favorable pre-exponential factor.

Fitted rate constants to important products in the benzylperoxy + H reaction, at pressures between 0.01 and 100 atm, are listed in Table 3. Rate constants to all considered product sets in the benzylperoxy + H mechanism are plotted in Figure 7, at 10 atm pressure (typical of autoignition conditions, where this reaction is expected to be of most significance). Collisional stabilization of benzyl hydroperoxide is the dominant channel at low temperatures, but it becomes unimportant at temperatures of 800 K and above, even at these relatively high pressures. From around 700 K and above, benzoxyl + OH (via barrierless O−OH dissociation) are by far the dominant products, due to the favorable enthalpy and entropy of this path. Even at high temperature, the benzoxyl + OH product is formed with rate constant close to 2 orders of magnitude higher than that of any other product set. At higher temperatures, the formation of benzyl + $HO_2$ plays a small role, contributing around 2% of the total products at 1500 K and above. The α-hydroxybenzyl + OH product set is the next most important, accounting for around 1% of the forward reaction at higher temperatures, followed by benzaldehyde plus $H_2O$. The reaction to benzaldehyde + $H_2O$ and all slower reactions are deemed to be unimportant and are not considered further.

It is clear from our results that under relevant autoignition conditions the dominant products of the benzylperoxy + H reaction are benzoxyl + OH. The potential significance of this reaction in toluene oxidation is discussed later in this contribution. For 10 atm we predict some formation of benzyl hydroperoxide at temperatures below 800 K, and the potential importance of these products is considered here. Figure 8 shows a plot of the branching ratio to benzyl hydroperoxide collisional stabilization in the benzylperoxy + H reaction, as a function of temperature and pressure. At pressures below 1 atm, benzyl hydroperoxide is never an important reaction product, even at very low temperatures (e.g., 500 K). At between 10 and 100 atm, representative of ignition conditions in an SI engine, quenching of the benzyl hydroperoxide adduct is significant at up to moderate temperatures, but drops away rapidly as we approach 800 K, which is at the lower end of ignition temperatures. Similar results were found previously for benzyl hydroperoxide formed in the benzyl + $HO_2$ mechanism.

We have investigated the effect of collisional energy transfer on benzyl hydroperoxide formation, simulating the reaction kinetics with $\Delta E_{\mathrm{down}}$ values of between 500 and 3000 $cm^{-1}$. While $N_2$ (which we used as the buffer gas in our simulations) is a poor collider, the presence of more efficient colliders like toluene and other hydrocarbons in an actual

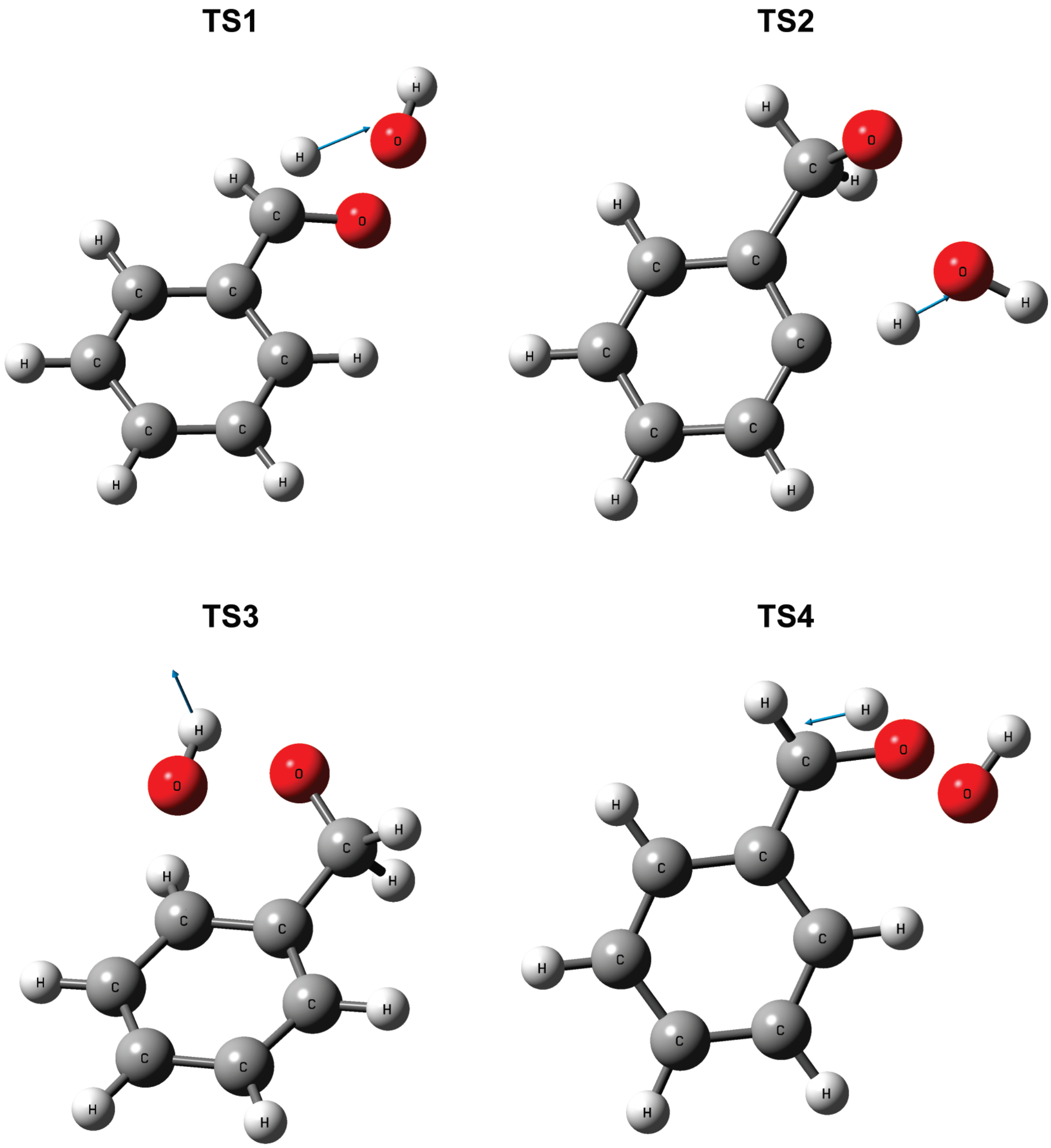


***Figure 6.*** Transition state structures for reactions in the benzylperoxy + H mechanism: benzaldehyde + $H_2O$ (**TS1**), benzyne + HCHO + $H_2O$ (**TS2**), 3-methoxy-4-hydroxy-2,5-cyclohexadien-1-yl (**TS3**), and α-hydroxybenzyl + OH (**TS4**). B3LYP/6-31G(d) level; displacement vectors illustrated.

flame can lead to improved collisional energy transfer. This phenomenon may be the reason why $\Delta E_{\mathrm{down}}$ values on the order of 2000 $\mathrm{cm}^{-1}$ were required to reproduce experimental falloff behavior in several unimolecular reactions that we recently studied.[2d,21] Branching ratios to benzyl hydroperoxide formation, as a function of $\Delta E_{\mathrm{down}}$, are plotted in Figure 9; while large values of $\Delta E_{\mathrm{down}}$ increase the yield of benzyl hydroperoxide at low temperatures, this product still becomes negligible for 800 K and above. At these temperatures, the benzyl hydroperoxide adduct lifetime is short toward decomposition to benzoxyl + OH, relative to collision stabilization, and cannot be quenched in any appreciable quantity. Accordingly, benzyl hydroperoxide formed via the benzylperoxy + H reaction (or benzyl + $HO_2$) is unlikely to play a role in toluene oxidation.

**Benzyl Hydroperoxide Decomposition.** Our kinetic simulations demonstrate that chemically activated benzyl hydroperoxide formed in the benzylperoxy + H association reacts to new products, with negligible collisional stabilization. The same applies for the benzyl + $HO_2$ reaction. Benzyl hydroperoxide can form, however, via other routes. Particularly, benzylperoxy will abstract a H atom from $HO_2$,[7] toluene, and other molecules with benzylic or allylic C−H bonds, forming benzyl hydroperoxide. While the $C_6H_5CH_2OO{-}H$ bond in benzyl hydroperoxide is weak (86.7 kcal $\mathrm{mol}^{-1}$), the $C_6H_5CH_2{-}H$ bond in toluene is similar (91.7 kcal $\mathrm{mol}^{-1}$), making this abstraction reaction almost thermoneutral (5 kcal $\mathrm{mol}^{-1}$ endothermic). Reaction of benzylperoxy with $HO_2$ is an important process in the atmospheric degradation of toluene and should also be of some significance in combustion systems.[7]

Benzyl hydroperoxide decomposes to benzoxyl + OH according to the high-pressure limit rate expression $k$ [$\mathrm{s}^{-1}$] = $3.29 \times 10^{13}T^{0.42}\exp(-20.08/T)$. This corresponds to an activation energy of 39.89 kcal $\mathrm{mol}^{-1}$ and a pre-expontial factor ($A'T^n$) of $6 \times 10^{14}$ $\mathrm{s}^{-1}$ at 1000 K. Rate constants have been calculated for this decomposition reaction from a steady-state solution of the master equation, at pressures

***Table 3.*** Apparent Rate Parameters to Important Product Sets in the Benzylperoxy + H Reaction as a Function of Pressure

| | $A'$ ($cm^3$ $mol^{-1}$ $s^{-1}$) | $n$ | $E_a$ (kcal $mol^{-1}$) |
|---|---|---|---|
| benzylperoxy + H → benzoxyl + OH (0.01 atm) | $2.90 \times 10^{12}$ | 0.41 | 0.47 |
| benzylperoxy + H → benzoxyl + OH (0.1 atm) | $1.49 \times 10^{13}$ | 0.21 | 0.87 |
| benzylperoxy + H → benzoxyl + OH (1 atm) | $3.80 \times 10^{14}$ | −0.19 | 1.89 |
| benzylperoxy + H → benzoxyl + OH (10 atm) | $1.36 \times 10^{17}$ | −0.87 | 4.49 |
| benzylperoxy + H → benzoxyl + OH (100 atm) | $8.26 \times 10^{15}$ | −0.42 | 5.83 |
| benzylperoxy + H → benzyl hydroperoxide (0.01 atm, <700 K) | $5.66 \times 10^{48}$ | −14.95 | 2.82 |
| benzylperoxy + H → benzyl hydroperoxide (0.01 atm, ≥ 700 K) | $1.49 \times 10^{252}$ | −84.99 | 13.99 |
| benzylperoxy + H → benzyl hydroperoxide (0.1 atm, <700 K) | $1.39 \times 10^{57}$ | −16.16 | 6.39 |
| benzylperoxy + H → benzyl hydroperoxide (0.1 atm, ≥700 K) | $1.28 \times 10^{141}$ | −51.10 | −44.94 |
| benzylperoxy + H → benzyl hydroperoxide (1 atm, <700 K) | $4.35 \times 10^{60}$ | −15.92 | 11.40 |
| benzylperoxy + H → benzyl hydroperoxide (1 atm, ≥700 K) | $4.65 \times 10^{125}$ | −46.44 | −60.47 |
| benzylperoxy + H → benzyl hydroperoxide (10 atm, <700 K) | $1.91 \times 10^{31}$ | −5.87 | 4.98 |
| benzylperoxy + H → benzyl hydroperoxide (10 atm, ≥700 K) | $4.77 \times 10^{287}$ | −95.60 | 6.05 |
| benzylperoxy + H → benzyl hydroperoxide (10 atm, <700 K) | $2.25 \times 10^{31}$ | −5.90 | 5.00 |
| benzylperoxy + H → benzyl hydroperoxide (100 atm, ≥700 K) | $4.78 \times 10^{307}$ | −101.60 | 15.47 |
| benzylperoxy + H → benzyl + $HO_2$ (0.01 atm) | $5.73 \times 10^{1}$ | 3.18 | −0.33 |
| benzylperoxy + H → benzyl + $HO_2$ (0.1 atm) | $8.48 \times 10^{3}$ | 2.55 | 0.81 |
| benzylperoxy + H → benzyl + $HO_2$ (1 atm) | $1.96 \times 10^{4}$ | 2.47 | 1.43 |
| benzylperoxy + H → benzyl + $HO_2$ (10 atm) | $1.63 \times 10^{9}$ | 1.07 | 5.06 |
| benzylperoxy + H → benzyl + $HO_2$ (100 atm) | $5.23 \times 10^{1}$ | 3.38 | 3.79 |
| benzylperoxy + H → α-hydroxybenzyl + OH (0.01 atm) | $9.76 \times 10^{\circ}$ | 3.30 | −0.87 |
| benzylperoxy + H → α-hydroxybenzyl + OH (0.1 atm) | $8.90 \times 10^{1}$ | 3.03 | −0.31 |
| benzylperoxy + H → α-hydroxybenzyl + OH (1 atm) | $4.22 \times 10^{3}$ | 2.56 | 0.93 |
| benzylperoxy + H → α-hydroxybenzyl + OH (10 atm) | $2.17 \times 10^{5}$ | 2.12 | 3.15 |
| benzylperoxy + H → α-hydroxybenzyl + OH (100 atm) | $1.43 \times 10^{1}$ | 3.44 | 3.36 |

between 0.01 and 100 atm; the results are plotted in Figure 10, with fitted rate expressions listed in Table 4. Benzyl hydroperoxide lifetimes are shorter than 1 ms at temperatures of around 900 K and above for all pressures, making this reaction likely to proceed in an internal combustion engine. At temperature and pressure conditions relevant to the troposphere (1 atm and 300 K), the lifetime of benzyl hydroperoxide toward unimolecular decomposition is around $1 \times 10^{15}$ s, making this reaction unimportant. Instead, benzyl hydroperoxide will be photolyzed, or will participate in bimolecular reactions with, for example, OH radicals. Several pathways to benzoxyl other than benzyl hydroperoxide decomposition are available at ambient conditions (for example, benzylperoxy + NO and the benzylperoxy self-reaction). Benzoxyl formed via these processes will decompose to benzaldehyde + H with a lifetime of around 1 s,[8] and this should be the dominant mechanism for benzoxyl radical removal in the troposphere.

## Discussion

It is apparent, from work presented here and elsewhere, that benzyl hydroperoxide and the benzoxyl radical are key intermediates in benzyl radical oxidation, particularly under conditions relevant to autoignition and atmospheric oxidation. Scheme 1 depicts the major reaction pathways expected to take place in benzyl radical oxidation under ignition and/or atmospheric conditions, based upon our current understanding (some species are formed as transient activated adducts and/or as stable quenched intermediates). Another potential

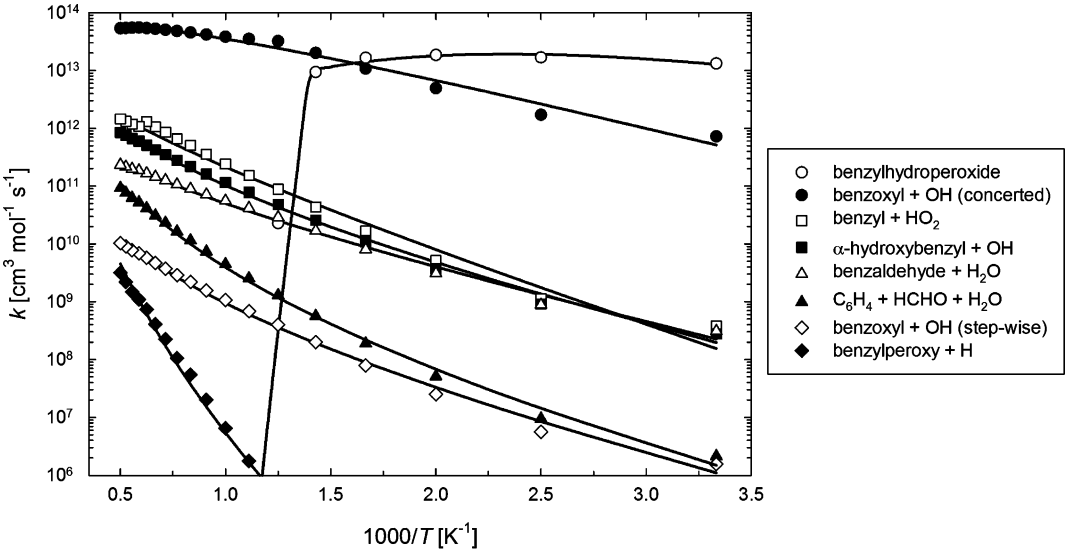


***Figure 7.*** Apparent rate constants at 10 atm to all considered product sets in the benzylperoxy + H reaction mechanism. Solid lines represent three-parameter Arrhenius fits.

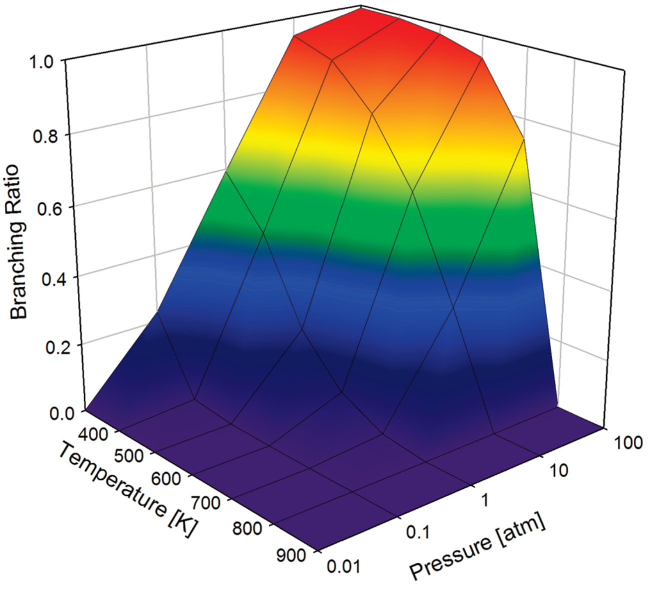


***Figure 8.*** Branching ratios to benzyl hydroperoxide as a function of temperature and pressure in the benzylperoxy + H reaction mechanism.

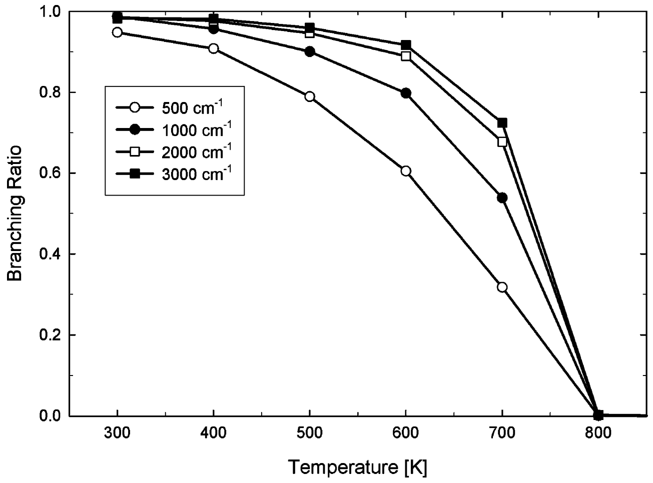


***Figure 9.*** Branching ratios to benzyl hydroperoxide at 10 atm, as a function of $\Delta E_{\mathrm{down}}$.

reaction not illustrated in Scheme 1 is that of benzylperoxy with OH, which should form benzoxyl + $HO_2$ (such reactions are known to be significant in alkyl radical oxidation kinetics).[22] Benzyl hydroperoxide will also react with OH via an addition mechanism. Ipso addition is expected to result in the formation of phenol plus the hydroperoxymethyl radical ($CH_2OOH$), which will dissociate to HCHO + OH. Free H atoms will also effect this addition/elimination sequence, resulting in benzene + HCHO + OH.

Most of the reactions included in Scheme 1 are now relatively well characterized, from both experiment and theory. While benzoxyl radical decomposition has recently been studied theoretically, further work is required to better understand the dissociation products of the highly activated benzoxyl radical that forms in the benzyl + O reaction (as well as the potential products of ring addition). Also, little information on the potentially important benzylperoxy + benzyl reaction is available (our thermochemistry predicts that this reaction will yield two benzoxyl radicals in a reaction that is 19.5 kcal $mol^{-1}$ exothermic).

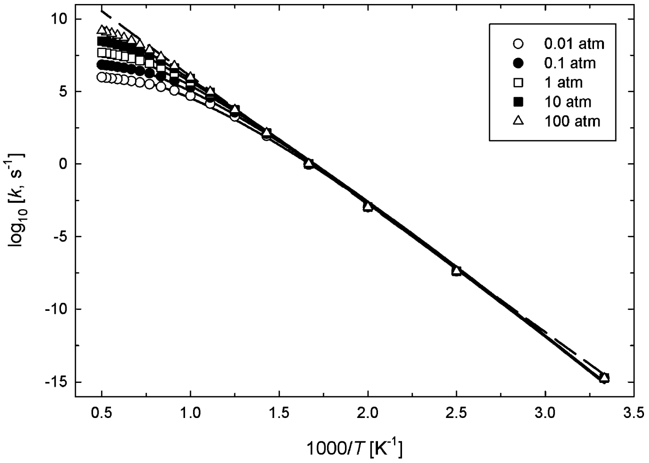


***Figure 10.*** Rate constants for benzyl hydroperoxide decomposition to benzoxyl + OH. Solid lines represent three-parameter Arrhenius fits, dashed line represents high-pressure limit.

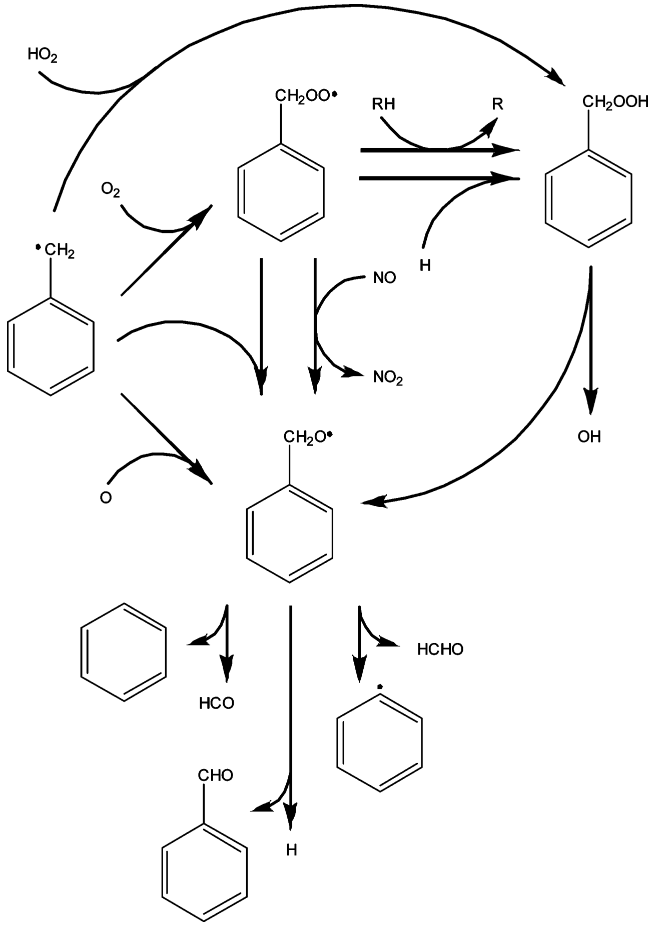
All of the reaction pathways illustrated in Scheme 1 ultimately produce the benzoxyl radical, highlighting the key role that this intermediate plays in aromatic oxidation chemistry. The further products of benzoxyl decomposition are benzaldehyde + H, benzene + HCO, and phenyl + HCHO, where branching among these three product sets is dependent on temperature and on the energy at which benzoxyl is formed. Benzyl hydroperoxide is also seen to play an important role in benzyl radical oxidation, both as an intermediate in the benzyl + $HO_2$ and benzylperoxy + H reactions and as a stable product from hydrogen abstraction by benzylperoxy (where RH is a hydrocarbon or $HO_2$). In combustion systems, benzyl hydroperoxide will decompose the benzoxyl + OH, but in the atmosphere, the further reactions of this species are less certain. Here, photolysis to benzoxyl + OH should be important, along with OH addition at the aromatic ring sites. Below, the potential role of benzyl hydroperoxide in the oxidation and ignition of toluene is explored in more detail.

The major reactions in the benzyl + $O_2$ + H reaction sequence are listed below, along with reaction enthalpies. Hydrogen atom addition to benzylperoxy produces the benzoxyl radical + OH in a considerably exothermic process. The benzoxyl radical is unstable, and is unlikely to exist for any significant lifetime in a thermal environment. Benzoxyl predominantly decomposes to benzaldehyde + H, although benzene + HCO and phenyl + HCHO are also important product sets, particularly at higher temperatures.[8] Irrespective, all three decomposition reactions are chain-propagating. When benzoxyl decomposes to benzaldehyde, the H atom initially consumed by benzylperoxy is regenerated. The overall reaction up to this point is benzyl + $O_2$ → benzaldehyde + OH ($\Delta H = -51.5$ kcal $mol^{-1}$). While this is a chain-propagating process and not directly the type of chain-branching reaction needed to initiate ignition, it does have the overall effect of converting the very unreactive benzyl radical into highly reactive OH, with a significant release of energy. Once OH is formed, it will readily abstract

***Table 4.*** Apparent Rate Parameters for Decomposition of Benzyl Hydroperoxide to Benzoxyl + OH, as a Function of Pressure

| | $A'$ ($s^{-1}$) | $n$ | $E_a$ (kcal $mol^{-1}$) |
|---|---|---|---|
| benzyl hydroperoxide → benzoxyl + OH (0.01 atm) | $8.33 \times 10^{28}$ | −4.39 | 45.32 |
| benzyl hydroperoxide → benzoxyl + OH (0.1 atm) | $1.56 \times 10^{38}$ | −7.33 | 48.16 |
| benzyl hydroperoxide → benzoxyl + OH (1 atm) | $2.03 \times 10^{47}$ | −10.27 | 50.71 |
| benzyl hydroperoxide → benzoxyl + OH (10 atm) | $1.39 \times 10^{55}$ | −12.87 | 52.60 |
| benzyl hydroperoxide → benzoxyl + OH (100 atm) | $1.36 \times 10^{61}$ | −14.95 | 53.65 |

***Scheme 1.*** Important Pathways in Benzyl Radical Oxidation



a H atom to form $H_2O$. For fuels like toluene (and other methylbenzenes), the predominant process will be to abstract a weak benzylic H atom (BDE of 86.7 kcal $mol^{-1}$, versus 118.8 kcal $mol^{-1}$ for HO−H). Including this reaction in our scheme, we arrive at an overall process in which toluene is oxidized by $O_2$ to benzaldehyde and water, along with a large production of energy. The exothermicity of this overall process aids ignition by increasing temperature, facilitating chain-branching decomposition reactions such as $H_2O_2$ → 2OH.

| | |
|---|---|
| $C_6H_5C{\bullet}H_2 + O_2 \leftrightarrow C_6H_5CH_2OO{\bullet}$ | $\Delta H = -22.9$ kcal $mol^{-1}$ |
| $C_6H_5CH_2OO{\bullet} + H{\bullet} \rightarrow C_6H_5CH_2O{\bullet} + O{\bullet}H$ | $\Delta H = -41.7$ kcal $mol^{-1}$ |
| $C_6H_5CH_2O{\bullet} \rightarrow C_6H_5CHO + H{\bullet}$ | $\Delta H = +12.7$ kcal $mol^{-1}$ |
| $C_6H_5CH_3 + O{\bullet}H \rightarrow C_6H_5C{\bullet}H_2 + H_2O$ | $\Delta H = -26.2$ kcal $mol^{-1}$ |
| $C_6H_5CH_3 + O_2 \rightarrow C_6H_5CHO + H_2O$ | $\Delta H = -78.0$ kcal $mol^{-1}$ |

The similar benzyl + $HO_2$ reaction process is known to play a key role in methylbenzene autoignition. In a recent experimental and kinetic modeling study on xylene autoignition, ignition delays were found to be most sensitive to the methylbenzyl + $HO_2$ association reactions.[23] The important reactions that we foresee taking place following the benzyl + $HO_2$ reaction are shown below. Benzyl reacts with $HO_2$ to form benzoxyl + OH in a mildly exothermic chain propagating reaction. Following decomposition of the benzyoxyl product, the free H atom can associate with $O_2$ to regenerate $HO_2$ (of course, other reactions are available), and if the reaction of toluene with OH is again included, we arrive at the same overall process as above.

| | |
|---|---|
| $C_6H_5C{\bullet}H_2 + HO_2{\bullet} \rightarrow C_6H_5CH_2O{\bullet} + O{\bullet}H$ | $\Delta H = -15.4$ kcal $mol^{-1}$ |
| $C_6H_5CH_2O{\bullet} \rightarrow C_6H_5CHO + H{\bullet}$ | $\Delta H = +12.7$ kcal $mol^{-1}$ |
| $H{\bullet} + O_2 \rightarrow HO_2{\bullet}$ | $\Delta H = -49.2$ kcal $mol^{-1}$ |
| $C_6H_5CH_3 + O{\bullet}H \rightarrow C_6H_5C{\bullet}H_2 + H_2O$ | $\Delta H = -26.2$ kcal $mol^{-1}$ |
| $C_6H_5CH_3 + O_2 \rightarrow C_6H_5CHO + H_2O$ | $\Delta H = -78.0$ kcal $mol^{-1}$ |

At moderate to high temperatures the H + $O_2$ reaction is effectively chain branching, producing OH + O in a mildly endothermic reaction. Under these conditions, the H atom formed in benzoxyl decomposition will produce OH and O, as shown below. The O atom formed in this scheme can further react with benzyl to yield benzaldehyde + H (among other products), and the benzyl + O → benzaldehyde + O + OH chain reaction will be a key process in toluene autoignition. We expect that aromatic ignition behavior will be highly sensitive to branching in the benzoxyl decomposition reaction between product sets such as benzaldehyde + H, phenyl + HCHO, and benzene + HCO.[8]

| | |
|---|---|
| $C_6H_5C{\bullet}H_2 + HO_2{\bullet} \rightarrow C_6H_5CH_2O{\bullet} + O{\bullet}H$ | $\Delta H = -15.4$ kcal $mol^{-1}$ |
| $C_6H_5CH_2O{\bullet} \rightarrow C_6H_5CHO + H{\bullet}$ | $\Delta H = +12.7$ kcal $mol^{-1}$ |
| $H{\bullet} + O_2 \rightarrow O{\bullet}H + O$ | $\Delta H = +16.4$ kcal $mol^{-1}$ |
| $C_6H_5C{\bullet}H_2 + HO_2{\bullet} + O_2 \rightarrow C_6H_5CHO + 2O{\bullet}H + O$ | $\Delta H = +13.7$ kcal $mol^{-1}$ |

In addition to reaction with H, the benzylperoxy radical can abstract a hydrogen atom to form stable benzyl hydroperoxide. The following scheme indicates the likely reactions that would occur in toluene combustion. Chain-branching decomposition of benzyl hydroperoxide follows, which we show to be rapid at temperatures above around 900 K, with subsequent pyrolysis of the benzoxyl radical to benzaldehyde + H. The overall process is now endothermic by 57.2 kcal $mol^{-1}$, but is highly chain branching with the formation of reactive OH radicals and O($^3$P) atoms.

| | |
|---|---|
| $C_6H_5C{\bullet}H_2 + O_2 \leftrightarrow C_6H_5CH_2OO{\bullet}$ | $\Delta H = -22.9$ kcal $mol^{-1}$ |
| $C_6H_5CH_2OO{\bullet} + C_6H_5CH_3 \rightarrow C_6H_5CH_2OOH + C_6H_5C{\bullet}H_2$ | $\Delta H = +6.0$ kcal $mol^{-1}$ |
| $C_6H_5CH_2OOH \rightarrow C_6H_5CH_2O{\bullet} + O{\bullet}H$ | $\Delta H = +45.0$ kcal $mol^{-1}$ |
| $C_6H_5CH_2O{\bullet} \rightarrow C_6H_5CHO + H{\bullet}$ | $\Delta H = +12.7$ kcal $mol^{-1}$ |
| $H{\bullet} + O_2 \rightarrow O{\bullet}H + O$ | $\Delta H = +16.4$ kcal $mol^{-1}$ |
| $C_6H_5CH_3 + 2O_2 \rightarrow C_6H_5CHO + 2O{\bullet}H + O$ | $\Delta H = +57.2$ kcal $mol^{-1}$ |

From the above reaction schemes we find that oxidation reactions proceeding via benzyl hydroperoxide aid in toluene autoignition, through exothermic and chain-branching processes. The benzyl radical can react with $HO_2$, or with $O_2$ and then H, in processes that ultimately result in the exothermic oxidation of toluene to benzaldehyde + $H_2O$. In order for these processes to take place, preliminary reactions producing benzyl, H, and $HO_2$ are required. Hydrogen abstraction from toluene by $O_2$ can also form $HO_2$ (+ benzyl), while toluene will react with most radicals to produce benzyl. At higher temperatures, toluene will pyrolyse to benzyl + H. The benzyl radical produced in any of the above processes will react with $O_2$ and then abstract a hydrogen atom to produce benzyl hydroperoxide, which decomposes in a chain-branching reaction at even low temperatures. Current kinetic models for the oxidation of toluene and other alkylated aromatics will be improved through inclusion of the reactions considered here.

**Supporting Information Available:** Cartesian coordinates and vibrational frequencies for benzylperoxy, benzyl hydroperoxide, and **TS1**−**TS4**; canonical rate constants for the benzylperoxy + H reaction as a function of transition state structure and temperature. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Li, Y.; Zhang, L.; Tian, Z.; Yuan, T.; Wang, J.; Yang, B.; Qi, F. *Energy Fuels* **2009**, *23*, 1473–1485.

(2) (a) Müller-Markgraf, M.; Troe, J. *J. Phys. Chem.* **1988**, *92*, 4899. (b) Jones, J.; Bacskay, G. B.; Mackie, J. C. *J. Phys. Chem. A* **1997**, *101*, 7105. (c) Oehlschlaeger, M. A.; Davidson, D. F.; Hanson, R. K. *J. Phys. Chem. A* **2006**, *110*, 6649. (d) da Silva, G.; Cole, J. A.; Bozzelli, J. W. *J. Phys. Chem. A* **2009**, *113*, 6111.

(3) (a) Elmaimouni, L.; Minetti, R.; Sawersyn, J. P.; Devolder, P. *Int. J. Chem. Kinet.* **1993**, *25*, 399. (b) Fenter, F. F.; Nozière, B.; Caralp, F.; Lesclaux, R. *Int. J. Chem. Kinet.* **1994**, *26*, 171. (c) Murakami, Y.; Oguchi, T.; Hashimoto, K.; Nosaka, Y. *J. Phys. Chem. A* **2007**, *111*, 13200.

(4) Hippler, H.; Reihs, C.; Troe, J. *Proc. Combust. Inst.* **1990**, *23*, 37.

(5) (a) Brezinsky, K.; Litzinger, T. A.; Glassman, I. *Int. J. Chem. Kinet.* **1984**, *16*, 1053. (b) Bartels, M.; Edelbuttel-Einhaus, J.; Hoyermann, K. *Proc. Combust. Inst.* **1989**, *22*, 1041. (c) Hoyerman, K.; Seeba, J.; Olzmann, M.; Viskolcz, B. *Ber. Bunsenges. Phys. Chem.* **1997**, *101*, 538.

(6) da Silva, G.; Bozzelli, J. W. *Proc. Combust. Inst.* **2009**, *32*, 287.

(7) (a) Nozière, B.; Lesclaux, R.; Hurley, M. D.; Dearth, M. A.; Wallington, T. J. *J. Phys. Chem.* **1994**, *98*, 2864. (b) El Dib, G.; Chakir, A.; Roth, E.; Brion, J.; Daumont, D. *J. Phys. Chem. A* **2006**, *110*, 7848. (c) Du, B.; Zhang, W.; Mu, L.; Feng, C.; Qin, Z. *Chem. Phys. Lett.* **2007**, *445*, 17.

(8) da Silva, G.; Bozzelli, J. W. *J. Phys. Chem. A* **2009**, *113*, 6979.

(9) Baboul, A. G.; Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **1999**, *110*, 7650.

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

(11) Chase, M. W., Jr. *J. Phys. Chem. Ref. Data, Monogr. 9* **1998**, 1.

(12) da Silva, G.; Moore, E. E.; Bozzelli, J. W. *J. Phys. Chem. A* **2009**, *113*, 10264.

(13) da Silva, G.; Bozzelli, J. W. *J. Phys. Chem. A* **2008**, *112*, 3566.

(14) Mokrushin, V.; Bedanov, V.; Tsang, W.; Zachariah, M.; Knyazev, V. *ChemRate, Version 1.5.2*; National Institute of Standards and Testing: Gaithersburg, MD, 2006.

(15) Ruscic, B.; Pinzon, R. E.; Morton, M. L.; Srinivasan, N. K.; Su, M.-C.; Sutherland, J. W.; Michael, J. V. *J. Phys. Chem. A* **2006**, *110*, 6592.

(16) Cox, J. D.; Wagman, D. D.; Medvedev, V. A. *CODATA Key Values for Thermodynamics*; Hemisphere Publishing Corp.: New York, 1984.

(17) da Silva, G.; Bozzelli, J. W.; Sebbar, N.; Bockhorn, N. *ChemPhysChem* **2006**, *7*, 1119.

(18) Ruscic, B.; Wagner, A. F.; Harding, L. B.; Asher, R. L.; Feller, D.; Dixon, D. A.; Peterson, K. A.; Song, Y.; Qian, X.; Ng, C.-Y.; Liu, J.; Chen, W.; Schwenke, D. W. *J. Phys. Chem. A* **2002**, *106*, 2727.

(19) da Silva, G.; Bozzelli, J. W. *J. Phys. Chem. A* **2009**, *113*, 8971.

(20) da Silva, G.; Bozzelli, J. W.; Liang, L.; Farrell, J. T. *J. Phys. Chem. A* **2009**, *113*, 8923.

(21) da Silva, G. *Chem. Phys. Lett.* **2009**, *474*, 13.

(22) Petway, S. V.; Ismail, H.; Green, W. H.; Estupinan, E. G.; Jusinski, L. E.; Taatjes, C. A. *J. Phys. Chem. A* **2007**, *111*, 3891.

(23) Shen, H.-P. S.; Oehlschlaeger, M. A. *Combust. Flame* **2009**, *156*, 1053.

CT900352F

JCTC Journal of Chemical Theory and Computation

# Martini Coarse-Grained Force Field: Extension to Carbohydrates

Cesar A. López,[†] Andrzej J. Rzepiela,[†] Alex H. de Vries,[†] Lubbert Dijkhuizen,[‡] Philippe H. Hünenberger,[§] and Siewert J. Marrink*[,†]

*Groningen Biomolecular Sciences and Biotechnology Institute & Zernike Institute for Advanced Materials, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands, Groningen Biomolecular Sciences and Biotechnology Institute (GBB), Centre for Carbohydrate Bioprocessing, University of Groningen, Kerklaan 30, 9751 NN Haren, The Netherlands, and Laboratorium für Physikalische Chemie, ETH Zürich, CH-8093 Zürich, Switzerland*

Received June 21, 2009

**Abstract:** We present an extension of the Martini coarse-grained force field to carbohydrates. The parametrization follows the same philosophy as was used previously for lipids and proteins, focusing on the reproduction of partitioning free energies of small compounds between polar and nonpolar phases. The carbohydrate building blocks considered are the monosaccharides glucose and fructose and the disaccharides sucrose, trehalose, maltose, cellobiose, nigerose, laminarabiose, kojibiose, and sophorose. Bonded parameters for these saccharides are optimized by comparison to conformations sampled with an atomistic force field, in particular with respect to the representation of the most populated rotameric state for the glycosidic bond. Application of the new coarse-grained carbohydrate model to the oligosaccharides amylose and Curdlan shows a preservation of the main structural properties with 3 orders of magnitude more efficient sampling than the atomistic counterpart. Finally, we investigate the cryo- and anhydro-protective effect of glucose and trehalose on a lipid bilayer and find a strong decrease of the melting temperature, in good agreement with both experimental findings and atomistic simulation studies.

## 1. Introduction

Carbohydrates (saccharides), the most abundant product of photosynthesis, play an important role in the energetic metabolism of living species and the signaling and immunological responses and are a fundamental component of the external cell wall of many organisms.[1] In addition, saccharides are present in a variety of emerging classes of biomimetic materials.[2] Furthermore, due to their cryo- and anhydro-protective properties, many sugars have been shown to be effective stabilizers of biological components, such as proteins and membranes, in the low-temperature or dehydrated states.[3–5] This class of compounds encompasses a huge variety of possible monomeric units (differing in stereochemistry and functionalization) that can be connected in chains presenting a virtually infinite number of possible residue sequences, linkage types, and degrees of branching. Despite their importance, the experimental characterization of the structural and dynamical properties of oligosaccharides in general has proven rather problematic. Unlike proteins, nucleic acids, and lipids, which tend to predominantly adopt a well-defined (native) conformation under the conditions where they are biologically functional, carbohydrates are typically associated with a high extent of conformational heterogenity. As a result of this structural diversity and conformational heterogenity, carbohydrates arguably represent the most challenging class of biomolecules in terms of experimental characterization and elucidation of structure−function relationships.[6–8] Static structures of carbohydrates may often be obtained from (X-ray) crystallography (of crystals or fibers), but it is always uncertain whether these

* Corresponding author e-mail: s.j.marrink@rug.nl.

[†] Zernike Institute for Advanced Materials.

[‡] Microbiology Department, Haren.

[§] ETH Zürich.

10.1021/ct900313w CCC: $40.75 © 2009 American Chemical Society
Published on Web 10/30/2009

molecules adopt similar conformations in solution as in the solid state. On the other hand, nuclear magnetic resonance (NMR) spectroscopy provides information about carbohydrates in solution but only in the form of averages over all the populated conformational states present in solution (i.e., over all the molecules or molecular segments in the sample as well as over the time scale of the NMR experiment). Many other experimental techniques (e.g., electron microscopy, light or neutron diffraction, circular dichroism, infrared spectroscopy, or rheology) provide useful but even more indirect information about carbohydrate conformations.

Molecular dynamics (MD) simulations can, in principle, provide the link between structure and physical properties that are more readily measured (i.e., radius of gyration, gel transition temperature; mostly concerning long polymers[1,9–13]). Many force fields have been extensively parametrized for carbohydrates[14–19] and have been used to provide details of the structure and dynamics at an all-atom (AA) level, for example, to explore the ring puckering of glucose,[20–22] conformational changes in disaccharides and trisaccharides,[14,23–25] and stability of oligosaccharides like amylose and Curdlan.[9,26–28] However, such studies are necessarily limited to small system sizes with a limited sampling of the potentially very large conformational space. Simulations of longer oligosaccharides or the association of these in colloidal systems are very challenging at the AA level.

An alternative to the AA approach is the use of coarse-grained (CG) force fields, which provide a useful methodology to study large systems on a long time scale at reasonable computational cost. CG models can capture the most fundamental physical and chemical properties after averaging out some of the atomistic information, both spatially and temporally. A large diversity of CG approaches for biomolecular systems is available. They range from qualitative, solvent-free models to models including chemical specificity.[29] Most of the effort has been canalized into the development of models for the simulation of proteins and lipids. The design of reliable coarse-grained models for carbohydrates is hindered by their high structural diversity and the limited amount of experimental data available. Pioneering efforts in the context of glucose-based carbohydrates have been undertaken by Liu et al.[30] and Molinero et al.,[31] in which the glucose ring is represented by three particle types and its behavior in water is optimized with respect to simulations at the AA level. Bonded interaction potentials are obtained by Boltzmann inversion of the distributions of the bonds, angles, and torsion from atomistic simulations of glucose. Amylose chains have been used as a test model in both cases, revealing excellent agreement with experimental data. Another possible approach is the one adopted by Bathe et al.,[32] in which explicit atom models of isolated disaccharides are used to generate pretabulated potentials of mean force for the glycosidic torsions. In addition, electrostatic and steric interactions between nonadjacent residues were included, making use of virtual sites. The model was optimized for application to glycosaminoglycans. In spite of some promising results, these approaches cannot be easily extended to other systems without a full reparameterization or be used in combination with other (bio)molecules. A more general force field for CG simulations has been developed by one of us,[33–35] coined the Martini force field. It is based on a four-to-one mapping scheme, implying that on average four heavy atoms and associated hydrogens are represented as a single CG site. The Martini model has been parametrized extensively by using a chemical building block principle. Its key feature is the reproduction of thermodynamic data, especially the partitioning of the building blocks between polar and nonpolar phases. It has been successfully applied to a range of lipid and protein systems.[36–38]

In this work, we extend the Martini force field to include carbohydrates. We base our parametrization on the conformational sampling of small carbohydrates with our CG model in comparison to AA simulations. In addition, octanol/water partitioning free energies are calculated to select the appropriate CG particle types. The set of carbohydrates used in the parametrization is illustrated in Figure 1 and comprises the monosaccharides glucose (G) and fructose (F) and the disaccharides sucrose (SUC), trehalose (T), maltose (M), cellobiose (C), kojibiose (K), sophorose (S), laminarabiose (L), and nigerose (N). The group of disaccharides includes the most important sugar−sugar linkages (1−1, 1−2, 1−3, 1−4) except for the 1−6 linkage, which proved difficult to model at the CG level. To test the transferability of the parameters to oligosaccharides, simulations of two different oligomers (amylose and Curdlan) are presented and compared with their AA counterpart. Finally, the compatibility of the carbohydrate parameters with the lipid parameters in the Martini force field is tested by looking at the stability of the liquid-crystalline phase of a dipalmitoyl-phosphatidylcholine (DPPC) bilayer in the presence of glucose and trehalose solutions.

At this point it is important to stress the inherent limitations of our CG carbohydrate model. First, the level of coarsening does not allow distinction between different ring conformations; the model represents the most populated chair $^4C_1$ puckering state. Second, no distinction is made between different anomers at the level of a reducing residue. Consequently, $\alpha$- and $\beta$-anomers are represented by the same topology. Note, however, that for the glycosidic linkage between sugar units, the $\alpha$- and $\beta$-linkages are distinguished at the CG level through the use of different angle and dihedral interaction potentials. Third, in its current state, the model can only represent a single conformation (denoted 'syn') for the glycosidic linkage. The more flexible 1−6 linkage is not considered at present. Finally, the model is aimed at simulations of saccharides in solution, not in a crystal state. For a more elaborate discussion of the scope and limitations of the model, we refer to the last section of the manuscript.

The rest of this article is organized as follows. The methods section is devoted to giving a detailed account of the parametrization procedure. It is followed by the results section, reporting the results obtained for mono-, di-, and oligosaccharides. A conclusive section, with limitations and outlook, ends this article.

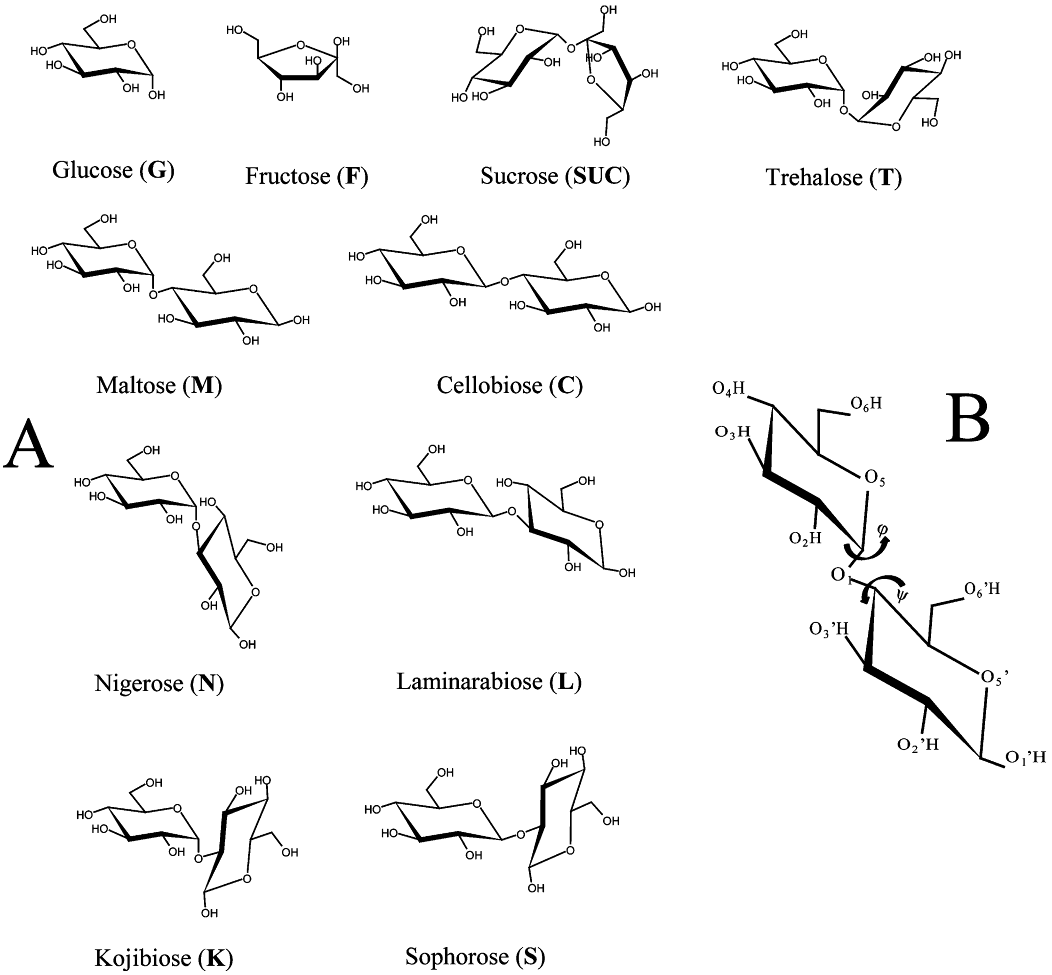


*Figure 1.* (A) Saccharides considered in this work: glucose (Glcα (G)), fructose (Fruβ (F)), maltose (M; Glcα(1−4)Glcβ), cellobiose (C; Glcβ(1−4)Glcβ), kojibiose (K; Glcα(1−2)Glcβ), sophorose (S; Glcβ(1−2)Glcβ), nigerose (N; Glcα(1−3)Glcβ), laminarabiose (L; Glcβ(1−3)Glcβ), sucrose (SUC; Glcα(1−2)Fruβ), and trehalose (T; Glcα(1−1)αGlc). (B) The definitions of the dihedral angles $\varphi$ ($O_5$−$C_1$−$O_1$−$C_n'$) and $\psi$ ($C_1$−$O_1$−$C_n'$-$C_{n-1}'$) are illustrated for maltose.

## 2. Computational Methods

**2.1. Model.** The Martini CG model is used for the basic parametrization of the carbohydrate force field, which is therefore fully compatible with the Martini lipid[35] and protein[34] models. In this section we provide a brief overview of the basic parametrization procedure followed for carbohydrates: definition of the mapping and parametrization of nonbonded and bonded interactions. More details about the basic Martini model can be found in the original articles.[34,35]

*2.1.1. Mapping of CG Sugars.* According to the mapping procedure for the Martini force field, on average four heavy particles are represented by one CG site. For a single sugar ring, consisting of 12 atoms (hydrogen atoms not counted), three particles are therefore required. This level of resolution preserves the geometrical shape of the rings (Figure 2A) and allows for a distinction between different types of monosaccharides through variations in the bond lengths, angles, and CG particle types. Disaccharides are modeled as two three-bead units connected by a single bond, which mimics the glycosidic linkage (Figure 2B). This geometry allows for the definition (and subsequent parametrization) of the glycosidic dihedral angles $\varphi$ and $\psi$ which determine the relative orientation of the two sugar residues and the flexibility of the linkage. The set of fine-grained particles represented by the CG beads is chosen to be different for a monosaccharide and for the individual residues in a disaccharide. This somewhat nonobvious choice confers to the model the ability to represent the typical polar/apolar character of the disaccharides with the apolar part corresponding to the central part along the glycosidic linkage. Oligosaccharides are constructed by connecting disaccharide residues through additional bonds (Figure 2C).

*2.1.2. Parameterization of Nonbonded Interactions.* Nonbonded interactions are described by a Lennard−Jones (LJ) 12−6 potential energy function

$$U_{\mathrm{LJ}}(r) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right] \qquad (1)$$

with $\sigma_{ij}$ representing the distance at zero energy (collision diameter) between two particles $i$ and $j$ and $\varepsilon_{ij}$ the strength of their interaction. The other nonbonded component of the Martini force field, the Coulomb interaction between charged particles, is not relevant for the saccharides considered in this work. The Martini model considers two different particles sizes: normal types and ring particle types, which differ in the $\sigma_{ii}$ value of 0.47 and 0.43 nm, respectively. The

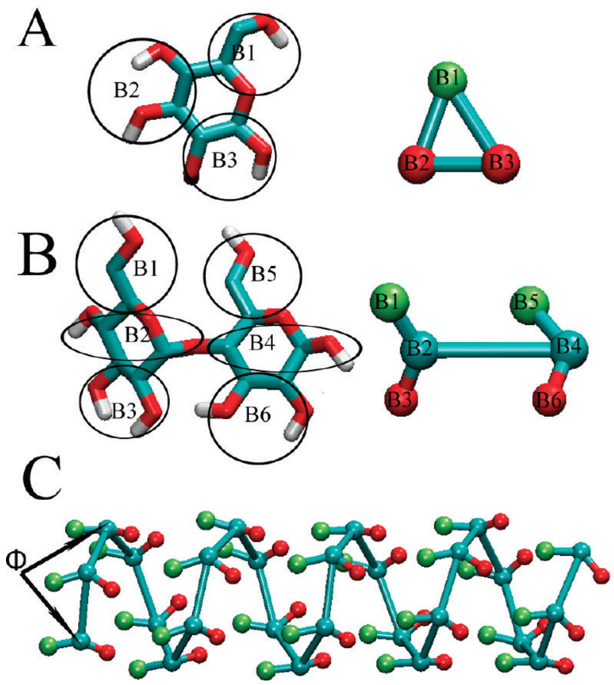


***Figure 2.*** Coarse-grained mapping for mono-, di-, and oligosaccharides. (A) Monosaccharides, like glucose, are represented by three beads B1−B3. (B) Disaccharides, such as trehalose, are composed of six beads B1−B6, with the bond between B2 and B4 representing the glycosidic bond. (C) Oligosaccharides are based on the disaccharide topology but using an extra angle potential (Φ) for three consecutive backbone beads. Different colors are used to indicate different levels of polarity of the CG beads, see Table 3.

strength of the pairwise particle−particle interaction is determined by the value of the LJ parameter $\varepsilon_{ij}$. Larger values (i.e., stronger attraction) mimic polar interactions, whereas smaller values (weaker attraction) are used to mimic the hydrophobic effect. In the full interaction matrix, four main types of interaction sites are differentiated: polar (P), nonpolar (N), apolar (C), and charged (Q). The special class of ring-type particles is further denoted by the letter “S” and has a reduced value of $\varepsilon_{ii}$. Within a main type, subtypes are distinguished either by a letter denoting the hydrogen-bonding capabilities (d = donor, a = acceptor, da = both, 0 = none) or by a number indicating the degree of polarity (from 1 = low polarity to 5 = high polarity). Each of these particle types is representative of a specific chemical building block, i.e., inferred from a class of small compounds with similar chemical properties. The Martini force field has been parametrized extensively to reproduce the correct partitioning free energies of small molecules between a diversity of polar and apolar solvents. The full interaction matrix $\varepsilon_{ij}$ can be found in the original publication.[35]

For the parametrization of the saccharides, the chemical nature of the underlying fine-grained structure is used to select the most appropriate building block and corresponding particle types. The division of the saccharides into building blocks can be done in multiple ways, however, leaving some room for adjustment. Therefore, the partitioning free energy of monosaccharides and disaccharides between water and octanol has been computed to fine tune the appropriate particle-type selection for the nonbonded interactions. Another alternative concerns the use of the normal particle type versus the ring particle type. On first thought, the ring particle type might seem more appropriate to model the sugar rings. However, the class of ring particle types has been parametrized based on unsubstituted ring compounds such as cyclohexane and benzene for which a four-to-one mapping is inadequate. In contrast, in the case of carbohydrate rings, the standard four-to-one mapping scheme applies (cf. Figure 2A). The properties of monosaccharide solutions over a large concentration range were used to test the two possible models. Note that the partioning free energies do not depend on this choice, which only affects sugar−sugar interactions.

*2.1.3. Parameterization of Bonded Interactions.* Three types of bonded interactions are considered for the carbohydrates. CG particles chemically connected are described by a harmonic potential $V_{\mathrm{bond}}(R)$

$$V_{\mathrm{bond}}(R) = \frac{1}{2}K_{\mathrm{bond}}(R - R_{\mathrm{bond}})^2 \tag{2}$$

with equilibrium distance $R_{\mathrm{bond}}$ and force constant $K_{\mathrm{bond}}$. LJ interactions between bonded neighbors are excluded. Since the degrees of freedom are reduced at the coarse-grained level, it is necessary to preserve the topology of differently linked sugars by using both angle and dihedral potentials. A cosine-harmonic potential $V_{\mathrm{angle}}(\theta)$ is used for the angles

$$V_{\mathrm{angle}}(\theta) = \frac{1}{2}K_{\mathrm{angle}}[\cos(\theta) - \cos(\theta_0)]^2 \tag{3}$$

where $K_{\mathrm{angle}}$ and $\theta_0$ are the force constant and equilibrium angle, respectively. For the dihedrals, a proper dihedral potential $V_{\mathrm{pd}}(\phi)$ is used with a multiplicity of 1

$$V_{\mathrm{pd}}(\phi) = K_{\mathrm{pd}}[1 + \cos(\phi - \phi_{\mathrm{pd}})] \tag{4}$$

In this case, $\phi$ denotes the angle between planes containing the beads $i$, $j$, $k$ and $j$, $k$, $l$, respectively, with equilibrium angle $\phi_{\mathrm{pd}}$ and force constant $K_{\mathrm{pd}}$.

The set of bonded parameters featured in eqs 2−4 has been parametrized by comparison to simulations of sugars at the AA level. To this end, the AA trajectories were converted to pseudo-CG trajectories using the center of mass of the appropriate fine-grained particles[39]

$$\mathbf{r}_i^{\mathrm{CG}} = \frac{\sum_{j=1}^{p} \mathbf{r}_j.m_j}{\sum_{j=1}^{p} m_j} \tag{5}$$

The vector $\mathbf{r}_i^{\mathrm{CG}}$ describes the position of the pseudo-CG bead, $p$ is the number of atoms mapped to a given coarse bead, $m_j$ is the mass of the atom $j$, and $\mathbf{r}_j$ is its coordinates. The mapping between the AA and CG representation is shown in Figure 2. From the AA trajectory the target distribution functions were obtained for the various bonds, angles, and dihedrals considered. In a couple of iterative steps, the CG parameters were adjusted manually to obtain as close a match as possible between the pseudo-CG and real CG distributions.

**2.2. Simulation Details.** *2.2.1. System Setup.* The following saccharides were modeled (cf. Figure 1). Monosaccharides: glucose (G; α-D-glucopyranose) and fructose (F; β-D-fructofuranose). Disaccharides: maltose (M; Glcα(1−4)-Glcβ), cellobiose (C; Glcβ(1−4)Glcβ), kojibiose (K; Glcα(1−2)Glcβ), sophorose (S; Glcβ(1−2)Glcβ), nigerose (N; Glcα(1−3)Glcβ), laminarabiose (L; Glcβ(1−3)Glcβ), sucrose (SUC; Glcα(1−2)Fruβ), trehalose (T; Glcα(1−1)αGlc). Oligosaccharides: maltoheptaose (7 D-glucopyranose monomers in α 1−4 linkage), amylose (26 D-glucopyranose monomers in α 1−4 linkage), laminaraheptabiose (7 D-glucopyranose monomers in β 1−3 linkage), and Curdlan (3 chains of 26 D-glucopyranose monomers in β 1−3 linkage).

For each of these sugars both AA and CG simulations were performed. The monosaccharide and disaccharide systems, used for the parametrization, consisted of a single sugar molecule either in pure water or in water-saturated octanol. The water-saturated octanol consists of a 0.255 water/octanol molar fraction.[40] The oligosaccharides were simulated either in water or in nonane. An additional set of CG simulations was performed in which the concentration of glucose was increased systematically up to supersaturated solutions (60 wt %). A pure glucose system was also simulated. The cryo- and anhydro-protection effect of sugars was investigated by simulating a pre-equilibrated DPPC bilayer consisting of 64 lipids per leaflet either in pure water or in saccharide solutions. Two sugars were used for this investigation: glucose at 4 M (664 sugars and 2166 CG water particles) and trehalose at 2 M (332 sugars and 2166 CG water particles) concentration. Moreover, a control system was used in which the membrane was completely hydrated (2166 CG waters). All simulations were performed using the Gromacs package, version 3.3.1.[41] Table 1 provides the complete list of the systems simulated in this work, including details about compositions and total simulation times. The details of the simulation parameters are given below, including a description of the method by which the partition coefficients were computed.

*2.2.2. Coarse-Grained Simulation Parameters.* In the simulations at the coarse-grained level, we followed the standard simulation protocol used in the Martini parametrization.[35] The nonbonded interactions are cut off at a distance $r_{\text{cut}}$ of 1.2 nm. To reduce generation of unwanted noise, the standard shift function of Gromacs[41] is used in which both the energy and the force smoothly vanish at the cutoff distance. The LJ potential is shifted from $r = 0.9$ nm to the cutoff distance. The time step used to integrate the equations of motion is 20 fs. Note that here and throughout the entire manuscript actual simulation time is reported (i.e., no scaling of the time axis has been applied to provide an effective time scale). The temperature is maintained at 310 K by weak coupling of the solvent and solute separately to a Berendsen heat bath[42] with a relaxation time of 1 ps. The pressure is maintained at 1.0 bar by weak coupling to a pressure bath via isotropic coordinate scaling with a relaxation time of 5 ps. Simulations of bilayers are performed at three different temperatures (270, 325, and 475 K) with a semi-isotropic coupling of the lateral and perpendicular box dimensions to a pressure of 1.0 bar. The topology and parameters for water and octanol are taken from the Martini force field[35] data set. In the case of nonane, the molecule is represented using two C1 particles similar to octane.

*2.2.3. All-Atom Simulation Parameters.* The AA simulations of hexopyranoses were performed using the latest Gromos force field parameters set for carbohydrates[19] (note that although the Gromos force field is a united-atom force field, it will be referred to as AA). In the case of furanose (i.e., in fructose and sucrose), parameters were adapted from the hexopyranose force field set (see Supporting Information). Verification of the furanose parameters was done by comparison to results from previous simulations[43,44] (Figures S1 and S2, Supporting Information) obtained with different force fields. For each simulation, the solute was placed in the center of a periodic cubic box with minimum wall−solute distances of 2 nm. The SPC water model[45] was used to solvate the system. For the simulations in nonane and octanol, the procedure to set up the systems was similar to that in the aqueous environment. The parameters for aliphatic hydrocarbons were taken from the Gromos 53a6 force field.[46] A steepest descent algorithm[41] was used to relax the internal interactions in vacuum. After that, the box was filled with the respective solvent and the minimization procedure was repeated. In all cases, a 2 fs time step was used to integrate Newton's equations of motion. The LINCS algorithm[47] was applied to constrain all bond lengths with a relative geometric tolerance of $10^{-4}$. Before production time, the systems were pre-equilibrated by slow heating up to 310 K. The temperature was maintained at 310 K by weak coupling of the solvent and solute separately to a Berendsen heat bath[42] with a relaxation time of 0.1 ps. Pressure coupling was maintained at 1.0 bar using an NPT ensamble by weak coupling via isotropic coordinate scaling with a relaxation time of 1 ps. Nonbonded interactions were handled using a twin-range cutoff[48] scheme. Within a short-range cutoff of 0.9 nm, the interactions were evaluated every time step based on a pair list recalculated every 5 time steps. The intermediate-range interactions up to a long-range cutoff radius of 1.4 nm were evaluated simultaneously with each pair list update and assumed constant in between. To account for electrostatic interactions beyond the long-range cutoff radius, a reaction field approach[49] was used with a relative dielectric permittivity of 66 for water, 2 for nonane, and 10.3 for octanol. Analysis of the dihedral distributions of the various disaccharides in an aqueous environment showed a good agreement with the ones obtained by Pereira et al.[25] using the same force field. Mapping of the AA trajectories to pseudo-CG trajectories was performed at a frequency of once per 40 ps. Table 1 provides a complete overview of the system composition and total simulation time for each of the simulations performed.

*2.2.4. Partitioning Free Energies.* In order to compute octanol/water partition coefficients $P_{\text{OW}}$ directly, the free energies of solvation of the sugar compounds were calculated in both aqueous and organic phases. Given the appropriate free energies of solvation, computation of the partition coefficient is straightforward. The difference between the solvation free energy in the aqueous ($\Delta G^{\text{W}}$) and organic phase ($\Delta G^{\text{O}}$) is the partitioning free energy ($\Delta\Delta G_{\text{OW}}$) of the carbohydrate between water-saturated octanol solution and water:

$$\Delta\Delta G_{\mathrm{OW}} = -2.3RT \log P_{\mathrm{OW}} \tag{6}$$

$\Delta G^{\mathrm{W}}$ and $\Delta G^{\mathrm{O}}$ were calculated as the free energy difference $\Delta F$ of the solute in vacuum (state A) and in the condensed phase (state B) using the thermodynamic integration (TI) procedure:[50]

$$\Delta F_{\mathrm{BA}} = F_{\mathrm{B}} - F_{\mathrm{A}} = \int_{\lambda_{\mathrm{A}}}^{\lambda_{\mathrm{B}}} \mathrm{d}\lambda \left\langle \frac{\partial U_{\mathrm{uv}}(\lambda)}{\partial \lambda} \right\rangle_{\lambda} \tag{7}$$

Here $U_{\mathrm{uv}}(\lambda)$ denotes the potential energy function describing the total solute−solvent interaction, the average ⟨...⟩ is taken over the MD trajectory, and $\lambda$ is a coupling parameter that regulates the strength of $U_{\mathrm{uv}}$ and varies linearly from zero ($\lambda_{\mathrm{A}} = 0$) to full ($\lambda_{\mathrm{B}} = 1$) interaction. All bonded interactions were interpolated linearly; on the other hand, to remove the singularities in the potentials for the nonbonded interactions a soft-core approach was used.[51] Calculations were performed at 25 intermediate $\lambda$ values until a smooth curve for the free energy derivative was obtained, which was then integrated numerically (by trapezoidal integration). For each individual $\lambda$ point, at least 10 (CG) or 6 ns (AA) was used for the analysis. Additional $\lambda$ points, especially in high-curvature regions, were required for the disaccharides at the AA level. Simulations in vacuum were performed using a stochastic dynamics approach with the same number of $\lambda$ points as used for the water and octanol systems. To estimate the error in the free energy calculation, each $\lambda$ set

***Table 1.*** Summary of the Simulations Performed in This Work

| | composition | | | | | |
|---|---|---|---|---|---|---|
| | water | carbohydrate | nonane | octanol | temp. (K) | time (ns) |
| (A) all-atom (AA) | | | | | | |
| *mapping procedure* | | | | | | |
| monosaccharides | 876 | 1 | | | 310 | 200 |
| disaccharides | 876 | 1 | | | 310 | 200 |
| *log P oct/water* | | | | | | |
| monosaccharides in vacuum | | 1 | | | 310 | 25 × 6[a] |
| monosaccharides in water | 1400 | 1 | | | 310 | 25 × 6 |
| monosaccharides in octanol | 66 | 1 | | 199 | 310 | 25 × 6 |
| disaccharides in vacuum | | 1 | | | 310 | 25 × 6 |
| disaccharides in water | 1400 | 1 | | | 310 | 25 × 6 |
| disaccharides in octanol | 66 | 1 | | 199 | 310 | 25 × 6 |
| *oligosaccharides* | | | | | | |
| maltoheptaose in water | 7106 | 1 | | | 310 | 40 |
| maltoheptaose in nonane | | 1 | 400 | | 310 | 40 |
| amylose in water | 21 223 | 1 | | | 310 | 40 |
| amylose in nonane | | 1 | 1825 | | 310 | 40 |
| laminaraheptabiose in water | 7106 | 1 | | | 310 | 40 |
| laminaraheptabiose in nonane | | 1 | 400 | | 310 | 40 |
| Curdlan in water | 21 223 | 1 | | | 310 | 40 |
| Curdlan in nonane | | 1 | 1077 | | 310 | 40 |
| *(B) coarse-grained (CG)* | | | | | | |
| *mapping procedure* | | | | | | |
| monosaccharides | 616 | 1 | | | 310 | 200 |
| disaccharides | 616 | 1 | | | 310 | 200 |
| *log P oct/water* | | | | | | |
| monosaccharides in vacuum | | 1 | | | 310 | 25 × 10 |
| monosaccharides in water | 1000 | 1 | | | 310 | 25 × 10 |
| monosaccharides in octanol | 43 | 1 | | 519 | 310 | 25 × 10 |
| disaccharides in vacuum | | 1 | | | 310 | 25 × 10 |
| disaccharides in water | 1000 | 1 | | | 310 | 25 × 10 |
| disaccharides in octanol | 43 | 1 | | 519 | 310 | 25 × 10 |
| *density profile* | | | | | | |
| glucose in solution | 0−100% w/w | 0−100% w/w | | | 310 | 50 × 50[b] |
| *oligosaccharides* | | | | | | |
| maltoheptaose in water | 905 | 1 | | | 310 | 40 |
| maltoheptaose in nonane | | 1 | 500 | | 310 | 40 |
| amylose in water | 5137 | 1 | | | 310 | 40 |
| amylose in nonane | | 1 | 2000 | | 310 | 40 |
| laminaraheptabiose in water | 905 | 1 | | | 310 | 40 |
| laminaraheptabiose in nonane | | 1 | 500 | | 310 | 40 |
| curdlan in water | 5137 | 1 | | | 310 | 40 |
| curdlan in nonane | | 1 | 2000 | | 310 | 40 |
| *cryo-protection effect* | | | | | | |
| 128 DPPC + 4 M glucose | 2166 | 664 | | | 270/325/475 | 500 |
| 128 DPPC + 2 M trehalose | 2166 | 332 | | | 270/325/475 | 500 |
| 128 DPPC | 2166 | | | | 270/325/475 | 500 |

[a] Multiple simulations at different $\lambda$ points were reported; see methods. [b] Fifty different sugar concentrations were used, simulated for 50 ns at each concentration level.

***Table 2.*** Force Field Parameters CG Carbohydrates[a]

| sugar | bond | $R_{bond}$ (nm) | $K_{bond}$ (kJ mol$^{-1}$nm$^{-2}$) | angles | $\theta_0$ (deg) | $K_{angle}$ (kJ mol$^{-1}$) | dihedrals | $\Phi_{pd}$ | $K_{pd}$ (kJ mol$^{-1}$) |
|---|---|---|---|---|---|---|---|---|---|
| glucose (G) | B1−B2 | 0.375 | 35 000 | | | | | | |
| | B1−B3 | 0.331 | 35 000 | | | | | | |
| | B2−B3 | 0.322 | 50 000 | | | | | | |
| fructose (F) | B1−B2 | 0.309 | 10 000 | | | | | | |
| | B1−B3 | 0.303 | 35 000 | | | | | | |
| | B2−B3 | 0.420 | 50 000 | | | | | | |
| sucrose (SUC) | B1−B2 | 0.222 | 30 000 | B1−B2−B4 | 130 | 10 | B1−B2−B4−B5 | 130 | 25 |
| | B2−B3 | 0.247 | 30 000 | B3−B2−B4 | 110 | 150 | B1−B2−B4−B6 | 80 | 2 |
| | B2−B4 | 0.429 | 30 000 | B5−B4−B2 | 20 | 50 | B3−B2−B4−B5 | −70 | 20 |
| | B4−B5 | 0.293 | 30 000 | B6−B4−B2 | 85 | 150 | | | |
| | B4−B6 | 0.372 | 30 000 | | | | | | |
| maltose (M) | B1−B2 | 0.222 | 30 000 | | | | | | |
| | B2−B3 | 0.246 | 30 000 | B1−B2−B4 | 150 | 50 | B1−B2−B4−B5 | 110 | 8 |
| | B2−B4 | 0.561 | 30 000 | B3−B2−B4 | 140 | 50 | B1−B2−B4−B6 | −20 | 5 |
| | B4−B5 | 0.239 | 30 000 | B5−B4−B2 | 70 | 100 | B3−B2−B4−B5 | −80 | 5 |
| | B4−B6 | 0.281 | 30 000 | B6−B4−B2 | 50 | 25 | | | |
| cellobiose (C) | B1−B2 | 0.242 | 30 000 | | | | | | |
| | B2−B3 | 0.284 | 30 000 | B1−B2−B4 | 126 | 50 | B1−B2−B4−B5 | 30 | 8 |
| | B2−B4 | 0.518 | 30 000 | B3−B2−B4 | 120 | 50 | B1−B2−B4-B6 | −150 | 5 |
| | B4−B5 | 0.234 | 30 000 | B5−B4−B2 | 60 | 100 | B3−B2−B4−B5 | −150 | 5 |
| | B4−B6 | 0.278 | 30 000 | B6−B4−B2 | 65 | 25 | | | |
| kojibiose (K) | B1−B2 | 0.222 | 30 000 | | | | | | |
| | B2−B3 | 0.247 | 30 000 | B1−B2−B4 | 127 | 50 | B1−B2−B4−B5 | 165 | 8 |
| | B2−B4 | 0.470 | 30 000 | B3−B2−B4 | 81 | 200 | B1−B2−B4−B6 | 110 | 10 |
| | B4−B5 | 0.358 | 30 000 | B5−B4−B2 | 75 | 400 | B3−B2−B4−B5 | 5 | 30 |
| | B4−B6 | 0.394 | 30 000 | B6−B4−B2 | 120 | 200 | | | |
| sophorose (S) | B1−B2 | 0.222 | 30 000 | | | | | | |
| | B2−B3 | 0.247 | 30 000 | B1−B2−B4 | 90 | 20 | B1−B2−B4−B5 | 40 | 8 |
| | B2−B4 | 0.432 | 30 000 | B3−B2−B4 | 125 | 200 | B1−B2−B4−B6 | 55 | 10 |
| | B4−B5 | 0.384 | 30 000 | B5−B4−B2 | 90 | 350 | B3−B2−B4−B5 | −135 | 5 |
| | B4−B6 | 0.399 | 30 000 | B6−B4−B2 | 125 | 300 | | | |
| nigerose (N) | B1−B2 | 0.222 | 30 000 | | | | | | |
| | B2−B3 | 0.247 | 30 000 | B1−B2−B4 | 87 | 5 | B1−B2−B4−B5 | −15 | 15 |
| | B2−B4 | 0.413 | 30 000 | B3−B2−B4 | 130 | 125 | B1−B2−B4−B6 | −22 | 15 |
| | B4−B5 | 0.294 | 30 000 | B5−B4−B2 | 50 | 250 | B3−B2−B4−B5 | 160 | 1 |
| | B4−B6 | 0.424 | 30 000 | B6−B4−B2 | 96 | 250 | | | |
| laminarabiose (L) | B1−B2 | 0.329 | 30 000 | | | | | | |
| | B2−B3 | 0.376 | 30 000 | B1−B2−B4 | 54 | 80 | B1−B2−B4−B5 | 20 | 15 |
| | B2−B4 | 0.356 | 30 000 | B3−B2−B4 | 124 | 200 | B1−B2−B4−B6 | 55 | 5 |
| | B4−B5 | 0.276 | 30 000 | B5−B4−B2 | 44 | 500 | B3−B2−B4−B5 | 42 | 5 |
| | B4−B6 | 0.372 | 30 000 | B6−B4−B2 | 67 | 800 | | | |
| trehalose (T) | B1−B2 | 0.222 | 30 000 | | | | | | |
| | B2−B3 | 0.246 | 30 000 | B1−B2−B4 | 150 | 100 | B1−B2−B4−B5 | −80 | 8 |
| | B2−B4 | 0.512 | 30 000 | B3−B2−B4 | 95 | 250 | B1−B2−B4−B6 | 123 | 5 |
| | B4−B5 | 0.231 | 30 000 | B5−B4−B2 | 120 | 80 | B3−B2−B4−B5 | −40 | 20 |
| | B4−B6 | 0.220 | 30 000 | B6−B4−B2 | 65 | 180 | | | |

[a] See Figure 2 for a definition of B1−B6.

was divided in five blocks and averages were calculated for each block. The total average error was calculated from the variance between averages over the individual blocks.

## 3. Results and Discussion

**3.1. Monosaccharides.** *3.1.1. Topology.* The monosaccharides considered in this work are glucose (G) and fructose (F). On the basis of the chosen mapping of three CG particles to represent the sugar rings (see Figure 2), the first step consisted in the calibration of the bonded parameters. In the case of the monosaccharides, only three bond potentials (eq 2) are required. The effective bond lengths and force constants were derived from bond length distributions obtained from AA simulations, considering the center of mass of the groups of atoms constituting a CG interaction site (see Methods section for details). The final parameters for the bonded interactions are summarized in Table 2. Bond lengths range between 0.3 and 0.42 nm, similar to the range of bond lengths used in the standard Martini protein force field.[34] The force constants are rather high, reflecting the limited flexibility of the sugar ring. Note that the AA simulations only sampled the $^4C_1$ chair conformation, as expected. In practice, such high force constants may lead to numerical instabilities with the rather large time steps used in CG simulations and are better replaced by constraints without noticeable consequences. The B1−B2 bond of fructose is somewhat weaker in comparison to glucose, reflecting the higher intrinsic flexibility of five- versus six-membered rings.

For the selection of the particle types, we used standard Martini particle types only to ensure compatibility with the other force field components. Considering the high number of hydroxyl groups around the ring, we focused on the polar particle types P4, P3, P2, and P1. The most polar of these, P4, is representative of the ethanediol building block; the

***Table 3.*** Particle-Type Selection for CG Carbohydrates[a]

| molecule | B1 | B2 | B3 | B4 | B5 | B6 |
|---|---|---|---|---|---|---|
| glucose (G) | P1 | P4 | P4 | | | |
| fructose (F) | P1 | P3 | P4 | | | |
| sucrose (SUC) | P1 | P2 | P4 | P1 | P1 | P4 |
| maltose (M) | P1 | P2 | P4 | P2 | P1 | P4 |
| cellobiose (C) | P1 | P2 | P4 | P2 | P1 | P4 |
| kojibiose (K) | P1 | P2 | P4 | P2 | P4 | P1 |
| sophorose (S) | P1 | P2 | P4 | P2 | P4 | P1 |
| nigerose (N) | P1 | P2 | P4 | P2 | P4 | P1 |
| laminarabiose (L) | P1 | P2 | P4 | P2 | P4 | P1 |
| trehalose (T) | P1 | P2 | P4 | P2 | P1 | P4 |

[a] A complete overview of non-bonded interactions parameters can be found in the original Martini force field article.[35]

less polar P1 particle type represents propanol. Since the mapping of the AA structure to the CG model is not unique, we tested various options trying to optimize the performance of the model with respect to reproducing the experimental density and partitioning data (see below). We considered both normal particle types and the special class of ring "S" particles. The latter choice seems more logical given the ring structure of the sugars; however, these ring parameters were originally derived for use in a 2-to-1 or 3-to-1 mapping scheme, whereas the current mapping of the sugars is 4-to-1, consistent with the standard mapping. Although the use of ring particles improves the density profile for our model (see the density section below), this particle type decreases the sugar−sugar self-interaction to such an extent that even pure sugar remains liquid at room temperature. We decided not to change the strength of the self-interaction because it plays a crucial role in the packing and recognition between different sugar groups. We eventually settled on the topology as shown in Table 3. The inhomogeneous distribution of the polarity around the sugar rings is well reflected by the combination of the two more polar particle types (P3,P4) with one less polar particle type (P1). The latter maps to the part of the sugar containing carbons in positions 5 and 6. As we will show below, using this topology, the experimental water/octanol partition coefficient as well as the density of sugar solutions is reasonably well approximated.

*3.1.2. Density.* To test if our CG model produces density of aqueous sugar solutions comparable with experimental data, a systematic set of simulations at different glucose concentrations was performed. Figure 3 shows the density of the solution as a function of glucose concentration. We compared our CG model to data obtained from the literature.[52] We find that the experimental densities are reproduced to within 10% in the condensed phase (e.g., to 60% w/w). At low concentrations (up to ~20% w/w), the agreement is even better. The underestimation of the density, especially at higher glucose concentrations, points to an effect due to sugar−sugar interactions, i.e., the packing of the CG glucose monomers is not as efficient as it is in reality. The situation can be improved somewhat by switching to the special class of ring particle types, which have a smaller effective size as set by the LJ parameter $\sigma_{ii}$. However, also the effective interaction strength, controlled by the value of $\varepsilon_{ii}$, is reduced for this class of particles; as already mentioned before, this has the unwanted consequence that even pure sugar systems are found in a liquid state. With the normal particle types, pure glucose forms a solid structure. Although we do not pretend to be able to reproduce the correct crystal packing with our CG model, at least the effective sugar−sugar interaction is strong enough to capture the transition from a solution toward the solid state upon increasing sugar concentration. Experimentally, the solubility of glucose at 300 K is around 47% w/w content. Using the normal particle types, visual inspection of the trajectories shows that our model becomes clearly more crystal-like around 50% w/w, with increasing aggregation of the solutes and phase separation of sugar rich clusters surrounded by layers of water.

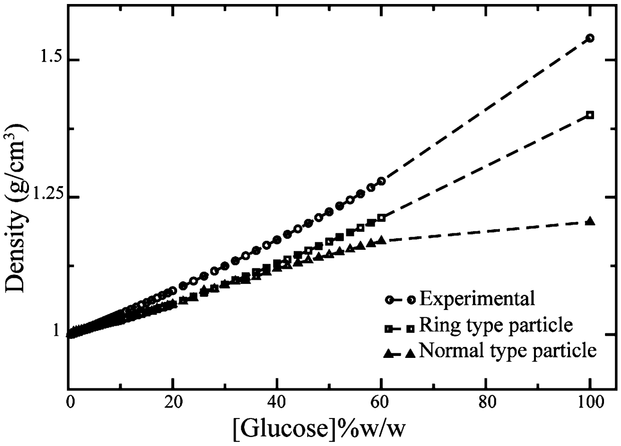


***Figure 3.*** Density of aqueous glucose solutions as a function of concentration. Results from CG simulations at 310 K, for both ring-type particles (squares) and normal-type particles (triangles), are shown. For comparison, experimental data[52] obtained at $T = 305$ K (circles) are included.

*3.1.3. Partitioning.* In the Martini force field development, reproducing realistic partitioning behavior is of central importance. To select the optimal particle types for the carbohydrates, we therefore calculated the octanol/water partitioning free energies of our basic building blocks, glucose and fructose, in order to compare the values to either experimental results or to results obtained from AA simulations. We used the thermodynamic integration approach to calculate the free energies of solvation of the sugars in both water and water-saturated octanol. From the difference, the partitioning free energy is obtained as explained in the Methods section. Figure 4 depicts a comparison of the thermodynamic integration profile for glucose at the AA and CG levels of modeling. The derivative of the Hamiltonian $H$ with respect to the integration parameter $\lambda$, as well as the running integrand, is plotted for successive $\lambda$ points (since only the solute−solvent interaction is $\lambda$ dependent, and the atomic masses are unchanged during the process, $\partial H/\partial\lambda = \partial U_{uv}(\lambda)/\partial\lambda$ in eq 7). It is interesting to note that the $\partial H/\partial\lambda$ profile looks very different for the AA versus the CG model, especially for the simulations in water. This is caused by the difference in particle size of the atomistic solvent molecules versus the CG water beads which unite four individual water molecules. The sharp drop in the profile at $\lambda = 0.6$ in the case of the CG water molecules and for both CG and AA octanol reflects the transition between configurations with overlapping sugar/solvent molecules to the nonoverlapping, normal, situation at full interaction strength. For the relatively small AA water molecules, this transition

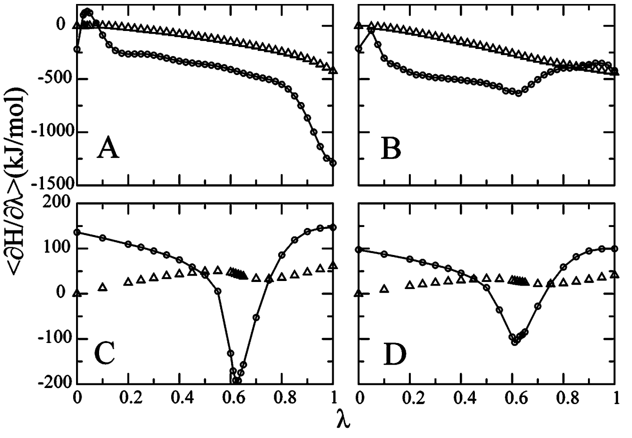


***Figure 4.*** Thermodynamic integration profiles as a function of the integration parameter $\lambda$ for glucose in water (A, C) and water-saturated octanol (B, D). Panels A and B are obtained with the AA model and panels C and D with the CG model. The solid line denotes the derivative of the free energy with respect to $\lambda$; the triangles represent the integrated curve. The integration has been performed using a trapezoidal scheme. The integrated value at $\lambda = 1$ corresponds to the free energy of hydration (A, C) or solvation in water-saturated octanol (B, D). The magnitude of the error at each $\lambda$ point is smaller than the diameter of the circles, i.e., less than 10 kJ/mol for the AA and 1 kJ/mol for the CG system.

takes place at a much smaller value of $\lambda$ (not clearly noticeable). The free energy, however, does not depend on the details of the integration path and corresponds to the integrated value at $\lambda = 1$.

The results of the free energy calculations are summarized in Table 4 and compared to available experimental values. The CG model reproduces the correct trend for free energies of solvation, although the actual values are systematically too low. This observation is in line with the results for different functional groups in the Martini force field.[35] As long as its application is aimed at studying the condensed phase and not at reproducing gas/fluid or solid/fluid coexistence regions, the most important thermodynamic property is the partitioning free energy. Importantly, the water/octanol partitioning of both monosaccharides can be accurately reproduced with the current parametrization of the CG model. A comparison to AA simulations and experimental data reveals a close agreement to within 2kT.

**3.2. Disaccharides.** *3.2.1. Topology.* Each of the disaccharides is modeled by two three-bead monosaccharide residues joined together by a single bond (see Figure 2). As discussed in the Method section, the monosaccharide units are represented differently from the individual monosaccharides. Analogous to the procedure followed for the monosaccharides, the parameters for the bonded interactions were obtained from a comparison to mapped AA trajectories. The full set of parameters is listed in Table 2. Bond lengths, defining the overall geometry of the molecules, vary somewhat depending on the sugar type, are found in the range 0.22−0.4 nm within each of the two sugar residues, and are slightly larger (up to 0.47 nm) for the B2−B4 bond representing the glycosidic linkage between them. Little variability was found for the stiffness of these bonds, so the force constant was set to 30 000 kJ mol$^{-1}$ nm$^{-2}$ for all of them. Similar to the case of the monosaccharides, these bonds can be replaced by constraints in practice.

The conformation of disaccharides is mainly determined by the populations of rotamers around the glycosidic linkage. At the AA level, these rotameric conformations can be described by the glycosidic torsional angles $\varphi$ ($O_5$−$C_1$−$O_1$−$C_n'$) and $\psi$ ($C_1$−$O_1$−$C_n'$−$C_{n-1}'$) around a $(1-n)$-linkage[25] (with $n$ = 2, 3, 4, 6 and $\varphi = \psi$ for trehalose). In the CG representation, the distinction between these dihedral angles is lost. Instead, the rotameric phase space available to the disaccharides needs to be represented by a set of dihedral angles (eq 4) and normal angles (eq 3) along the B2−B4 glycosidic bond. Upon transformation of the atomistic trajectories to effective CG trajectories, specific distributions for the angles and dihedrals were obtained, reflecting the conformational freedom of the disaccharides in water. As an example, Figures 5 and 6 show these distributions for every angle and dihedral of trehalose. The distributions obtained with the CG model are also shown, revealing that the atomistic configurations can be quite accurately mapped by our simplistic CG topology. The probability distributions associated with the $\varphi$ and $\psi$ glycosidic dihedral angles are essentially unimodal. Full rotation around these angles is observed at most once or twice for $\psi$ (never for $\varphi$) on the 50 ns time scale in the AA simulation studies performed by Pereira et al.,[25] corroborated with our AA simulations (data not shown). The exception is the distribution for 1−6 linked disaccharides (isomaltose and gentiobiose) for which a bimodal distribution is observed.[25] Since it is impossible to represent such a distribution with the dihedral approach used here (eq 4), the 1−6 linked sugars were omitted from the current study and left for future refinement.

The assignment of particle types follows the assignment done for the monosaccharides, with the exception of the B2 and B4 beads involved in the glycosidic bond. The formation of a glycosidic bond between two glucose residues decreases the number of hydroxyl groups, which are largely responsible for the polarity of the molecule. To mimic this effect, a less polar particle type, P2, was used for the B2 and B4 beads (compared to P4 for the monosaccharides, see Table 3). For sucrose, the polarity was even further decreased for the correct reproduction of the partition coefficient, see below.

*3.2.2. Partitioning.* The water/octanol partitioning free energy, corresponding to the difference in solvation energy in water and octanol, is summarized in Table 4 for the different disaccharides considered in this study. Note that we only calculated the free energies for sucrose and maltose explicitly; the other sugars share the same particle assignment with maltose. Test runs showed that the geometrical fine details and overall conformational flexibility do not affect the free energies to within the error estimate of 1 kJ mol$^{-1}$. For only two of the sugars (sucrose and trehalose) experimental data is available,[53] which is well matched by our CG model. Regarding the other disaccharides, comparison can be made to the atomistic model, for which we also computed the free energies of solvation and partitioning as listed in Table 4. There is a general good agreement between the results obtained at the atomistic and coarse-grained level. The atomistic model reveals little effect of the chemical

***Table 4.*** Thermodynamic Parameters of Solvation and Partitioning Calculated for CG and AA Carbohydrates[a]

| molecule | $\Delta G^W$ (AA) (kJ mol$^{-1}$) | $\Delta G^O$ (AA) (kJ mol$^{-1}$) | $\Delta\Delta G_{OW}$ (AA) (kJ mol$^{-1}$) | log $P_{OW}$(AA) | $\Delta G^W$ (CG) (kJ mol$^{-1}$) | $\Delta G^O$ (CG) (kJ mol$^{-1}$) | $\Delta\Delta G_{OW}$ (CG) (kJ mol$^{-1}$) | log $P_{OW}$(CG) | log $P_{OW}$ (exp) |
|---|---|---|---|---|---|---|---|---|---|
| glucose (G) | −89 | −74 | 15 | −2.5 | −60 | −43 | 17 | −2.9 | −2.8 |
| fructose (F) | −80 | −69 | 11 | −2.0 | −60 | −44 | 16 | −2.7 | |
| sucrose (SUC) | −107 | −89 | 18 | −3.0 | −103 | −83 | 20 | −3.4 | −3.3 |
| maltose (M) | −121 | −96 | 25 | −4.2 | −120 | −96 | 24 | −4.0 | |
| cellobiose (C) | −114 | −90 | 24 | −4.0 | −120 | −96 | 24 | −4.0 | |
| kojibiose (K) | −121 | −93 | 28 | −4.7 | −120 | −96 | 24 | −4.0 | |
| sophorose (S) | −120 | −88 | 32 | −5.4 | −120 | −96 | 24 | −4.0 | |
| nigerose (N) | −119 | −89 | 30 | −5.0 | −120 | −96 | 24 | −4.0 | |
| laminarabiose (L) | −120 | −91 | 29 | −5.0 | −120 | −96 | 24 | −4.0 | |
| trehalose (T) | −120 | −92 | 28 | −5.0 | −120 | −96 | 24 | −4.0 | −3.78 |

[a] The partition coefficient of octanol−water log $P_{OW}$ is based on the difference between the independently calculated free energy of hydration ($\Delta G_W$) and free energy of solvation in water-saturated octanol ($\Delta G_O$), according to eq 6. Simulation data were obtained at 310 K, whereas the temperature of the experimental data (log $P_{OW}$(exp)) is 300 K.[53] The statistical accuracy of the free energies obtained from CG simulations is 1 kJ mol$^{-1}$. The CG values for cellobiose, kojibiose, sophorose, nigerose, laminarabiose, and trehalose were set equal to the values for maltose (as the topologies are based on the same kind of coarse particle types).

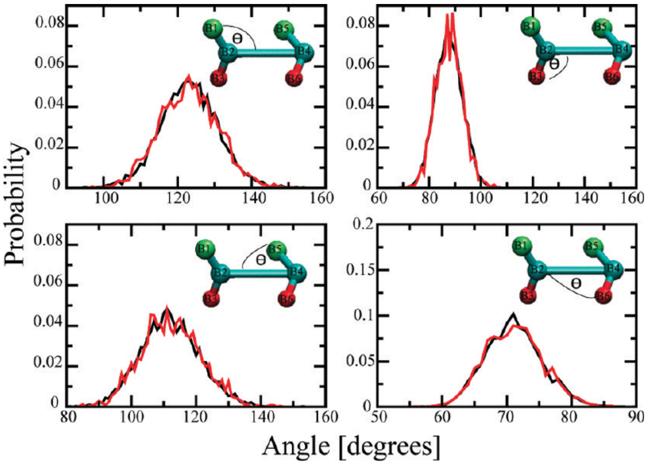


***Figure 5.*** Angle distributions ($\theta$) for trehalose obtained from AA simulations after mapping to CG particles (black line) and from CG simulations (red line). Four angles, indicated by the insets, were used to preserve the characteristic rotameric states of the disaccharide at the coarse-grained level.

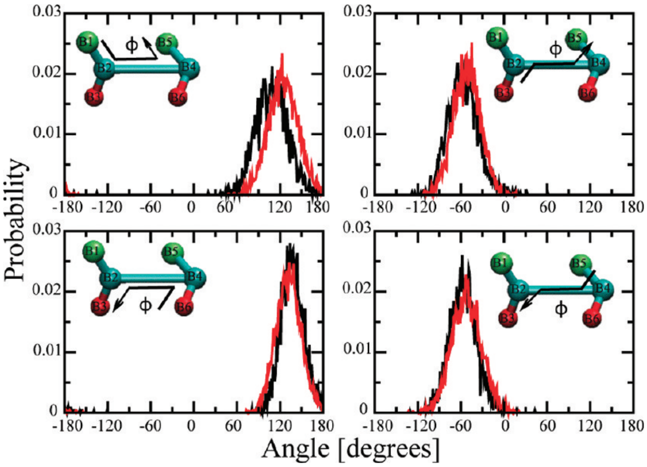


***Figure 6.*** Dihedral angle distributions ($\phi$) for trehalose obtained from AA simulations after mapping to CG particles (black line) and from CG simulations (red line). In addition to the angle potentials (cf. Figure 5), four dihedral angles, indicated by the insets, were used to preserve the characteristic rotameric states of the disaccharide at the coarse-grained level.

details of the sugars on their partitioning, justifying our choice of the same particle types for all disaccharides except sucrose.

**3.3. Oligosaccharides.** Molecular systems in which carbohydrates are involved are not limited to monosaccharides or disaccharides but also include long polymeric sugar chains exhibiting a large variety in monosaccharide composition, linkage type, and degree of branching. For this reason, the following part of the article illustrates how the parameters derived for the simulation of mono- and disaccharides can be applied to study the structure and dynamics of oligosaccharides. As an example, two different oligosaccharides have been studied at both the CG and the AA levels. First, two amylose-type chains of different lengths are considered as an example of α 1−4 linked polymers. Second, a triple-helix structure representing Curdlan is considered as an example of a $\beta$ 1−3 linked structure.

*3.3.1. Amylose.* Before considering the full amylose chain, we studied the short oligosaccharide maltoheptaose. Maltoheptaose consists of seven glucose monomers, connected by α 1−4 glycosidic linkages just like amylose. The behavior of maltoheptaose was studied both in aqueous and in nonpolar environments (nonane). Results from the CG simulations are compared to results obtained with an AA using a similar setup consisting of a single molecule in excess solvent.

In our first attempt, we took the parameters derived for the disaccharides and simply extended the topology to model the heptamer. However, with this setup, the configurations sampled at the CG and AA levels did not overlap (data not shown). Therefore, an extra angle potential (eq 3) was used to reproduce the correct structural shape of this molecule in water as well as in nonane. Three consecutive backbone particles were subject to an angle potential, as illustrated in Figure 2C. We found that the optimal angle parameters are an equilibrium angle ($\theta_0$) of 154° in water and 120° in nonane with a force constant ($K_{\text{angle}}$) of 100 and 250 kJ mol$^{-1}$ respectively. Including these additional ‘three-sugar’ potentials, the CG representation nicely matches the structure observed in the AA simulation, as is illustrated by the snapshots shown in Figure 7B for the case of maltoheptaose in water. Figure 7A also provides a comparison between the

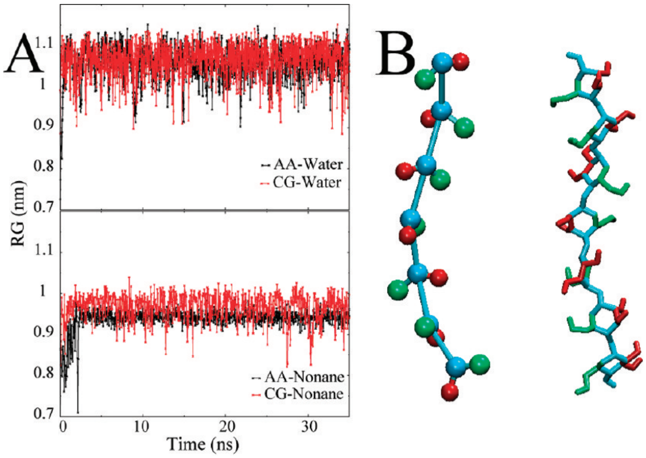


***Figure 7.*** Structure and dynamics of the maltoheptaose oligomer. (A) Radius of gyration (RG) as a function of the simulation time for the AA (black lines) and CG (red lines) systems. (B) Snapshot obtained from the CG (left) and AA (right) simulation in water.

temporal evolution of the radius of gyration (RG) of maltoheptaose in both water and nonane. The average value for both the AA and the CG simulation in water is 1.05 nm, which is in agreement with the previous value obtained by Shimada et al.,[13] revealing a more extended structure compared to the structure obtained by small-angle X-ray scattering. In nonane, the structure is somewhat more compact with an average RG of 0.95 (AA) and 0.98 nm (CG). Judging from the fluctuations of the RG over time, the AA structure is slightly more rigid compared to the CG one.

Our results show that, with the addition of an angle term acting between three consecutive sugar moieties, the short oligosaccharide maltoheptaose can be quite accurately modeled. The question is, will this suffice to also model the longer oligosaccharide amylose, the principal component of starch? Like maltoheptaose, amylose is a linear oligosaccharide of 1−4 α linked glucose monomers. In aqueous solution it behaves as a flexible random coil with stretches of left-handed helical segments, which are more pronounced at low hydration levels.[54] In fact, AFM experiments have shown that after the assisted unfolding of the molecule, it tries to refold again to a helical conformation. However, this refolding was not complete[55] unless a less polar solvent (butanol) was used. In general, stable secondary conformations, known as A-, B-, and V-amylose,[54] are formed in either ionic solutions or less polar solvents. The A and B allomers consist of parallel left-handed double helices with six glucopyranosil units per turn, differing only in the number of helices packed in the unit cell. V-Amylose, cocrystallized with compounds such as iodine, DMSO,[56] alcohols, or fatty acids, reveals a strict structure of left-handed helix[57] with six to eight glucose residues per turn. Multiple helices form a central channel in which the additives are complexed. In fact, a large number of V-amylose crystalline structures have been obtained, depending on the exact crystallization conditions.[54]

On the basis of the parameter set for maltoheptaose, several CG simulations were performed for a 26-glucose amylose chain. For comparison, AA simulations were performed as well. Both water and nonane were used as solvents. Figure 8A shows a comparison of the structure formation of the amylose chain in nonane at the AA and CG level. The initial conformation was a regular helical structure characterized by the torsional angles $\varphi$ ($O_5-C_1-O_1-C_4'$) = 87.27°, $\psi$ ($C_1-O_1-C_4'-C_3'$) = 101°, and $\omega'$ ($O_5-C_5-C_6-O_6$) = −125.97°. After 25 ns simulation, both simulations (AA and CG) refold to the same helical form characteristic of V-amylose. This structure proved stable for the remainder of the simulation (40 ns in total). Figure 9 shows a close-up view of both the AA and the CG equilibrium structure, with the values for the pitch of the helix and diameter of the channel indicated. The atomistic and coarse-grained structures are nearly indistinguishable. Experimentally, the V-amylose conformation is observed by X-ray studies[11,57] in complex with nonpolar solvents, in agreement with our simulations. Moreover, the experimentally determined pitch value, counting 6 glucose residues per turn, gives an average value of 7.9 Å. In our simulations, the average value is 7.5 Å (for both AA and CG), in good agreement with the experiment, given the differences between the experimental and simulation conditions.

In water, the amylose chain remains largely unfolded during the simulation, also in agreement with the experimental observation.[58] Figure 8B shows typical snapshots of AA and CG amylose, revealing a somewhat extended, fully solvated structure. To quantify the degree of extension, the radius of gyration was calculated. Averaged over 40 ns, the radius of gyration of the amylose chain is found to be similar at both levels of resolution, namely, 3.2 and 3.0 nm (±0.1 nm) for the CG and AA system, respectively.

*3.3.2. Curdlan.* In recent years, there has been a great interest in the $\beta$-1,3-D-glucan series (Curdlan) because they show antitumor and anti-HIV viral activity in humans.[56] After a rigorous purification step, the structure of Curdlan has been identified as a right-handed triple-helical complex under aqueous conditions.[59] A number of additional crystal structures[60,61] also reveal a triple helix formed by three parallel independent chains and stabilized by both inter- and intramolecular hydrogen bonds.[28]

The CG model of Curdlan was based on the laminarabiose disaccharide. Similar to the case of extending maltose into oligomers, an extra angle potential was used for three consecutive backbone beads (cf. Figure 2C). The value of this angle was obtained from matching CG to AA conformations of a chain consisting of seven $\beta$ 1−3 glucose monomers in water, sampled over a 40 ns trajectory. The optimal parameters were found to be $\theta_0 = 136°$ with $K_{\text{angle}}$ of 500 kJ mol$^{-1}$. A simulation of the same molecule was performed in nonane in order to determine if this angle depends on the solvent. Contrary to the case of the α 1−4 linked sugars described above, here we found no significant dependency of the additional angle potential on solvent environment.

Next, three chains composed of 26 $\beta$ 1−3 glucose monomers, representing Curdlan, were simulated in nonane. The additional angle potential was included. The starting structure was taken from Deslandes' CUR data,[59] determined by X-ray crystallography. Figure 10A shows the final

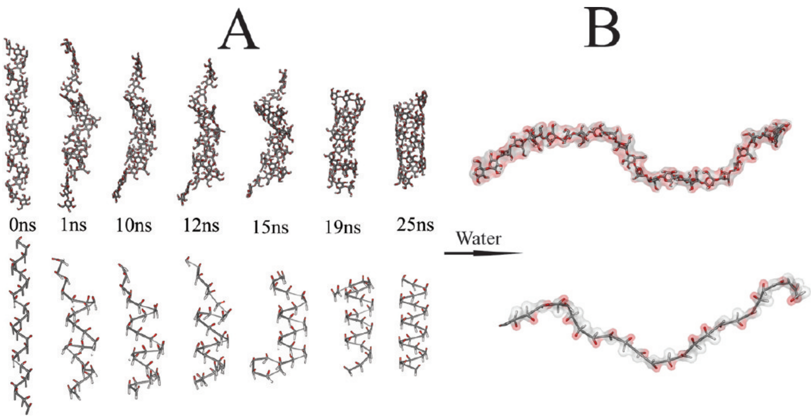


***Figure 8.*** Structure of amylose in nonane and water. (A) Snapshots of a 26-mer of amylose, simulated in nonane, at AA (top drawings) and CG (bottom drawings) levels. In both cases, starting from an elongated helix, a transition to a stable V shape is observed during 25 ns of simulation. (B) After transferring this molecule to an aqueous solution, amylose evidences an increased flexibility with unfolding of the helical structure.

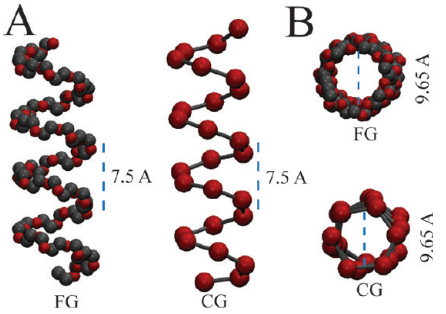


***Figure 9.*** Structural characteristics of amylose in nonane. Only backbone particles (gray/red spheres) are shown. The pitch value (A) and helix diameter (B) of both AA and CG representations are identical, as indicated.

structures obtained after 40 ns both at the AA and CG levels of resolution. The characteristic triple-helical structure of Curdlan is well preserved. No significant difference was detected between the average CG and the AA structure. This is illustrated in Figure 10B, showing root mean squared deviation (rmsd) between the AA (mapped to CG coordinates) and the CG representation. Moreover, we calculated the value for the pitch of each turn, which is defined as the length between the six main-chain glucose units along the helix *c* axis.[59] The result is also shown in Figure 10B. The average pitch value for both AA and CG structure, 2.0 nm, is in good agreement with the experimental value of 1.8 nm,[59] despite the differences in environment (apolar solvent versus crystal). The pitch value obtained in the simulations also agrees well with previous simulation studies.[28] In addition, we tested the stability of the triple-helix configuration of Curdlan in water at both levels of resolution. We basically found the same results as observed with nonane as a solvent, indicating that our CG approach can effectively mimic the strength of the intra- and intermolecular hydrogen bonds, preserving Curdlan's triple-helical structure regardless of the solvent.

**3.4. Cryo- and Anhydro-Protection Effects.** The ability of sugars to act as cryo- and anhydro-protective agents has been well established.[3,62] Several organisms make use of this property of sugars; by increasing their intracellular sugar concentration they have been found to survive under low-temperature or low-hydration conditions over extended periods of time. The origin of the cryo- or anhydro-protective effect of membranes is usually explained by different mechanisms, namely, (i) replacement of the lipid−water hydrogen bonds by lipid−sugar hydrogen bonds, (ii) entrapment of lipid hydration water, and (iii) vitrification effects. There is an ongoing debate about which of these mechanisms dominates in the modulation of many bilayer properties.[63] One of these properties is the main phase transition temper-

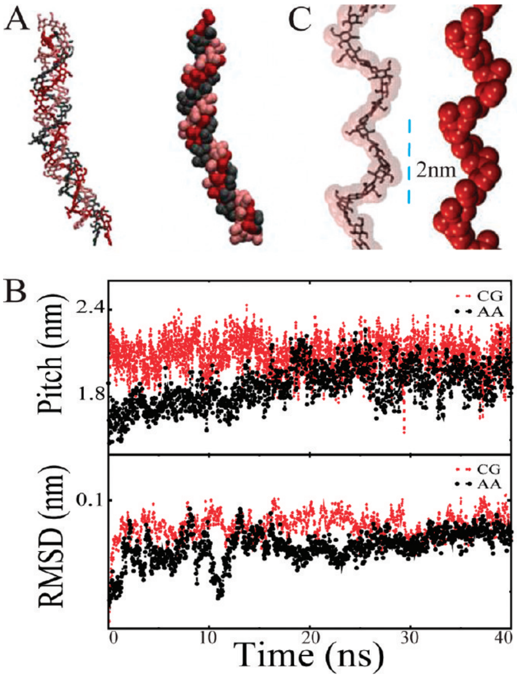


***Figure 10.*** Comparison of the AA and CG structure of Curdlan in nonane. (A) Snapshots of the triple-helix structure, stable for both the AA (left) and the CG (right) systems (each helix represented by a different color type). (B) Temporal evolution of the helical pitch (top graph) and rmsd with respect to the starting structure (bottom graph) for both levels of resolution. (C) Indication of the pitch distance for a single helical strand at AA (left) and CG (right) resolution. For best comparison, only one helix at the AA level (stick representation with transparent spheres) and one at the CG representation (red balls) were considered.

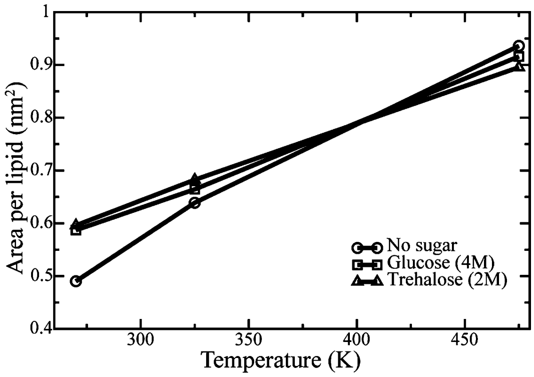


***Figure 11.*** Average area per lipid of a DPPC membrane as a function of temperature. Circles: system with no sugars. Squares: 4 M glucose solution. Triangles: 2 M trehalose solution.

ature of the lipid bilayer, which is lowered in the presence of sugars. Thus, the biologically relevant liquid-crystalline phase is stabilized, protecting the cell membrane from freezing.

In order to test the capabilities of our model to reproduce this effect, we performed several 500 ns CG simulations of a DPPC bilayer in the presence of glucose and trehalose at 4 and 2 M, respectively, varying the temperature in the range 270−475 K. A pure DPPC bilayer without sugars was also simulated for comparison. The area per lipid was used to monitor the protective effect of the sugars. In Figure 11 the dependency of the average area per lipid on the temperature is shown for each of the three systems. A clear difference between the pure DPPC membrane, on the one hand, and the DPPC membrane in the presence of sugars, on the other hand, can be appreciated. Whereas the pure system adopts an area per lipid of 0.48 nm$^2$ at 270 K, characteristic of a gel phase, both trehalose and glucose manage to keep the area per lipid at a value similar to the value at 325 K, in a liquid-crystalline state. At elevated temperatures, the sugars appear to have less effect, although the thermal expansivity (i.e., the slope of the curves in Figure 11) is somewhat smaller in the presence of the sugars. A close view of the structural effect of the sugars on the bilayers at the low-temperature range is depicted in Figure 12. In the pure system, the DPPC tails at 270 K have adopted a straight conformation, and the system has transformed into a gel phase. Note that the Martini model does not reproduce the experimentally observed tilt of the lipid tails in the gel state.[64] However, the bilayer remains in a liquid-crystalline state when a large amount of glucose is added; this is clearly noticeable by the disordered acyl chains. The same phenomenon was also observed when trehalose was added (using one-half the concentration of glucose).

Although a real vitrification process cannot be directly observed at the CG level, our simulations point to a direct interaction between the sugars and the DPPC lipid head groups. Figure 13 shows the density profiles of the lipids, water, and sugars across the bilayer normal. The profile for glucose shows that the sugars are able to bind to the lipid/water interface and can penetrate the membrane up to the level of the carbonyl groups. Consequently, the amount of water in the interface is reduced. By intercalating between the head groups, the sugar molecules are replacing part of the hydrating water and the freezing of the acyl chains is avoided. The same characteristic was also observed for trehalose. These results are most easily interpreted in the context of the water-replacement hypothesis, as has been proposed by Pereira et al.[63,65] and Sum et al.,[66] based on atomistic simulations.

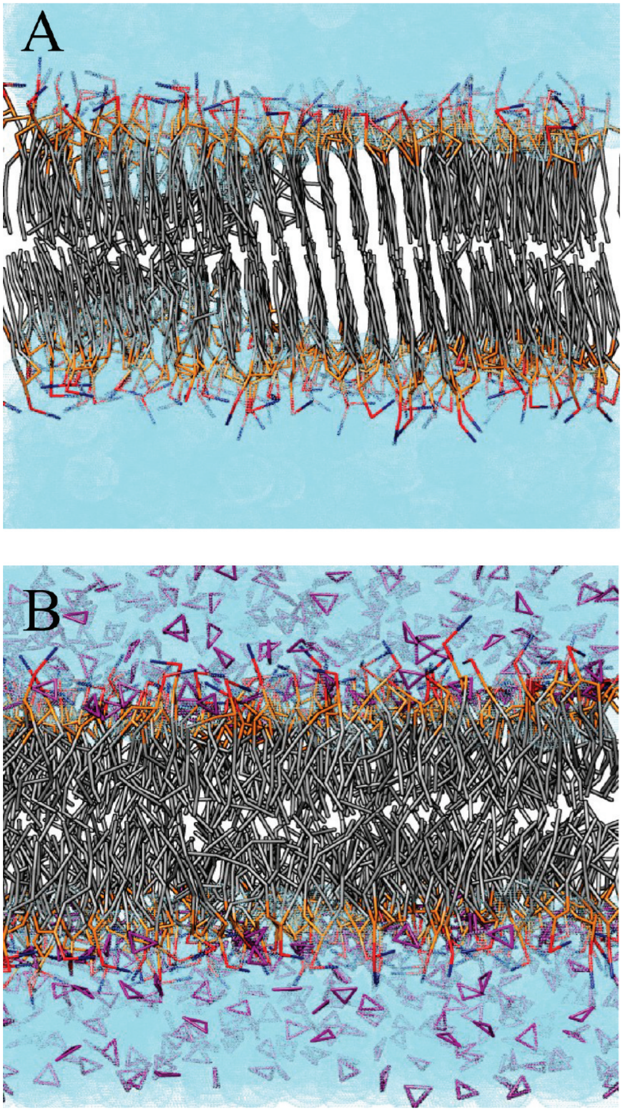


***Figure 12.*** Snapshots of CG DPPC bilayers at $T = 270$ K. (A) Gel phase formed in pure water. (B) Fluid phase stabilized in the presence of 4 M glucose. Lipid tails are depicted in gray, head groups are shown blue/orange for choline/phosphate moieties, and sugars are shown in purple. Water is represented by semitransparent blue dots.

## 4. Limitations and Outlook

The potential range of applications of the carbohydrate Martini model is very broad. Along the lines of the parametrization as presented in this article, the model could be extended toward other (oligo)saccharides. There are, however, certain important limitations which should be kept in mind. An obvious limitation is that the limited resolution of our CG model prevents the distinction between α and $\beta$ anomers of reducing ends as well as the stereochemistry of

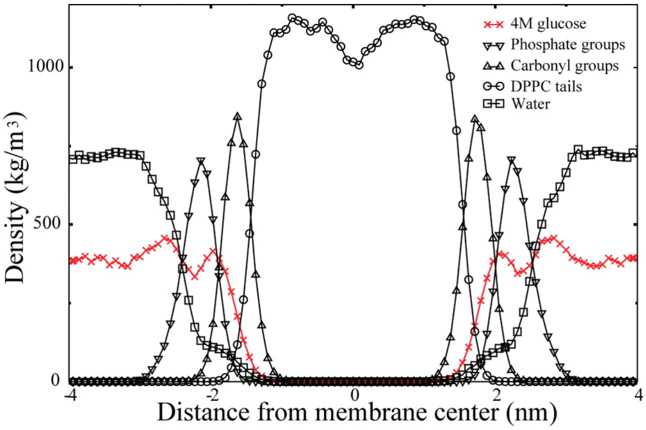


***Figure 13.*** Mass density profiles across a CG DPPC bilayer in the presence of 4 M glucose at $T = 270$ K. Crosses: glucose. Triangles down: phosphate group. Triangles up: carbonyl groups. Circles: lipid tails. Squares: water.

exocyclic OH groups. In aqueous solution, the sugar hydroxyl groups are directed outward to oxygen atoms of adjacent water molecules, thus destroying the intramolecular hydrogen bonding.[67] Furthermore, an important simplification of the model is that ring puckering (i.e., chair−chair or chair−boat transformations) in hexopyranoses is completely neglected. Only the $^4C_1$ chair conformation is considered. This is not a real problem for the simulation of long polymers, since the $^4C_1$ chair conformation is the dominant puckering state[68,69] and the conformations of oligosaccharides are mainly determined by the accessible rotameric states around the glycosidic linkage. There are, however, exceptions, notably idose and some sulfated sugars.[22] In the case of furanoses, a single conformation is less of a problem; five-membered rings are floppy, but the overall shape does not change substantially as the ring undergoes pseudo-rotation. Special attention is also required in systems with high sugar density, e.g., at low hydration conditions. We have shown an important deviation of the carbohydrate packing density under these conditions. This effect is most severe in the limit of a pure crystalline sugar phase but might also show up, for instance, in the condensing efficiency of long oligosaccharide chains in poor solvents.

The disaccharides used in this work exhibit primarily a single state for the glycosidic bond, which we showed to be easily represented using a dihedral potential at the CG level. Whether this is true for oligosaccharides in general is questionable. The glycosidic linkage is indeed in a single state (denoted syn) for all α-linked disaccharides but not necessarily for $\beta$-linkages. There is evidence from NMR that the other conformations are populated in solution, and crystal structures of protein/sugar complexes indicate antistates.[70–72] The underlying AA force field may also exhibit these states, but revealing them might require more extensive (nonequilibrium) sampling. Multiple states are only sampled in the case of the 1−6 bond. As previously found for the rotation of the hydroxymethyl group in hexopyranoses,[18] there is a clear preference for the *gg* ($\omega = 180°$) and *gt* ($\omega = -60°$) rotamers which are characterized by nearly identical free energies. At the CG level, this behavior is not easily represented by a single well potential, and therefore, the 1−6 linkage has not been considered here. An alternative to overcome the above limitations is the use of tabulated two- or three-well angle or dihedral potentials, which is currently under investigation. Anyhow, we have to keep in mind that the underlying atomistic force field may also have its shortcomings, despite the fact that it has been tested thoroughly.[16,19,22,25] We also note that the current parametrization is restricted to D-D sugars; L-L sugars are easily represented by changing the sign of the dihedral reference value, but for the L-D sugars, reparametrization of the bonded parameters is required.

Another important issue is transferability of the CG parameters. We found that the disaccharide glycosidic bond parameters were easily extrapolated to oligomers, except for the use of an additional angle potential required to adopt the appropriate structure (as judged from AA simulations). This angle potential is likely to be nontransferable, i.e., it needs to be parametrized for different type of linkages. Besides, the optimal parameters were found to depend on the type of solvent, at least in the case of 1−4 α linked sugars. This is not a desirable situation. Especially for applications of oligosaccharides near interfaces, the current parametrization is expected to be problematic.

The parametrization presented in this article should be viewed only as a first step toward a comprehensive carbohydrate force field. Improvements are expected to take place hand in hand with the ongoing development of the Martini force field. The model is easily extendable to include polymers containing more than one type of sugar−sugar linkage or featuring branched sugars. Potential applications include a variety of oligosaccharides such as cellulose and dextran and mixed systems such as membranes containing glycolipids and the bacterial cell wall.

As a spin off of the parametrization of the CG force field, we found that fast folding of oligosaccharides can be observed in nonpolar solvents, even at an atomistic level of resolution. As an example, the folding of a 26 sugar residue amylose chain was presented (cf. Figure 9). Within 25 ns, the molecule changed its fold from an extended conformation toward the experimentally observed crystal structure, the V-shape. In simulations, crystallization conditions of sugars are seemingly effectively reproduced by the use of low-dielectric solvents. In contrast to proteins, which possess side chains of different polarity, sugars can be forced to fold toward the native crystal structure by replacing the hydration shell by nonpolar solvents.

In summary, in this article, an extension of the Martini force field parameters to carbohydrates has been presented. On the basis of atomistic simulations, a complete set of bonded parameters was extracted to model the dynamics and structure of several mono- and disaccharides at the CG level. Standard particle types of the Martini force field were used for the nonbonded interactions, assuring that the carbohydrate model is fully compatible with the other biomolecular components. Since most applications of the CG model are naturally in the condensed phase, the reproduction of the correct partitioning free energies between polar and nonpolar phases is essential. We demonstrated that our model predicts values for water/octanol partitioning in close agreement with atomistic data and where available with experimental measurements. With an additional angle potential to account for the collective effect of three linked sugar residues, the model appears well suited for application to oligosaccharides. Both an amylose chain and a triple-helical Curdlan structure were modeled; despite the lack of explicit hydrogen bonding at the CG level, the conformation and dynamics were found to be in good agreement with simulations at the all-atom level. In addition, our CG model is able to reproduce semiquantitatively the modulating effect of sugars on lipid bilayers, in particular their cryo- and anhydro-protective effect.

Keeping in mind certain inherent limitations of the CG carbohydrate model, such as the inability to represent ring puckering or some of the complex rotameric states exhibited by certain sugar links, the model shows great promise for exploring the phase space of carbohydrate systems which are computationally too costly at full atomistic resolution. Moreover, the sugar parameters are fully compatible with the other parameters in the Martini force field, opening the way to explore a large variety of sugar-containing biomolecular systems at an unprecedented scale.

**Acknowledgment.** The authors would like to acknowledge Cristina Pereira for her kind help with the AA simulations.

**Supporting Information Available:** The complete list of parameters for the simulation of furanose rings (fructose and sucrose) at full atom resolution is provided as supporting information. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Bush, C. A.; Martin-Pastor, M.; Imberty, A. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 269–293.

(2) Ladmiral, V.; Melia, E.; Haddleton, D. M. *Eur. Polym. J.* **2004**, *40*, 431–449.

(3) Dashnau, J. L.; Sharp, K. A.; Vanderkooi, J. M. *J. Phys. Chem. B* **2005**, *109*, 24152–24159.

(4) Crowe, J. H.; Oliver, A. E.; Tablin, F. *Integr. Comp. Biol.* **2002**, *42*, 497–503.

(5) Crowe, L. M. *Comp. Biochem. Physiol., A: Mol. Integr. Physiol.* **2002**, *131*, 505–513.

(6) Wormald, M. R.; Petrescu, A. J.; Pao, Y.-L.; Glithero, A.; Elliott, T.; Dwek, R. A. *Chem. Rev.* **2002**, *102*, 371–386.

(7) Duus, J.; Gotfredsen, C. H.; Bock, K. *Chem. Rev.* **2000**, *100*, 4589–4614.

(8) Yu, F.; Prestegard, J. *Biophys. J.* **2006**, *91*, 1952–1959.

(9) Momany, F. A.; Willett, J. L. *Biopolymers* **2002**, *63*, 99–110.

(10) Corzana, F.; Motawia, M. S.; Du Penhoat, C. H.; Perez, S.; Tschampel, S. M.; Woods, R. J.; Engelsen, S. B. *J. Comput. Chem.* **2004**, *25*, 573–86.

(11) Buléon, A.; Delage, M. M.; Brisson, J.; Chanzy, H. *Int. J. Biol. Macromol.* **1990**, *12*, 25–33.

(12) Li, H.; Rief, M.; Oesterhelt, F.; Gaub, H. E.; Zhang, X.; Shen, J. *Chem. Phys. Lett.* **1999**, *305*, 197–201.

(13) Shimada, J.; Kaneko, H.; Takada, T.; Kitamura, S.; Kajiwara, K. *J. Phys. Chem. B* **2000**, *104*, 2136–2147.

(14) Behrends, R.; Kaatze, U. *ChemPhysChem* **2005**, *6*, 1133–1145.

(15) Green, D. F. *J. Phys. Chem. B* **2008**, *112*, 5238–5249.

(16) Kräutler, V.; Müller, M.; Hünenberger, P. H. *Carbohydr. Res.* **2007**, *342*, 2097–2124.

(17) Guvench, O.; Greene, S. N.; Kamath, G.; Brady, J. W.; Venable, R. M.; Pastor, R. W.; Mackerell, A. D. *J. Comput. Chem.* **2008**, *29*, 2543–2564.

(18) Kony, D.; Damm, W.; Stoll, S.; van Gunsteren, W. F. *J. Comput. Chem.* **2002**, *23*, 1416–1429.

(19) Lins, R. D.; Hünenberger, P. H. *J. Comput. Chem.* **2005**, *26*, 1400–12.

(20) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeirino, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008**, *29*, 622–655.

(21) Spiwok, V.; Lipovová, P.; Skálová, T.; Vondráčková, E.; Dohnálek, J.; Hašek, J.; Králová, B. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 887–901.

(22) Hansen, H. S.; Hünenberger, P. H. *J. Comput. Chem.* **2009**. DOI: 10.1002/jcc.21253.

(23) Stenutz, R.; Widmalm, G. *Glycoconjugate J.* **1998**, *15*, 415–418.

(24) Kozar, T.; Tvaroska, I.; Carver, J. P. *Glycoconjugate J.* **1998**, *15*, 187–191.

(25) Pereira, C. S.; Kony, D.; Baron, R.; Müller, M.; van Gunsteren, W. F.; Hünenberger, P. H. *Biophys. J.* **2006**, *90*, 4337–4344.

(26) Sugiyama, H.; Nitta, T.; Horii, M.; Motohashi, K.; Sakai, J.; Usui, T.; Hisamichi, K.; Ishiyama, J. I. *Carbohydr. Res.* **2000**, *325*, 177–182.

(27) Yu, H.; Amann, M.; Hansson, T.; Köhler, J.; Wich, G.; van Gunsteren, W. F. *Carbohydr. Res.* **2004**, *339*, 1697–1709.

(28) Okobira, T.; Miyoshi, K.; Uezu, K.; Sakurai, K.; Shinkai, S. *Biomacromolecules* **2008**, *9*, 783–788.

(29) Voth, G. A. *Coarse-Graining of Condensed Phase and Biomolecular Systems*; CRC Press: Boca-Raton, 2008.

(30) Liu, P.; Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 11566–11575.

(31) Molinero, V.; Goddard, W. A. *J. Phys. Chem. B* **2004**, *108*, 1414–1427.

(32) Bathe, M.; Rutledge, G. C.; Grodzinsky, A. J.; Tidor, B. *Biophys. J.* **2005**, *88*, 3870–3887.

(33) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750–760.

(34) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. *J. Chem. Theory Comput.* **2008**, *4*, 819–834.

(35) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

(36) Risselada, H. J.; Marrink, S. J. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 17367–17372.

(37) Baoukina, S.; Monticelli, L.; Risselada, H. J.; Marrink, S. J.; Tieleman, D. P. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 10803–10808.

(38) Yefimov, S.; van der Giessen, E.; Onck, P. R.; Marrink, S. J. *Biophys. J.* **2008**, *94*, 2994–3002.

(39) Rzepiela, A. J.; Schafer, L. V.; Goga, N.; Risselada, H. J.; de Vries, A. H.; Marrink, S. J. *J. Comput. Chem.* **2009**, in press.

(40) Best, S. A.; Merz, K. M.; Reynolds, C. H. *J. Phys. Chem. B* **1999**, *103*, 714–726.

(41) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(42) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(43) Costa, M. T. M. *Carbohydr. Res.* **2005**, *340*, 2185–2194.

(44) Engelsen, S.; Herve du Penhoat, C.; Perez, S. *J. Phys. Chem.* **1995**, *99*, 13334–13351.

(45) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces, in Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981.

(46) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

(47) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(48) van Gunsteren, W. F.; Berendsen, H. J. C. *Angew. Chem., Int. Ed.* **1990**, *29*, 992–1023.

(49) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451–5459.

(50) van Gunsteren, W. F.; Berendsen, H. J. C. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 171–176.

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

(52) Ji, P.; Feng, W.; Tan, T. *J. Chem. Eng. Data* **2007**, *52*, 135–140.

(53) Mazzobre, M. F.; Román, M. V.; Mourelle, A. F.; Corti, H. R. *Carbohydr. Res.* **2005**, *340*, 1207–1211.

(54) Gessler, K.; Uson, I.; Takaha, T.; Krauss, N.; Smith, S. M.; Okada, S.; Sheldrick, G. M.; Saenger, W. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 4246–4251.

(55) Lu, Z.; Nowak, W.; Lee, G.; Marszalek, P. E.; Yang, W. *J. Am. Chem. Soc.* **2004**, *126*, 9033–9041.

(56) di Luzio, N. R.; Williams, D. L.; McNamee, R. B.; Edwards, B. F.; Kitahama, A. *Int. J. Cancer* **1979**, *24*, 773–779.

(57) Helbert, W.; Chanzy, H. *Int. J. Biol. Macromol* **1994**, *16*, 207–213.

(58) Maciejewska, W.; Kaczmarski, M. *J. Raman Spectrosc.* **1989**, *20*, 413–418.

(59) Deslandes, Y.; Marchessault, R. H.; Sarko, A. *Macromolecules* **1980**, *13*, 1466–1471.

(60) Okuyama, K.; Otsubo, A.; Fukuzawa, Y.; Ozawa, M.; Harada, T.; Kasai, N. *J. Carbohydr. Chem.* **1991**, *10*, 645–656.

(61) Chuah, C. T.; Sarko, A.; Deslandes, Y.; Marchessault, R. H. *Macromolecules* **1983**, *16*, 1375–1382.

(62) Montiel, P. *Cryo Lett.* **2000**, *21*, 83–90.

(63) Pereira, C. S.; Lins, R. D.; Chandrasekhar, I.; Freitas, L. C. G.; Hünenberger, P. H. *Biophys. J.* **2004**, *86*, 2273–2285.

(64) Marrink, S. J.; Risselada, J.; Mark, A. E. *Chem. Phys. Lipids* **2005**, *135*, 223–244.

(65) Pereira, C. S.; Hünenberger, P. H. *Mol. Simul.* **2008**, *34*, 403–420.

(66) Sum, A. K.; Faller, R.; de Pablo, J. J. *Biophys. J.* **2003**, *85*, 2830–2844.

(67) Kabayama, M. A.; Paterson, D. *Can. J. Chem.* **1958**, *36*, 563–573.

(68) Reeves, R. E. *J. Am. Chem. Soc.* **1950**, *72*, 1499–1506.

(69) Franks, F.; Lillford, P. J.; Robinson, G. *J. Chem. Soc., Faraday Trans.* **1989**, *85*, 2417–2426.

(70) Höög, C.; Landersjö, C.; Widmalm, G. *Chemistry* **2001**, *7*, 3069–3077.

(71) Vidal, P.; Vauzeilles, B.; Bieriot, Y.; Sollogoub, M.; Sinay, P.; Jimenez-Barbero, J.; Espinosa, J. F. *Carbohydr. Res.* **2007**, *342*, 1910–1917.

(72) DeMarco, M. L.; Woods, R. J. *Glycobiology* **2009**, *19*, 344–355.

CT900313W

JCTC Journal of Chemical Theory and Computation

# Versatile Object-Oriented Toolkit for Coarse-Graining Applications

Victor Rühle, Christoph Junghans, Alexander Lukyanov, Kurt Kremer, and Denis Andrienko*

*Max Planck Institute for Polymer Research, Ackermannweg 10, 55128 Mainz, Germany*

Received July 17, 2009

**Abstract:** Coarse-graining is a systematic way of reducing the number of degrees of freedom representing a system of interest. Several coarse-graining techniques have so far been developed, such as iterative Boltzmann inversion, force-matching, and inverse Monte Carlo. However, there is no unified framework that implements these methods and that allows their direct comparison. We present a versatile object-oriented toolkit for coarse-graining applications (VOTCA) that implements these techniques and that provides a flexible modular platform for the further development of coarse-graining techniques. All methods are illustrated and compared by coarse-graining the SPC/E water model, liquid methanol, liquid propane, and a single molecule of hexane.

## 1. Introduction

Computational materials science deals with phenomena covering a wide range of length- and time-scales from Ångstrøms (typical bond lengths) and femtoseconds (bond vibrations) to micrometers (crack propagation) and milliseconds (a single polymer chain relaxation). Depending on the characteristic time- and length-scales involved, the system description can vary from first principles and atomistic force fields to coarse-grained models and continuum mechanics. The role of bottom-up coarse-graining, in a broad sense, is to provide a systematic link between these levels of description.

Here we focus on coarse-graining techniques that link two particle-based descriptions with a different number of degrees of freedom. The system with the larger number of degrees of freedom we denote as the reference system. The system with the reduced number of the degrees of freedom is referred to as the coarse-grained system. An example is an all-atom (reference) and a united-atom (coarse-grained) molecular representation, where the number of the degrees of freedom is reduced by embedding hydrogens into heavier atoms.[55] Another example, which is treated in detail here, is an all-atom (three sites) and a single site model of water. Other examples can be readily found in the literature.[1−12]

We also assume that the following prerequisites are satisfied:

(i) Both the reference and the coarse-grained descriptions are represented by a set of point sites, $\boldsymbol{r} = \{\boldsymbol{r}_i\}$, $i = 1, 2, ..., n$, in case of the reference system, and $\boldsymbol{R} = \{\boldsymbol{R}_j\}$, $j = 1, 2, ..., N$, in case of the coarse-grained system.[56]

(ii) A mapping scheme, i.e., a relation between $\boldsymbol{r}$ and $\boldsymbol{R}$, can be expressed as $\boldsymbol{R} = \hat{\mathbf{M}}\boldsymbol{r}$, where $\hat{\mathbf{M}}$ is a $n \times N$ matrix.[57]

(iii) For the reference system, we have the coordinates and the forces of a trajectory that samples a canonical ensemble (or that part of it we are interested in reproducing on a coarse-grained level).

Then the prime task of systematic coarse-graining is to devise a potential energy function of the coarse-grained system, $U(\boldsymbol{R})$.

To do this, one can use several coarse-graining approaches. From the point of view of implementation, these approaches can be divided into iterative and noniterative methods. Boltzmann inversion is a typical example of a noniterative method.[1] In this method, which is exact for independent degrees of freedom, coarse-grained interaction potentials are calculated by inverting the distribution functions of the coarse-grained system. Another example of a noniterative method is force matching, where the coarse-grained potential

* Corresponding author. E-mail:denis.andrienko@mpip-mainz.mpg.de.

10.1021/ct900369w CCC: $40.75 © 2009 American Chemical Society
Published on Web 11/09/2009

is chosen in such a way that it reproduces the forces on the coarse-grained beads.[5,13] Configurational sampling,[14] which matches the potential of mean force, also belongs to this category. Boltzmann inversion and force matching only require a trajectory for a reference system.[58] Once that is known, coarse-grained potentials can be calculated for any mapping matrix $\hat{\mathbf{M}}$.

Iterative methods refine the coarse-grained potential $U(\boldsymbol{R})$ by reiterating coarse-grained simulations and by calculating corrections to the potential on the basis of the reference and the coarse-grained observables (e.g., radial distribution function or pressure). The simplest example is the iterative Boltzmann inversion method,[15] which is an iterative analogue of the Boltzmann inversion method. More sophisticated (in terms of the update function) is the inverse Monte Carlo approach.[16]

One can also classify systematic coarse-graining approaches by the micro- and macroscopic observables they use to derive the coarse-grained potential, such as structure-,[1,16,17] force-,[5,13,18] and potential-based approaches,[19] where the name identifies the observable used for coarse-graining. Note that hybrids of these methods are also possible.[3,12]

With a rich zoo of methods plus their combinations available at hand, it is natural to ask about an optimal method for a specific class of systems. On a more fundamental level, one might question whether the different methods provide the same coarse-grained potential and whether it is possible to formulate a set of (even empirical) rules favoring one method with respect to another. It is obvious this is a difficult task to be treated analytically, especially for realistic systems. To assess the quality of a particular coarse-graining technique, one needs to apply all available methods to a certain number of systems and to compare and quantify the degree of discrepancy between the coarse-grained and the reference descriptions. This is, however, cumbersome due to the absence of a single package where all these methods are implemented with the same accuracy and same level of technical detail.

The main aim of this work is to introduce such a coarse-graining package. The paper is organized as follows: We first describe the basic ideas behind each method, paying special attention to the technical issues one has to overcome when implementing them. We then illustrate these methods by coarse-graining systems of different complexities: a three-site SPC/E water, methanol, propane, and hexane.

## 2. Methods

Before starting with brief recapitulations of the coarse-graining methods, we refer the reader to a (far from complete) list of reviews which cover various aspects of generating coarse-grained potentials.[20−26]

**2.1. Boltzmann Inversion.** Boltzmann inversion is the simplest method one can use to obtain coarse-grained potentials.[1] It is mostly used for bonded potentials, such as bonds, angles, and torsions. Boltzmann inversion is structure-based and only requires positions of atoms.

The idea of Boltzmann inversion stems from the fact that in a canonical ensemble independent degrees of freedom $q$ obey the Boltzmann distribution, i. e.:

$$P(q) = Z^{-1}\exp[-\beta U(q)] \tag{1}$$

where $Z = \int \exp[-\beta U(q)]\,\mathrm{d}q$ is a partition function, $\beta = 1/k_{\mathrm{B}}T$. Once $P(q)$ is known, one can invert eq 1 and obtain the coarse-grained potential, which, in this case, is a potential of mean force:

$$U(q) = -k_{\mathrm{B}}T\ln P(q) \tag{2}$$

Note that the normalization factor $Z$ is not important since it would only enter the coarse-grained potential $U(q)$ as an irrelevant additive constant.

In practice, $P(q)$ is computed from the trajectory of the reference system, which is sampled either by Monte Carlo, molecular dynamics, stochastic dynamics, or any other integrator that ensures a canonical distribution of states.

Boltzmann inversion is simple to implement, however, one has to be careful with the rescaling of the probability $P$ due to orientational entropy as well as computational issues. The probability rescaling can be explained on a particular example of coarse-graining of a single polymer chain by beads with bond, angle and torsion potentials. In this case the coarse-grained potential $U$ depends on three variables, bond length $r$, angle $\theta$, and torsion angle $\varphi$.

Assuming, as before, a canonical distribution and independence of the coarse-grained degrees of freedom, we can write:

$$P(r,\theta,\varphi) = \exp[-\beta U(r,\theta,\varphi)] \tag{3}$$

$$P(r,\theta,\varphi) = P_r(r)P_\theta(\theta)P_\varphi(\varphi) \tag{4}$$

If we now compute the histograms for the bonds $H_r(r)$, angle $H_\theta(\theta)$, and torsion angle $H_\varphi(\varphi)$, then we must rescale them in order to obtain the volume normalized distribution functions.[59]

$$P_r(r) = \frac{H_r(r)}{4\pi r^2},\quad P_\theta(\theta) = \frac{H_\theta(\theta)}{\sin\theta},\quad P_\varphi(\varphi) = H_\varphi(\varphi) \tag{5}$$

The coarse-grained potential can then be calculated by Boltzmann inversion of the distribution functions:

$$\begin{aligned} U(r,\theta,\varphi) &= U_r(r) + U_\theta(\theta) + U_\varphi(\varphi) \\ U_q(q) &= -k_{\mathrm{B}}T\ln P_q(q), \qquad q = r,\theta,\varphi \end{aligned} \tag{6}$$

On the technical side, the implementation of the Boltzmann inversion method requires smoothing of $U(q)$ to provide a continuous force. Splines can be used for this purpose. Poorly and unsampled regions, that is regions with high $U(q)$, shall be extrapolated. Since the contribution of these regions to the canonical density of states is small, the exact shape of the extrapolation is less important.

Another crucial issue is the cross-correlation of the coarse-grained degrees of freedom. Independence of the coarse-grained degrees of freedom is the main assumption that allows factorization of the probability distribution, eq 4, and the potential, eq 6, hence, one has to carefully check whether this assumption holds in practice. This can be done by performing coarse-grained simulations and by comparing cross-correlations for all pairs of degrees of freedom in atomistic and coarse-grained resolution, e.g., using a two-dimensional histogram, analogous to a Ramachandran plot.[60]

**2.2. Iterative Boltzmann Inversion.** Iterative Boltzmann inversion (IBI) is a natural extension of the Boltzmann inversion method. Since the goal of the coarse-grained model is to reproduce the distribution functions of the reference system as accurately as possible, one can also iteratively refine the coarse-grained potentials using some numerical scheme. Depending on the update function, this can be done by using either the iterative Boltzmann inversion[15] or the inverse Monte Carlo[16,17] method. We will first discuss the iterative Boltzmann inversion method.

In the iterative Boltzmann inversion, the coarse-grained potential is refined according to the following scheme:[61]

$$\begin{aligned} U^{(n+1)} &= U^{(n)} + \Delta U^{(n)} \\ \Delta U^{(n)} &= k_B T \ln \frac{P^{(n)}}{P_{\text{ref}}} = U_{\text{PMF}}^{\text{ref}} - U_{\text{PMF}}^{(n)} \end{aligned} \tag{7}$$

One can easily see that convergence is reached as soon as the distribution function $P^{(n)}$ matches the reference distribution function $P_{\text{ref}}$, or, in other words, the potential of mean force, $U_{\text{PMF}}^{(n)}$ converges to the reference potential of mean force.

IBI can be used to refine both bonded and nonbonded potentials. It is primarily used for simple fluids with the aim of reproducing the radial distribution function of the reference system in order to obtain nonbonded interactions.[15] It can have convergence problems for multicomponent systems, since it does not account for cross-correlation correction terms, that is the updates for $P_{\text{AA}}$, $P_{\text{AB}}$, and $P_{\text{BB}}$ are not coupled (the subscript enumerates a single component in a multicomponent system). For such systems, the inverse Monte Carlo method works better. The scheme can be stabilized by multiplying the update function, $\Delta U^{(n)}$, by a factor $\eta \in [0..1]$.

On the implementation side, IBI has the same issues as the inverse Boltzmann method, i.e., smoothing and extrapolation of the potential must be implemented.

We shall also mention that, according to the Henderson theorem,[25,27] which is a classical analogue of the Hohenberg–Kohn theorem, the pairwise coarse-grained potential $U(r)$ is unique up to an additive constant and exists,[28,29] which, in principle, states that all structure-based iterative methods must converge to the same coarse-grained potential, provided that their aim is to exactly reproduce pair correlation functions of the reference system. As we will see later, this is often not the case in practice, since small changes in the radial distribution function often lead to big changes in the pair potential, i.e., it is difficult to control systematic errors during the calculation of the potential update.

Another issue of coarse-graining is that coarse-grained models cannot reproduce all the statistical or thermodynamic properties of the reference system. Pressure, compressibility, or viscosity[30] are often very different from those of the reference system. In some cases, however, one can correct for some of these. For example, the viscosity can be adjusted by tuning the parameters of the thermostat,[31] and the pressure can be corrected iteratively by adding a linear term to the nonbonded potential:

$$\Delta U^{\text{pressure}} = A\left(1 - \frac{r}{r_{\text{cut}}}\right) \tag{8}$$

where $A$ is either a constant, e.g., $-0.1k_BT$,[15] or can be estimated from the virial expansion.[32] Compressibility and pressure, however, cannot be corrected simultaneously.

**2.3. Inverse Monte Carlo.** Inverse Monte Carlo (IMC) is another iterative procedure that refines the coarse-grained potentials until the coarse-grained model reproduces a set of reference distribution functions. It is very similar to IBI except that the update of the potential, $\Delta U$, is calculated using rigorous thermodynamic arguments.

The name “inverse Monte Carlo” is somehow confusing and is due to the fact that the original algorithm was combined with Monte Carlo sampling of the phase space.[16] However, practically any sampling method can be used (e.g., molecular or stochastic dynamics) as long as it provides a canonical sampling of the phase space.

A detailed derivation of the IMC method can be found in ref 16. Here we briefly recapitulate the more compact version for nonbonded interactions, which is outlined in ref 25 emphasizing technical problems encountered during implementation and application of the method.

The idea of IMC is to express the potential update $\Delta U$ in a thermodynamically consistent way in terms of measurable statistical properties, e.g., radial distribution function $g(r)$. Considering a single-component system as an example, we can write the Hamiltonian of the system as

$$H = \sum_{i,j} U(r_{ij}) \tag{9}$$

where $U(r_{ij})$ is the pair potential, and we assume that all interactions depend only on the distance, $r_{ij}$, between particles $i$ and $j$. We further assume that this potential is short-ranged, i.e., $U(r_{ij}) = 0$, if $r_{ij} \geq r_{\text{cut}}$.

The next step is to tabulate the potential $U(r)$ on a grid of $M$ points, $r_\alpha = \alpha\Delta r$, where $\alpha = 0, 1, ..., M$, and $\Delta r = r_{\text{cut}}/M$ is the grid spacing. Then the Hamiltonian, eq 9, can be rewritten as

$$H = \sum_{\alpha} U_\alpha S_\alpha \tag{10}$$

where $S_\alpha$ is the number of particle pairs with interparticle distances $r_{ij} = r_\alpha$, which correspond to the tabulated value of the potential $U_\alpha$.

On one hand, the average value of $S_\alpha$ is related to the radial distribution function $g(r)$:

$$\langle S_\alpha \rangle = \frac{N(N-1)}{2} \frac{4\pi r_\alpha^2 \Delta r}{V} g(r_\alpha) \tag{11}$$

where $N$ is the number of atoms in the system, ((1/2)$N(N-1)$ is then the number of all pairs), $\Delta r$ is the grid spacing, $r_{\text{cut}}/M$, and $V$ is the total volume of the system.

On the other hand, $\langle S_\alpha \rangle$ is a function of the potential $U_\alpha$ and, hence, can be expanded in a Taylor series with respect to small perturbations of $U_\alpha$, $\Delta U_\alpha$

$$\Delta\langle S_\alpha\rangle = \sum_\gamma \frac{\partial\langle S_\alpha\rangle}{\partial U_\gamma}\Delta U_\gamma + \mathcal{O}(\Delta U^2) \tag{12}$$

The derivatives $\partial\langle S_\alpha\rangle/\partial U_\gamma$ can be obtained by using the chain rule:

$$\begin{aligned} A_{\alpha\gamma} &= \frac{\partial\langle S_\alpha\rangle}{\partial U_\gamma} \\ &= \frac{\partial}{\partial U_\gamma}\frac{\int dq S_\alpha(q)\exp[-\beta\sum_\lambda U_\lambda S_\lambda(q)]}{\int dq \exp[-\beta\sum_\lambda U_\lambda S_\lambda(q)]} \\ &= \beta(\langle S_\alpha\rangle\langle S_\gamma\rangle - \langle S_\alpha S_\gamma\rangle) \end{aligned} \tag{13}$$

Equations 11−13 allow us to calculate the correction for the potential by solving a set of linear equations:

$$\langle S_\alpha\rangle - S_\alpha^{\text{ref}} = A_{\alpha\gamma}\Delta U_\gamma \tag{14}$$

where $S_\alpha^{\text{ref}}$ is given by the target radial distribution function. The procedure is then repeated until convergence is reached.

A clear advantage of the IMC compared to the IBI method is that the update of the potential is rigorously derived using statistical mechanics, and hence, the iterative procedure shall converge faster with the IMC update than with the empirical IBI update. Another advantage is that, in the case of multicomponent mixtures, IMC takes into account correlations of observables, that is updates for $U_{\text{AA}}$, $U_{\text{AB}}$, and $U_{\text{BB}}$ are interdependent (A and B denote different particle types). In the IBI method, these updates are independent which often leads to convergence problems for multicomponent systems.

The advantages come, of course, at a computational cost. As it is clear from eq 13, one has to calculate cross-correlations of $S_\alpha$. This requires much longer runs to get statistics that are good enough to calculate the potential update to a similar accuracy as IBI. The accuracies of the update functions of IMC and IBI methods are compared in Section 4.1 for the case of a coarse-grained model of water.

Another issue of the IMC method is the stability of the scheme. Several factors can influence it: the first, and rather technical, point is that $g^{\text{ref}}(r_\alpha)$ has to be calculated using exactly the same convention for the grid as $S_\alpha$ (e.g., the function value should be assigned to the middle of the interval), otherwise the scheme becomes unstable. Second, inversion of $A_{\alpha\gamma}$ requires that it shall be well-defined. This means that one has to remove the regions which are not sampled, such as those at the beginning of the radial distribution function. The convergence can be significantly improved if a smoothing of the potential update $\Delta U$ is used. Note that it is better to do smoothing of the update function, not the potential itself, since the latter has more features which can be lost due to too aggressive smoothing. The convergence can also be improved by introducing a multiplicative prefactor for the update function or by using a regularization procedure by adding thermodynamic constraints.[33]

Finally, we have also noticed that the systematic error in $\langle S_\alpha S_\gamma\rangle$ is always higher in the vicinity of the cutoff, which leads to a shift in the tail of the interaction potential and, as a result, to a large offset of pressure. The cross-correlation term $\langle S_\alpha S_\gamma\rangle$ is also very sensitive to the box size, and special care must be taken in order to converge the results with respect to system size. Finite size effects are discussed in detail in Section 4.2, where we coarse-grain liquid methanol.

**2.4. Force Matching.** Force matching (FM) is another approach to evaluate corse-grained potentials.[5,13,34] In contrast to the structure-based approaches, its aim is not to reproduce various distribution functions, but instead try to match forces on coarse-grained beads as closely as possible.[62] FM is a noniterative method and, hence, is less computationally demanding.

The method works as follows: we first assume that the coarse-grained force field (and hence the forces) depends on $M$ parameters $g_1, ..., g_M$. These parameters can be prefactors of analytical functions, tabulated values of the interaction potentials, or coefficients of splines used to describe these potentials.

In order to determine these parameters, the reference forces on coarse-grained beads are calculated by properly reweighting the forces on the atoms:

$$\boldsymbol{f}_i^{\text{ref}} = M_i \sum_\alpha \frac{w_\alpha \boldsymbol{f}_\alpha}{m_\alpha} \tag{15}$$

where $M_i = (\sum_\alpha w_\alpha{}^2/m_\alpha)^{-1}$ is the mass of the bead $i$, index $\alpha$ numbers all atoms belonging to this bead, $\boldsymbol{f}_\alpha$ is the force on the atom $\alpha$, $m_\alpha$ is its mass, $w_\alpha$ are mapping coefficients used to obtain the position of the coarse-grained bead, $\boldsymbol{R}_i = \sum_\alpha w_\alpha \boldsymbol{r}_\alpha$. If the center of mass is used in the mapping, then eq 15 simplifies to the sum of the forces.

By calculating the reference forces for $L$ snapshots, we can write down $N \times L$ equations:

$$\boldsymbol{f}_{il}^{\text{cg}}(g_1, ..., g_M) = \boldsymbol{f}_{il}^{\text{ref}}, \quad i = 1, ..., N, \quad l = 1, ..., L \tag{16}$$

Here $\boldsymbol{f}_{il}^{\text{ref}}$ is the force on the bead $i$, $\boldsymbol{f}_{il}^{\text{cg}}$ is the coarse-grained representation of this force. Index $l$ enumerates snapshots picked for coarse-graining. By running the simulations long enough one can always ensure that $M < N \times L$. In this case, the set of eqs 16 is overdetermined and can be solved in a least-squares sense.

Though the underlying idea of FM is very simple, implementation-wise it is the most complicated method. Here we briefly outline the problems, which are then discussed in more detail in Appendix A.

Going back to the set of eqs 16, one can see that $\boldsymbol{f}_{il}^{\text{cg}}$ is, in principle, a nonlinear function of its parameters $\{g_i\}$. It is, therefore, useful to represent the coarse-grained force field in such a way that eqs 16 become linear functions of $\{g_i\}$. This can be done using splines to describe the functional form of the forces.[5]

An adequate sampling of the system requires a large number of snapshots $L$. Hence, the applicability of the method is often constrained by the amount of available memory. To remedy the situation, one can split the trajectory into blocks, find the coarse-grained potential for each block and then perform averaging over the blocks. More details on the technical implementation of force matching using cubic splines is given in Appendix A.

## 3. Implementation

**3.1. Coarse-Graining Engine.** In a nutshell, coarse-graining is nothing more than an analysis of the canonical ensemble of a reference (high resolution) system. In addition to this analysis, iterative methods require canonical sampling of the coarse-grained system, which can be done using either molecular dynamics (MD), stochastic dynamics (SD), or Monte Carlo (MC) techniques. The latter are implemented in many standard simulation packages. Rather than implementing its own MD/SD/MC modules, the toolkit allows swift and flexible integration of existing programs in such a way that sampling is performed in the program of choice. Only the analysis needed for systematic coarse-graining is done using the package tools.

The tools include calculations of probability distributions of bonded and nonbonded interactions, correlation and autocorrelation functions, and updates for the coarse-grained pair potential. Analysis tools of the MD package can also be integrated into the coarse-graining workflow, if needed.

The package offers a flexible framework for reading, manipulating, and analyzing of MD/SD/MC topologies and trajectories. Its core is modular, and new file formats can be integrated without changing the existing code. At the moment, an interface for GROMACS[35] topologies and trajectories is provided. An interface to ESPResSo++[63] is planned.

The coarse-graining procedure itself is controlled by several extensible markup language (XML) input files, which contain mapping and other options required for the workflow control. In the mapping, it is possible to select groups of interactions which will be used for coarse-graining or analysis.

**3.2. Iterative Workflow Control.** The workflowchart is shown in Figure 1. The workflow is implemented as a shell script which can, in principle, be run on all available operating systems and provides the flexibility needed to call external (or overload existing) scripts and programs written in other programming languages. An interface to read values from the steering XML files in C++, Perl, and shell is also provided.

During the global initialization, the initial guess for the coarse-grained potential is calculated from the reference radial distribution function or converted from a given potential guess to the internal format. The actual iterative step starts with an iteration initialization. It searches for possible checkpoints and copies and converts files from the previous step and the base directory. Then the simulation run is prepared by converting potentials to the format required by the external sampling program, and actual sampling is performed. Currently, an interface with GROMACS[35] is implemented, and an extension to other packages is straightforward. After sampling the phase space, potential update $\Delta U$ is calculated. Often the update requires postprocessing, such as smoothing, interpolation, extrapolation, or fitting to an analytical form. A simple pressure correction[15] can also be seen as a postprocessing of $\Delta U$ due to the fact that it only adds a linear interparticle separation function. Finally, the new potential is determined and postprocessed. If the iterative process continues, then the next iterative step starts to initialize.

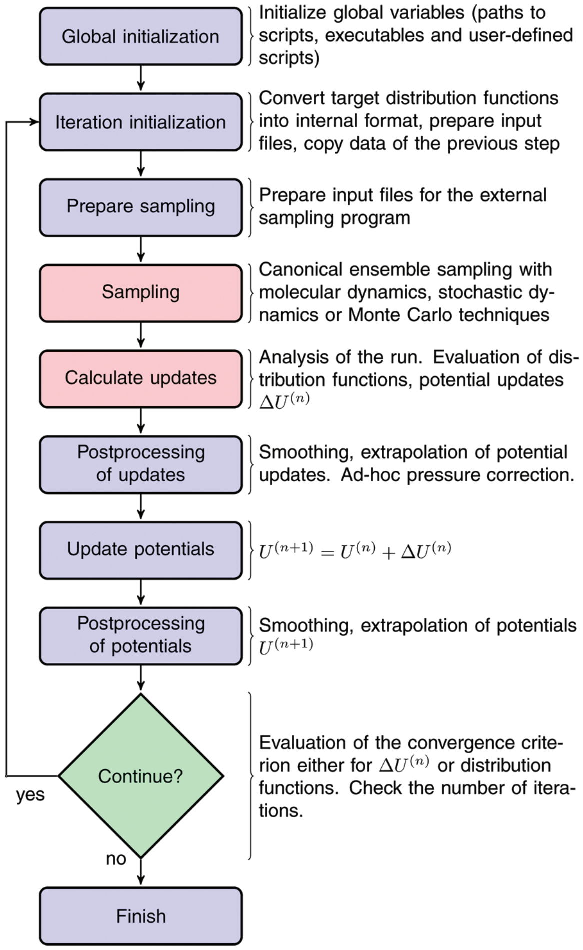

Global initialization
Initialize global variables (paths to scripts, executables and user-defined scripts)
Iteration initialization
Convert target distribution functions into internal format, prepare input files, copy data of the previous step
Prepare sampling
Prepare input files for the external sampling program
Sampling
Canonical ensemble sampling with molecular dynamics, stochastic dynamics or Monte Carlo techniques
Calculate updates
Analysis of the run. Evaluation of distribution functions, potential updates $\Delta U^{(n)}$
Postprocessing of updates
Smoothing, extrapolation of potential updates. Ad-hoc pressure correction.
Update potentials
$U^{(n+1)} = U^{(n)} + \Delta U^{(n)}$
Postprocessing of potentials
Smoothing, extrapolation of potentials $U^{(n+1)}$
Continue?
yes
no
Evaluation of the convergence criterion either for $\Delta U^{(n)}$ or distribution functions. Check the number of iterations.
Finish

***Figure 1.*** Block-scheme of the workflow control for the iterative methods. The most time-consuming parts are marked in red.

## 4. Examples

We illustrate the package functionality using four systems: SPC/E water, liquid methanol, liquid propane, and a single chain of hexane. The systems are chosen in such a way that the corresponding coarse-grained potentials have already been obtained using one or more techniques, providing a good reference point for comparison.

**4.1. Coarse-Graining of Water.** Water is one of the most studied liquids from the point of view of both all-atom representations and coarse-grained models.[36,37] Here we coarse-grain one of the all-atom models of water, the SPC/E[38,39] water model. The corresponding parameters of this three-site model are given in the caption to Figure 2. Note that this is a rigid model, i.e., the distances between two hydrogens as well as oxygen and hydrogens are constrained during the molecular dynamics runs. For the coarse-grained representation, we use a one-site representation with a pair potential $U(R_{ij})$, where $R_{ij}$ connects the centers of mass of water molecules $i$ and $j$.

The all-atom system consisting of 2180 water molecules was first equilibrated in the NPT ensemble at 300K and 1

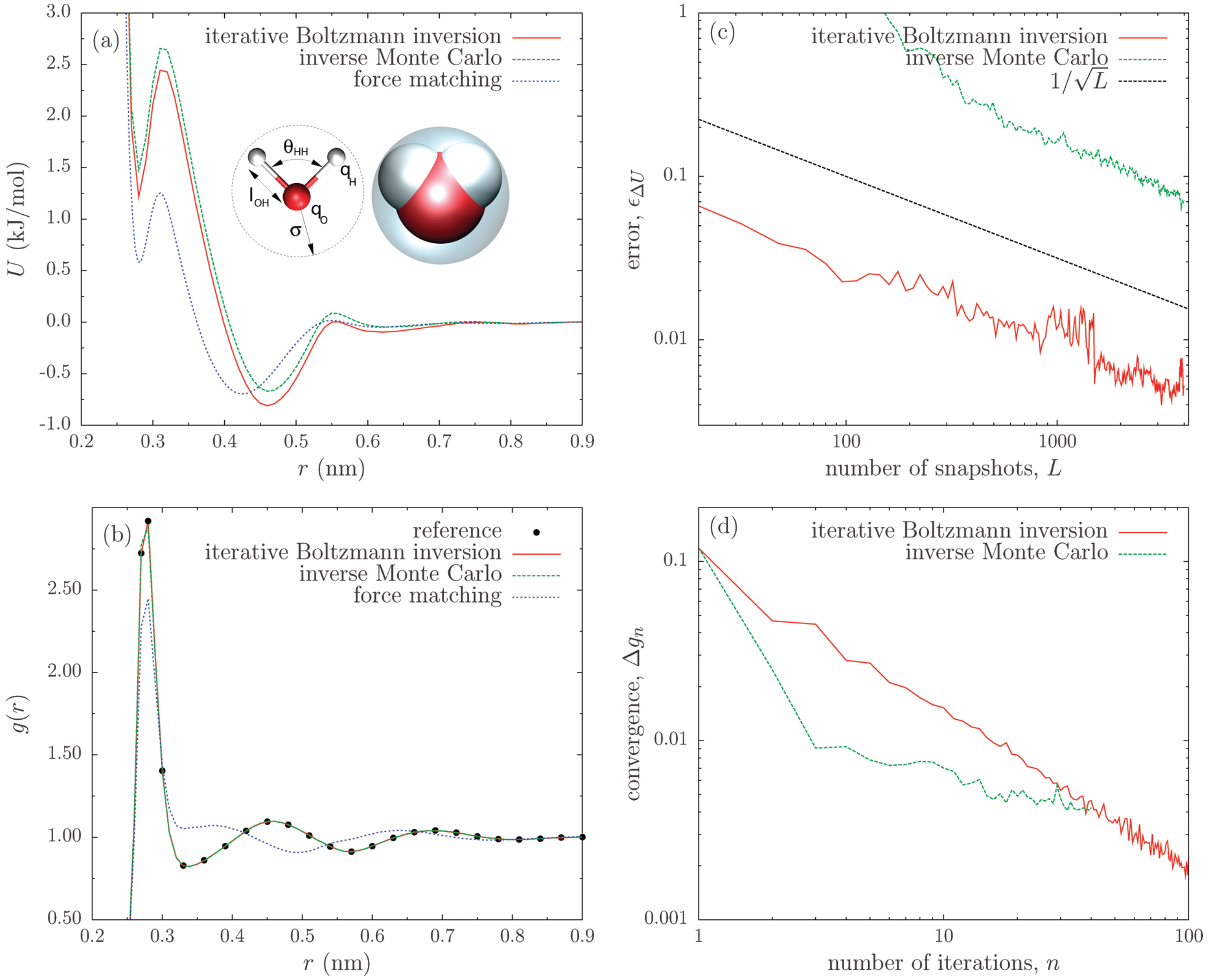


***Figure 2.*** Water: (a) Coarse-grained potentials for SPC/E water obtained using different coarse-graining techniques. (b) Corresponding radial distribution functions. (c) Average error of the potential update function versus number of snapshots used for calculating the update function. (d) Root-mean-square deviation of reference and current radial distribution function versus iteration step. One can see that IMC converges faster than that of IBI. Inset of (a) shows van der Waals excluded volume and coarse-grained representations of a single water molecule as well as parameters used: $\sigma = 3.166$ Å, $\varepsilon = 0.650$ kJ mol$^{-1}$, $l_{OH} = 1.0000$ Å, $q_H = +0.4238e$, $q_O = -0.8476e$, $\theta_{HH} = 109.47°$.

bar for 100 ns using the Berendsen thermostat and barostat.[40] The last 80 ns were used to determine the equilibrium box size of 4.031 nm, which was then fixed during the 45 ns production run in the NVT ensemble using a stochastic dynamics algorithm.[41] For all further analysis, only the last 40 ns were used. The radial distribution function was calculated using a 0.01 nm grid spacing. The snapshots were output every 0.4 ps.

Force matching potentials were calculated using blocks of six snapshots each. Spline grid spacing of 0.02 nm was used in the interval from 0.24 to 1 nm. For the iterative procedures, the potential of mean force was taken as an initial guess for the interaction potential. The coarse-grained box had the same system size as in the atomistic simulations. Simulations of the coarse-grained liquid were done using a stochastic dynamics algorithm.[41] When using IBI, 300 iterations of 100 ps each were performed. For IMC, we used 10 iterations of 500 ps each. Additionally, two iterations of triangular smoothing were applied to the IMC potential update, $\Delta U$. The cutoff was chosen at 0.9 nm with a grid spacing of 0.01 nm.

The reference radial distribution function, $g^{\text{ref}}(r)$, coarse-grained potentials, and corresponding radial distribution functions are shown in Figure 2a,b. IBI and IMC give practically the same interaction potential. Although the force-matched potential has a very similar structure with two minima, the corresponding radial distribution function is very different from the target one. Possible reasons for these discrepancies are discussed in refs 23, 25, and 34, and stem from the fact that FM aims to reproduce the many-body potential of mean force, which does not necessarily guarantee perfect pairwise distribution functions, considering the fact that the basis sets in the coarse-grained force field may be limited.

Note that all three methods lead to a different pressure of the coarse-grained system: 8000 bar (IBI), 9300 bar (IMC), and 6500 bar (FM). Different pressures for the iterative methods are due to a different accuracy of the potential update. Indeed, small changes of pressure can significantly affect the potential, especially its long tail.[15,42] However, they hardly change the radial distribution function due to the small compressibility of water. One can improve the

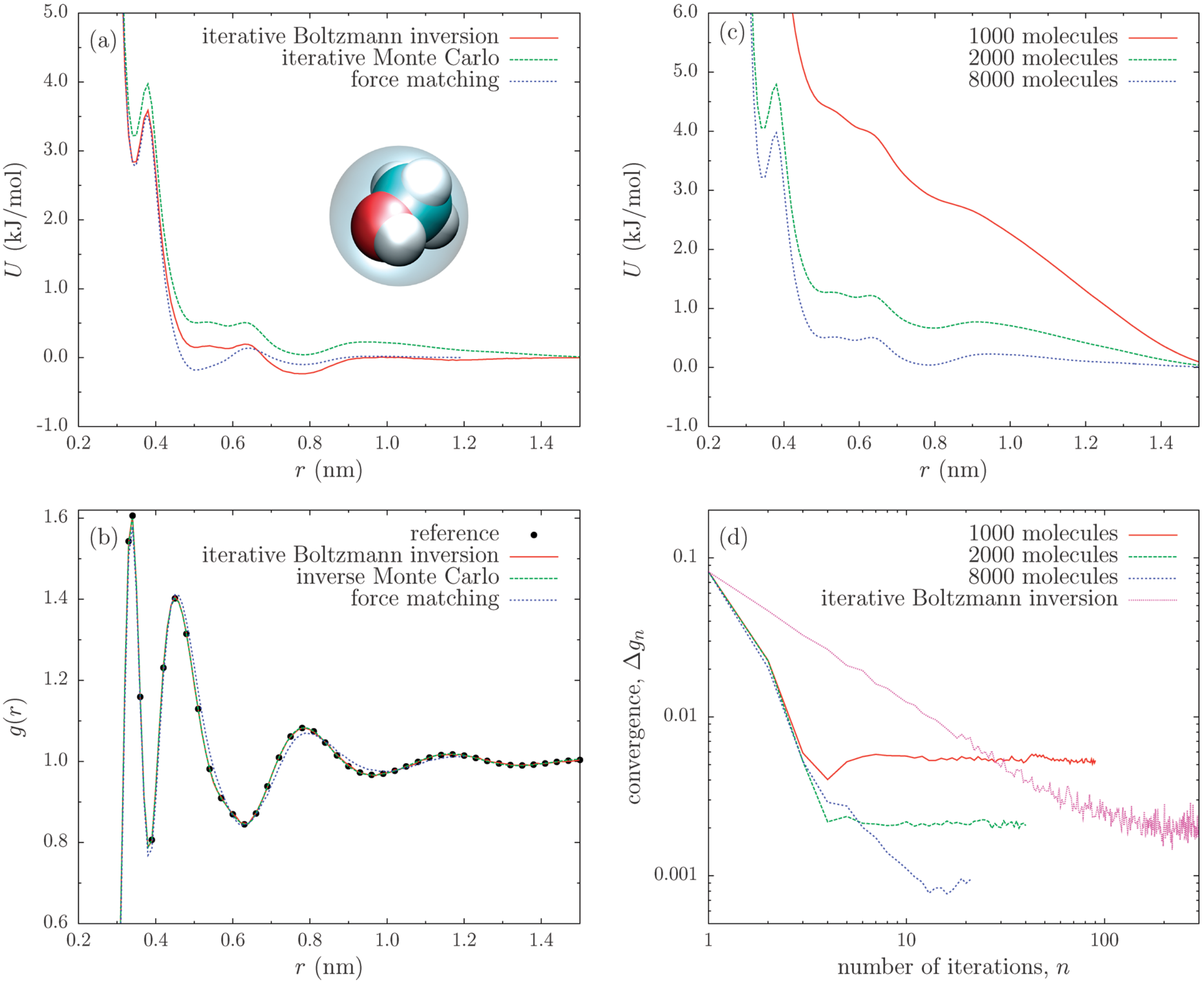


***Figure 3.*** Methanol: (a) Coarse-grained potentials. (b) Corresponding radial distribution functions. (c) coarse-grained potentials using 10 IMC iterations for simulation boxes with 1000, 2000, and 8000 methanol molecules (box size 4.09, 5.15308, and 8.18 nm) equilibrated at the same density. (d) Root-mean-square deviation of reference and the current radial distribution function versus number of iterations. Similar to liquid water, IMC converges faster than IBI. The convergence saturates and the saturation error strongly depends on the system size. The inset of (a) shows the van der Waals excluded volume and coarse-grained representations of a methanol molecule.

agreement between the iterative methods by using pressure correction terms for the update.

The performance of the iterative methods depends on two factors: (i) the average (over all bins) error of the potential update $\varepsilon_{\Delta U}$; and (ii) the number of iterations required for convergence. We define the average error as

$$\varepsilon_{\Delta U} = \frac{1}{N}\sum_{i=0}^{N} \varepsilon(\Delta U(r_i)) \tag{17}$$

where $N$ is the number of bins and $\varepsilon(\Delta U(r_i))$ is the error of the update function at a separation $r_i$. $\varepsilon(\Delta U(r_i))$ was calculated using a Jackknife analysis.[43]

The average error of the potential update is shown in Figure 2c as a function of the run length. One can see that, for both methods, the error decreases as $1/\sqrt{L}$, where $L$ is the number of snapshots used for averaging. However, the prefactor for the IBI update error, which is based on the radial distribution function, is at least 10 times smaller than of the IMC update error, which makes use of cross-correlations of $S_\alpha$. This observation implies that, in order to have the same accuracy of the update function, IMC needs significantly longer sampling.

This disadvantage is, of course, compensated by the efficiency of the update function, which is assessed by computing the root-mean-square deviation, $\Delta g_n$, of the current and target radial distribution functions:

$$\Delta g_n^2 = \int [g^{\text{ref}}(r) - g^{(n)}(r)]^2\, \mathrm{d}r \tag{18}$$

$\Delta g_n$ is plotted as a function of the number of iterations, $n$, in Figure 2d. It is clear that IMC converges much faster than IBI, though the root-mean-square deviation saturates after some number of iterations.

**4.2. Coarse-Graining of Methanol.** Liquid methanol (see the inset in Figure 3) is the second example of coarse-graining of nonbonded interactions that we present here. In fact, FM has already been used to coarse-grain this system,[42] and contrary to water, the liquid structure (radial distribution function) is well reproduced by the FM coarse-grained potential. In addition, the excluded volume of methanol is larger than that of water, and the undulations of the radial distribution function extend up to 1.5 nm. As we will see, this leads to pronounced finite size effects for IMC, since it has a nonlocal potential update. FM and IBI do not have this problem, since the IBI potential energy update is local, and FM is based on pair forces. The range of the latter is much shorter than the correlation length of structural properties (such as undulations of the radial distribution function), which may propagate over the boundaries for small boxes.

Simulation parameters were taken from ref 42, and OPLS[44,45] all-atom force field was used. Atomistic simulations were performed with 1000 methanol molecules in a cubic box (4.09 nm box size) at 300K using the Nosé−Hoover thermostat.[46,47] The system was equilibrated for 2 ns followed by a production run of 18 ns. The reference radial distribution function was calculated using snapshots every 0.5 ps and is shown in Figure 3b.

The FM potential was calculated using blocks of six frames each and using a spline grid of 0.02 nm. With this potential, coarse-grained simulations were performed using a stochastic dynamics integrator and using 1000 beads with the same box size and the same temperature as in the atomistic simulations. The system was equilibrated for 40 ps followed by a production run of 160 ps. Snapshots were stored every 5 ps and used to calculate the radial distribution function.

For the iterative procedures, the potential of mean force was taken as an initial guess. The cutoff was chosen at 1.54 nm with a grid spacing of 0.01 nm. For IBI, 300 iterations were performed using stochastic dynamics with the same parameters used in the FM-based procedure. The IMC iterations were performed with 8000 molecules and a box size of 8.18 nm. The total length of the run was 1 ns, and snapshots were stored every 0.2 ps. Two smoothing steps were used at each iteration for the potential update, $\Delta U$.

The coarse-grained potentials for all methods are shown in Figure 3a. In spite of small differences between the coarse-grained potentials, the agreement between the reference and the coarse-grained radial distribution functions is excellent, as can be seen from Figure 3b.

It is important to mention that the IMC method, which has a nonlocal update, is prone to systematic errors due to finite size effects and, hence, requires much larger simulation boxes in order to calculate the potential update. This is due to artificial cross-correlations of $S_\alpha$ at large distances, which lead to a small difference of tails between the coarse-grained and the reference radial distribution functions, and, as a consequence, to a much higher pressure of the coarse-grained system and a significantly different coarse-grained potential. In contrast, IBI and FM work well with system sizes of the order of two radial distribution function cutoff lengths.

To illustrate this point, we prepared simulation boxes of three different sizes, with 1000, 2000, and 8000 methanol molecules (box size of 4.09, 5.15308, and 8.18 nm and simulation times of 3, 2, and 1 ns, respectively). The IMC iterative procedure was repeated until the potentials converged, and these are shown in Figure 3c. One can see that the potentials significantly differ from each other. These differences lead to small deviations in the tail of the radial distribution function, which, however, vanish in a systematic way for bigger boxes, as illustrated in Figure 3d where we plot the integral of the difference of the reference and the current distribution functions.[64]

To summarize, IMC should be used with care for small systems. The potential update (or the coarse-grained potential) must be converged with respect to the simulation box size. In the case of methanol coarse-graining, a box of size three times the radial distribution function cutoff was not enough to achieve the converged potential for IMC, even though this is sufficient for IBI and FM methods.

**4.3. Liquid Propane: From an All- To an United-Atom Description.** So far we have illustrated coarse-graining of nonbonded degrees of freedom using liquid water and methanol as examples. Here we show how bonded interactions can be coarse-grained by deriving a united-atom model (i.e., hydrogens embedded into heavier atoms) from an all-atom model of liquid propane.[65] The mapping scheme as well as the bonded coarse-grained variables (two bonds, *b*, and one angle, $\theta$) are shown in the inset of Figure 4. Note that this coarse-graining scheme has two different bead types: an inner bead, of type B, with two hydrogens, and two outer beads, of type A, with three hydrogens. As a result, three types of nonbonded interactions, $U_{AA}$, $U_{BB}$, and $U_{AB}$ must be determined.

As before, atomistic simulations were performed using the OPLS all-atom force field.[44,45] A box of liquid propane was first equilibrated at 200K and 1 bar in the NPT ensemble for 10 ns, using the Berendsen thermostat and barostat.[40] The equilibrated box of the size 4.96337 × 5.13917 × 4.52386 $nm^3$ was then simulated for 10 ns in the NVT ensemble at 200K using velocity rescaling.[48] No bond constraints were used during the simulations, and hence, the integration time step was 1 fs. Snapshots were written every 1 ps.

In the case of iterative methods, the bonded potentials (bond and angle) were calculated by Boltzmann-inverting the corresponding distribution functions of a single molecule in vacuum, according to eq 5. The propane molecule in vacuum was simulated in an stochastic dynamics run[41] for 100 ns with snapshots stored every 2 ps. Nonbonded potentials were iteratively refined by using IBI with a grid spacing of 0.01 nm and a cutoff of 1.36 nm (1.38 nm) for A−A, A−B (B−B) interaction types, respectively. The run length for each iteration was 50 ps with snapshots written every 0.5 ps. At every iteration step, only one interaction type was corrected. When using the FM method, both bonded and nonbonded potentials were obtained at the same time, since FM does not require the explicit separation of bonded and nonbonded interactions.

The obtained potentials are shown in Figure 4a, c, and d. FM and Boltzmann inversion-derived bond and angle potentials (Figure 4c and d) perfectly agree with each other. The nonbonded potentials, shown in Figure 4a, are not completely identical but have similar shapes and barrier heights. This, of course, results in a good reproducibility of the propane liquid structure by the FM-based coarse-grained potentials, as can bee seen from the radial distribution functions shown in Figure 4b. Again, as expected, IBI reproduces the reference radial distribution functions exactly.

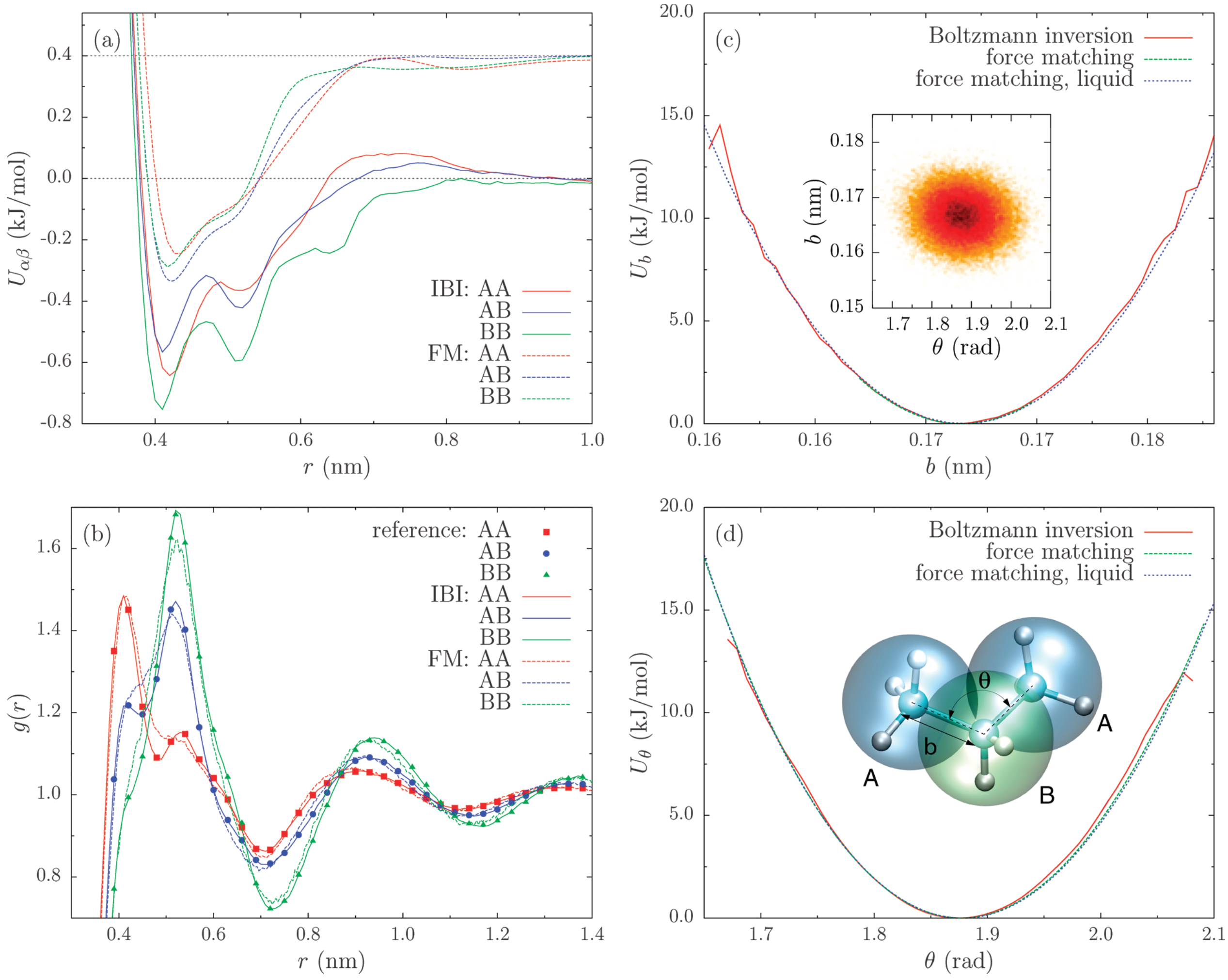


***Figure 4.*** Propane: (a) Nonbonded interaction potentials $U_{\mathrm{AA}}$, $U_{\mathrm{BB}}$, and $U_{\mathrm{AB}}$ obtained with IBI and FM methods. For clarity, FM potentials are offset along the *y*-axis. (b) Corresponding radial distribution functions plotted together with the atomistic radial distribution function. (c) Bond potential obtained for a single molecule in vacuum by Boltzmann-inverting the corresponding distribution function, using FM for a single propane molecule in vacuum and using force matching for liquid propane. (d) Angular coarse-grained potentials. The inset of (c) shows the correlations of $b$ and $\theta$. The inset of (d) shows all-atom and coarse-grained representations of a propane molecule, bead types, and coarse-grained bonded degrees of freedom (bond $b$ and angle $\theta$).

To summarize, the united-atom model of liquid propane is an ideal example of coarse-graining where the structure- and force-based methods result in similar bonded and nonbonded interaction potentials. As we will see later, this is due to: (i) the completeness of the basis set used to construct the coarse-grained force field; and (ii) independence of bond and angular degrees of freedom. The latter can be understood with the help of a histogram showing the correlation of $b$ and $\theta$, depicted in the inset of Figure 4c.

In the next section, we will look at coarse-graining of a single molecule of hexane, for which this is not the case.

**4.4. Angular Potential of a Hexane Molecule.** The final example we would like to discuss here is the angular potential of a hexane coarse-grained into a three-bead chain, with two carbon atoms per bead (see the inset in Figure 5a). Atomistic simulations of a single hexane molecule in vacuum were performed using an all-atom OPLS force field and a stochastic dynamics integrator.[41] The run length was 1000 ns, and the snapshots were stored every 2 ps.

The coarse-grained angular potential was again obtained by Boltzmann-inverting the angular distribution function or by using the FM method (we used blocks of 50000 frames each, spline grid of 0.05 nm, and sampling in the $\theta \in [1.6, 3.14]$ interval). Both coarse-grained potentials are shown in Figure 5a. The corresponding distribution functions, together with the reference function obtained from the atomistic simulations, are shown in Figure 5b.

It is obvious that the distribution, which corresponds to simple Boltzmann inversion, is practically identical to the reference distribution, while the FM-based distribution samples small angles much more often, which is a direct consequence of a very deep local minimum in the angular potential at these angles. It is easy to understand why FM fails to predict the relative height of this minimum. On a coarse-grained level, the change of the angle from large to small values corresponds to the reorientation of the dihedral angles at the atomistic level. This reorientation results in instantaneous forces, $\boldsymbol{f}_1$, $\boldsymbol{f}_2$, and $\boldsymbol{f}_3$, on the beads which have

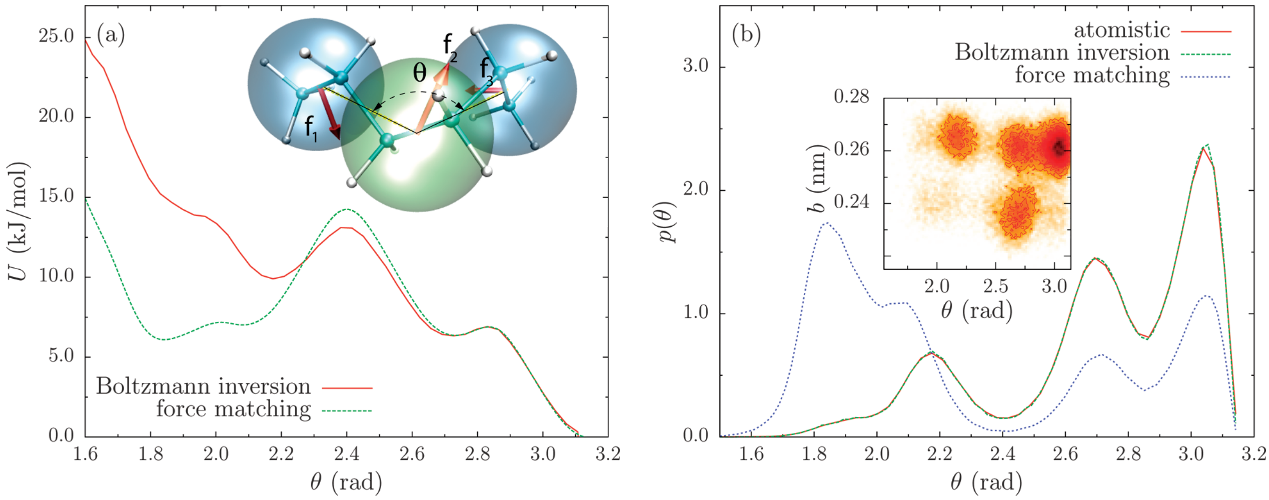


***Figure 5.*** Hexane: (a) Coarse-grained angular potentials obtained using Boltzmann inversion (no iterations) and using FM for a single hexane molecule in vacuum. The inset of (a) shows the hexane molecule and its coarse-grained representation. Arrows indicate the directions of the forces on three beads for a specific snapshot. (b) Probability density (probability distribution normalized by the interval) obtained from the atomistic run as well as from the runs using coarse-grained angular potentials. The inset of (b) shows the correlation of $b$ and $\theta$.

an out of plane component, where the plane is defined by the centers of the beads (see also the inset of Figure 5a). The coarse-grained potential, however, has only an angular term, $U_\theta$, and can only capture forces which lie in the plane in which the angle $\theta$ is defined. Hence, only the projections of the forces on this plane are used in FM, and this clearly leads to underestimation of the position of the second minimum, since the work conducted by the out-of-plane forces is completely ignored.[66]

Additionally, this mapping scheme does not have independent variables, e.g., bond and angle degrees of freedom are coupled, as can be seen from the Ramachandran plot shown in the inset of Figure 5b. This means that, even though Boltzmann inversion reproduces correct distributions, sampling of the configurational space is incorrect because of the lack of cross-correlation terms in the coarse-grained potential.

This example clearly shows that coarse-graining shall be used with understanding and caution, the methods should be cross-checked with respect to each other as well as with respect to the reference system.

## 5. Conclusions

To conclude, we have presented a flexible toolkit for developing and testing coarse-graining methods. Three of them, namely iterative Boltzmann inversion, inverse Monte Carlo, and force matching, have been implemented. With the help of the developed toolkit, we have coarse-grained liquid water, methanol, and propane and a single molecule of hexane. We have also illustrated several advantages as well as shortcomings of the implemented methods. For example, inverse Monte Carlo has an update function which is more efficient than that of the iterative Boltzmann inversion method. On the other hand, inverse Monte Carlo is very sensitive to the system size and the statistical averaging. We have also discussed problems one might encounter when using force matching due to incompleteness of the basis set used to represent the coarse-grained potential energy surface. It should always be kept in mind that the coarse-grained systems are physically different to the reference systems and that the coarse-graining methods cannot be used as a black box and require thorough cross-checking.

We shall also mention that the toolkit has an interface to the fast molecular orbital overlap calculations library and kinetic Monte Carlo code. Combined, these three packages have already been used to study self-assembly and charge transport in organic semiconductors.[49,50]

The source code of VOTCA is available on request and will soon be released under a public license.

**Acknowledgment.** This work was partially supported by DFG via IRTG program between Germany and Korea, DFG grant AN 680/1-1 and by the Multiscale Modeling Initiative of the Max Planck Society. POLYMAT graduate program (V.R.) and Eurosim Early Stage Training project of Marie Curie actions (A.L.) are acknowledged for the financial support. We are grateful to Christine Peter, Will Noid, Brad Lambeth, Markus Deserno, and Karen Johnston for useful discussions. V.R. is thankful to Alexander Lyubartsev for numerous discussions during his stay at Stockholm University.

## Appendix

**A. Force Matching Using Cubic Splines.** Implementations of force matching using different basis functions (linear splines, cubic splines, and step functions) and different methods for solving the least-squares problem (QR decomposition, singular value decomposition, iterative techniques, and normal matrix approach) are discussed in detail in ref 45.

Here, we outline the implementation using cubic splines as basis functions, QR-decomposition for solving the least-squares problem, and block averaging to sample large trajectories.

In our implementation the force $\boldsymbol{f}_{\gamma_i}(\{\boldsymbol{r}_k\})$ acting on bead $i$ due to an interaction $\gamma_i$ with the potential $U_{\gamma_i}$ can be written as

$$\begin{aligned} \boldsymbol{f}_{\gamma i}(\{\boldsymbol{r}_k\}) &= -\nabla_i U(\kappa(\{\boldsymbol{r}_k\})) \\ &= -\frac{\partial U}{\partial \kappa}\nabla_i \kappa(\{\boldsymbol{r}_k\}) \\ &= -f_{\gamma i}\nabla_i \kappa(\{\boldsymbol{r}_k\}) \end{aligned} \tag{19}$$

where $\kappa = r,\ b,\ \theta$, and $\varphi$ denotes the type of interaction and $\nabla_i$ is the gradient with respect to the coordinates $\boldsymbol{r}_i$ of bead $i$. The variable $\kappa$ can label nonbonded interactions, bonds, angles, or dihedral angles, which are given by the distance between the two beads, the bond length, and the angle, which depends on three beads or on the dihedral angle defined using four beads, respectively. Now, the total force $\boldsymbol{f}_i^{\mathrm{cg}}$, acting on coarse-grained bead $i$, can be expressed in terms of the coarse-grained interactions, and eq 16 can be rewritten as

$$\sum_{\gamma_i} f_{\gamma_i}(\kappa)\nabla_i \kappa(\{\boldsymbol{r}_{kl}\}) = \boldsymbol{f}_{il}^{\mathrm{ref}} \tag{20}$$

where $\gamma_i$ enumerates all interactions acting on bead $i$.

$f(\kappa)$ is interpolated using cubic splines connecting a set of points $\{\kappa_k\}$:

$$\begin{aligned} S_n(\kappa, \{\kappa_k\}, \{f_k\}, \{f_k''\}) &= A_n(\kappa)f_n \\ &+ B_n(\kappa)f_{n+1} \\ &+ C_n(\kappa)f_n'' \\ &+ D_n(\kappa)f_{n+1}'' \end{aligned} \tag{21}$$

where $\{f_k\}$and $\{f_k''\}$ are tabulations of $f(\kappa)$ and its second derivative on the grid $\{\kappa_k\}$, the parameters $\{f_k\}$ and $\{f_k''\}$ are obtained from the fit, $\kappa \in [\kappa_n, \kappa_{n+1}]$, and the coefficients $A_n$, $B_n$, $C_n$, and $D_n$ have the following form:

$$\begin{aligned} A_n(\kappa) &= 1 - \frac{\kappa - \kappa_n}{h_{n+1}} \\ B_n(\kappa) &= \frac{\kappa - \kappa_n}{h_{n+1}} \\ C_n(\kappa) &= \frac{1}{2}(\kappa - \kappa_n)^2 - \frac{1}{6}\frac{(\kappa - \kappa_n)^3}{h_{n+1}} - \frac{1}{3}h_{n+1}(\kappa - \kappa_n) \\ D_n(\kappa) &= \frac{1}{6}\frac{(\kappa - \kappa_n)^3}{h_{n+1}} - \frac{1}{6}h_{n+1}(\kappa - \kappa_n) \end{aligned} \tag{22}$$

where $h_n = \kappa_{n+1} - \kappa_n$.

An additional requirement on the spline coefficients is the continuity of the first derivatives:

$$\begin{aligned} &A_n(\kappa_{n+1})'f_n + B_n(\kappa_{n+1})'f_{n+1} + C_n(\kappa_{n+1})'f_n'' + \\ &\quad D_n(\kappa_{n+1})'f_{n+1}'' = A_{n+1}(\kappa_{n+1})'f_{n+1} + B_{n+1}(\kappa_{n+1})'f_{n+2} + \\ &\qquad C_{n+1}(\kappa_{n+1})'f_{n+1}'' + D_{n+1}(\kappa_{n+1})'f_{n+2}'' \end{aligned} \tag{23}$$

If the total number of grid points is $N + 1$ ($n = 0$, 1, ..., $N$), then these conditions are specified for the points $n = 0$, 1, ..., $N - 1$. For nonperiodic potentials, the end points are treated using normal boundary conditions, i.e., $f_0'' = 0$ and $f_N'' = 0$.

Due to the spline fitting, eq 20 simplifies to a set of linear equations with respect to the fitting parameters $f_n$ and $f_n''$. The complete set of equations to solve, therefore, consists of eq 20 and constraints, eq 23. Strictly speaking, this set of equations cannot be solved in a least-squares sense using simple QR decomposition. The reason is that the constraints shall be satisfied exactly to ensure the continuity of the first derivative of the potential, which is not the case if they are treated in a least-squares sense. To solve the problem, one, in principle, has to use a constrained least-squares solver.[51] From a practical point of view, however, it is simpler to treat the constraints in a least-squares sense for each block. This will only lead to a piecewise smooth potential, but the smoothness can be "recovered" by averaging over the blocks.

## References

(1) Tschöp, W.; Kremer, K.; Batoulis, J.; Burger, T.; Hahn, O. *Acta Polym.* **1998**, *49*, 61–74.

(2) Shelley, J.; Shelley, M.; Reeder, R.; Bandyopadhyay, S.; Klein, M. *J. Phys. Chem. B* **2001**, *105*, 4464–4470.

(3) Abrams, C.; Kremer, K. *Macromolecules* **2003**, *36*, 260–267.

(4) Murtola, T.; Falck, E.; Patra, M.; Karttunen, M.; Vattulainen, I. *J. Chem. Phys.* **2004**, *121*, 9156–9165.

(5) Izvekov, S.; Voth, G. *J. Chem. Phys.* **2005**, *123*, 134105.

(6) Sun, Q.; Faller, R. *J.Chem. Theo. Comp.* **2006**, *2*, 607–615.

(7) Harmandaris, V.; Adhikari, N.; van der Vegt, N.; Kremer, K. *Macromolecules* **2006**, *39*, 6708–6719.

(8) Yelash, L.; Müller, M.; Wolfgang, P.; Binder, K. *J. Chem. Theor. Comp.* **2006**, *2*, 588–597.

(9) Shih, A.; Arkhipov, A.; Freddolino, P.; Schulten, K. *J. Phys. Chem. B* **2006**, *110*, 3674–3684.

(10) Lyubartsev, A. *Eur. Biophys. J.* **2005**, *35*, 53–61.

(11) Zhou, J.; Thorpe, I.; Izvekov, S.; Voth, G. *Biophys. J.* **2007**, *92*, 4289–4303.

(12) Villa, A.; van der Vegt, N.; Peter, C. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2068–2076.

(13) Ercolessi, F.; Adams, J. B. *Europhys. Lett.* **1994**, *26*, 583–588.

(14) Hess, B.; Holm, C.; van der Vegt, N. *Phys. Rev. Lett.* **2006**, *96*, 147801.

(15) Reith, D.; Pütz, M.; Müller-Plathe, F. *J. Comput. Chem.* **2003**, *24*, 1624–1636.

(16) Lyubartsev, A.; Laaksonen, A. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1995**, *52*, 3730–3737.

(17) Soper, A. *Chem. Phys.* **1996**, *202*, 295–306.

(18) Izvekov, S.; Parrinello, M.; Burnham, C. J.; Voth, G. A. *J. Chem. Phys.* **2004**, *120*, 10896–10913.

(19) Toth, G. *J. Phys. Cond. Mat.* **2007**, *19*, 335222.

(20) Baschnagel, J.; Binder, K.; Doruker, P.; Gusev, A. A.; Hahn, O.; Kremer, K.; Mattice, W. L.; Müller-Plathe, F.; Murat, M.; Paul, W.; Santos, S.; Suter, U. W.; Tries, V. *Advances in Polymer Science: Viscoelasticity, Atomistic Models, Statistical Chemistry*; Springer Verlag: Heidelberg, Germany, 2000.

(21) Kremer, K. In *Soft and fragile matter, nonequilibrium dynamics, metastability and flow*; Cates, M. E., Evans, M. R., Eds.; J. W. Arrowsmith Ltd.: Bristol, U.K., 2000.

(22) Müller-Plathe, F. *Chem. Phys. Phys. Chem.* **2002**, *3*, 754–769.

(23) Johnson, M.; Head-Gordon, T.; Louis, A. *J. Chem. Phys.* **2007**, *126*, 144509.

(24) *Coarse-Graining of Condensed Phase and Biomolecular Systems*; Voth, G. A., Ed.; CRC Group: Boca Raton, FL, 2008.

(25) Murtola, T.; Bunker, A.; Vattulainen, I.; Deserno, M.; Karttunen, M. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1869–1892.

(26) Peter, C.; Kremer, K. Soft Matter 2009, accepted. DOI: 10.1039/b912027k.

(27) Henderson, R. *Phys. Lett. A* **1974**, *A49*, 197–198.

(28) Chayes, J.; Chayes, L. *J. Stat. Phys.* **1984**, *36*, 471–488.

(29) Chayes, J.; Chayes, L.; Lieb, E. *Comm. Math. Phys.* **1984**, *93*, 57–121.

(30) Leon, S.; van der Vegt, N.; Delle Site, L.; Kremer, K. *Macromolecules* **2005**, *38*, 8078–8092.

(31) Junghans, C.; Praprotnik, M.; Kremer, K. *Soft Matter* **2008**, *4*, 156–161.

(32) Wang, H.; Junghans, C.; Kremer, K. *Eur. Phys. J. E* **2009**, *28*, 221–229.

(33) Murtola, T.; Falck, E.; Karttunen, M.; Vattulainen, I. *J. Chem. Phys.* **2007**, *126*, 075101.

(34) Noid, W.; Chu, J.; Ayton, G.; Voth, G. *J. Phys. Chem. B* **2007**, *111*, 4116–4127.

(35) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theo. Comp.* **2008**, *4*, 435–447.

(36) Nezbeda, I.; Slovak, J. *Mol. Phys.* **1997**, *90*, 353–372.

(37) Wallqvist, A.; Mountain, R. D. *Rev. Comp. Chem.* **2007**, *13*, 183–247.

(38) Kusalik, P. G.; Svishchev, I. M. *Science* **1994**, *65*, 1219–1221.

(39) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(40) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(41) Gunsteren, W. F. V.; Berendsen, H. J. C. *Mol. Sim.* **1988**, *1*, 173.

(42) Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.; Ayton, G.; Izvekov, S.; Andersen, H. C.; Voth, G. *J. Chem. Phys.* **2008**, *128*, 244115.

(43) Janke, W. *Statistical Analysis of Simulations: Data Correlations and Error Estimation*, Lecture notes; Grotendorst, J., Marx, D., Muramatsu, A., Eds.; John von Neumann Institut für Computing (NIC) Series, Vol. 10; NIC: Jülich, Germany, 2002; pp 423−445.

(44) Jorgensen, W.; Tirado-Rives, J. *J. Chem. Soc., Abstr.* **1998**, *216*, U696–U696.

(45) Jorgensen, W.; Tirado-Rives, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665–6670.

(46) Nosé, S. *Mol. Phys.* **1984**, *52*, 255–268.

(47) Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695.

(48) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.

(49) Kirkpatrick, J.; Marcon, V.; Nelson, J.; Kremer, K.; Andrienko, D. *Phys. Rev. Lett.* **2007**, *98*, 227402.

(50) Feng, X.; Marcon, V.; Pisula, W.; Hansen, M.; Kirkpatrick, J.; Grozema, F.; Andrienko, D.; Kremer, K.; Müllen, K. *Nat. Mat.* **2009**, *8*, 421–426.

(51) Golub, G. H.; Van Loan, C. F. *Matrix Computations*, 3rd ed.; Johns Hopkins University Press: Baltimore, MD, 1996.

(52) Harmandaris, V. A.; Reith, D.; Van der Vegt, N. F. A.; Kremer, K. *Macromol. Chem. Phys.* **2007**, *208*, 2109–2120.

(53) Villa, A.; Peter, C.; van der Vegt, N. F. A. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2077–2086.

(54) Noid, W. G.; Chu, J.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G.; Das, A.; Andersen, H. C. *J. Chem. Phys.* **2008**, *128*, 244114.

(55) See also footnote 65.

(56) Note that the coordinates $\{\boldsymbol{R}_j\}$ which are obtained from an atomistic trajectory shall not be confused with the coordinates of a trajectory obtained from coarse-grained simulations.

(57) Note that here we only consider the special case of a linear relation between the $\boldsymbol{r}$ and $\boldsymbol{R}$. $\hat{\mathbf{M}}$ is a block-diagonal matrix and to construct it is enough to specify building blocks for each molecule type. For polymers it is enough to specify $\hat{\mathbf{M}}$ for one repeat unit only.

(58) Note that this is often a "special" trajectory which is designed to decouple the degrees of freedom of interest, e. g., a single polymer chain in vacuum with appropriate exclusions.[1]

(59) Note that, as before, we ignored an irrelevant normalization prefactor Z.

(60) Checking the linear correlation coefficient does not guarantee statistical independence of variables, for example $c(x, x^2) = 0$ if $x$ has a symmetric probability density $P(x) = P(-x)$. This case is often encountered in systems used for coarse-graining.[52,53] The concept is illustrated in section IV for liquid propane and a single molecule of hexane.

(61) Note that eq 7 is nothing else but a numerical scheme that allows one to match the coarse-grained and the reference distribution functions. It can be seen as a firstorder correction to the interaction potential with respect to a gas of non-interacting particles. Indeed, in an ideal gas, the probability of finding two particles at a distance $r$ is $P^{(0)} = 4\pi r^2$, which is equivalent to the statement that the radial distribution function of an ideal gas is 1. Substituting $P^{(0)}$ into eq 7 we obtain the first iteration $U^{(1)} = -k_BT \ln(P_{\text{ref}}/4\pi r^2)$, which is the potential of mean force, eq 2.

(62) A formal statistical mechanical framework of force matching applied to a liquid state, or a multiscale coarse-graining method, is given in ref 54.

(63) http://www.espresso-pp.de.

(64) More detailed analyses have shown that, for small boxes, an additional linear term in the potential update at large separations appear. To determine the origin of this term, $\Delta U$ was calculated using the full matrix $\mathbf{A}_{\alpha\beta}$ as well as only its diagonal elements. The potential after 50 IBI iterations was taken as an initial guess. Without the off-diagonal elements $\Delta U$ was small once the reference and coarse-grained radial distribution functions were matching each other. Inclusion of the off-diagonals elements always resulted in an additional, practically linear, term in the potential update which became smaller for large boxes. Based on this observation we concluded that the off-diagonal elements of the matrix $\mathbf{A}_{\alpha\beta}$ systematically change with the box size.

(65) The united atom model we use here shall not be confused with the united atom models commonly used in the atomistic force-field community, for example OPLS-UA forcefield.[44,45] The latter models map the potentials, which are analytical functions of bonds, angles, and dihedral angles, onto thermo-

dynamic properties of the corresponding substances. In our case coarse-grained potentials are tabulated functions of coarse-grained variables and only the mapping (hydrogens embedded into heavier atoms) is similar to that of the united atom force-fields.

(66) For condensed phase systems, the error introduced by the off-plane component of the force might be compensated by some other pair interactions. In this particular case, however, coarse-graining of liquid hexane with both bonded and non-bonded degrees of freedom treated simultaneously results in a very similar angular distribution to that of a single molecule in vacuum.

CT900369W

JCTC Journal of Chemical Theory and Computation

# First Principles Studies of Fe-Containing Aluminosilicate and Aluminogermanate Nanotubes

Fernando Alvarez-Ramírez

*Instituto Mexicano del Petróleo, Programa de Ingeniería Molecular, Eje Central Lázaro Cárdenas 152, 07730, México, Distrito Federal, México.*

Received June 09, 2009

**Abstract:** A theoretical study of the electronic effects of the inclusion of iron on aluminosilicates and aluminogermanates nanotubes with imogolite-like structure was carried out by unrestricted all-electron density functional theory calculations of periodic boundary models. The iron ion was incorporated to the imogolitic models by an isomorphic substitution of Al by Fe and by the adsorption of the Fe ion in the inner and outer nanotube structure in the octahedral hydrated configuration. Additionally, the effects of the Fe concentration in the interval $0.05 \leq x \leq 0.1$ were analyzed. We observe a drastic reduction of the bandgap value from 4.6 to 2.6 eV and from 4.2 to 1.0 eV for the silicon and germanium respectively. Finally, in all the models there is a shift of the Fermi energy toward the gap region as a result of the inclusion of iron electronic states in the bandgap region.

## Introduction

Imogolite denotes mineral nanotubes compounds with chemical formula $(OH)_3Al_2O_3SiOH$, which manifests the atomic layer arrangement going from the exterior to the interior of the nanotube wall. Mineral imogolite is mostly found in soils derived from volcanic ash and in weathered pumices and spodosols.[1,2] Imogolite-like aluminosilicates (Si-Imo) and aluminogermanates (Ge-Imo) have been synthesized by hydrothermal reactions at relatively low temperature of stirred diluted solutions of aluminum cloride [$AlCl_3$] as source of aluminum and tetraethylorthosilicate $Si(OC_2H_5)_4$ and tetraethylorthogermanate $Ge(OC_2H_5)_4$ as sources of silicon and germanium respectively.[3−5] Recently renovated efforts have been carried out toward a more efficient qualitative and quantitative synthesis procedure of imogolite-like structures.[6−8] Natural imogolites and its synthetic nanotube analog structures ($(OH)_3Al_2O_3SiOH$, Si-Imo) and ($(OH)_3Al_2O_3GeOH$, Ge-Imo) can be applied in the field of chemical sorption reactivity,[9] membrane,[10] humidity controlling,[11] catalysis support.[12,13]

Despite the imogolite-like structures being proposed as good candidates for catalytic applications, because of properties such as large surface areas, which vary from 200 $m^2/g$ to above 700 $m^2/g$, depending on absorbate[4,10] and the charge properties in the inner-outer walls, in most cases these volclays exhibit little or no catalysis because of chemical stability. Therefore, the incorporation of different metallic elements on the imogolites surface (Cr, Mo, W, V, Fe, Ni, Co),[14] (Cd, Cu, Pd),[9] Ag,[15] or active catalytic molecules[16] is necessary to improve their catalytic activity. The imogolite-metal ion dimer can be synthesized by a direct impregnation on the imogolite surface of the metallic ions being the imogolite surface the support for the catalytic metal species.[9,13,14,16] An alternative route for the incorporation of the metallic ions to the imogolite structure/surface is a direct addition of the metal ions in the gel-like precursors.[13,17]

Natural imogolite is commonly found in association with short-range-order materials like allophane and ferrihydrite in many Andisols, where the iron is one of the most recurrent contaminants of imogolite and other soil clays.[18−20] However, the phenomenon of the incorporation or segregation of $Fe^{3+}$ ions from and toward the nanotube is not completely answered because the difficulties in the interpretation of the configuration and atomic environment of the iron ions in the soil-like materials. On the basis of electron spin resonance (ESR) spectroscopy McBride et al.[17] suggested that little or no $Fe^{3+}$ is incorporated to the imogolite structure, observing the tendency of the Al and the Fe to segregate into imogolite and ferrihydrite structures respectively. In their studies, McBride et al. observed the formation of imogolite fibres in preparations with Al/Fe ratios of larger than 9 and the

10.1021/ct9004992 CCC: $40.75 © 2009 American Chemical Society
Published on Web 11/04/2009

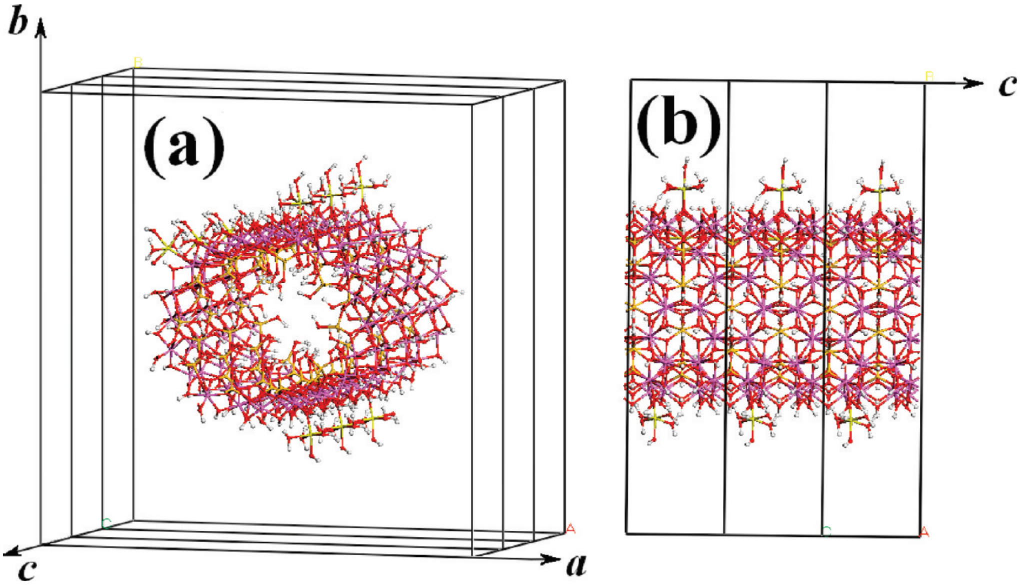


***Figure 1.*** (a) Periodic boundary conditions used in the Fe-imogolite calculations. (b) Periodic conditions along the fiber axis, *c*

formation of ferrihydrite for Al/Fe ratio of 1 and 0.25, deducing that the tube formation is inhibited by Al/Fe ratios of one or less.

On the other hand, Ookawa et al.[13] reported the synthesis of iron containig imogolites with atomic ratio between $x = \mathrm{Fe}/(\mathrm{Fe} + \mathrm{Al}) = 0.05$ and 0.1 (Al/Fe = 9 and 19, respectively), where the iron ions are added to the fibres by two routes. In the first route the $Fe^{3+}$ ions source, $FeCl_3$, are incorporated directly to aqueous solutions of $Na_4SiO_4$ and $AlCl_3 \cdot 6H_2O$, denoting this route as (Imo-Fe). In the second route the $Fe^{3+}$ ions are absorbed directly on the imogolite denoting this process as Imo/Fe. Based on UV−vis and XANES spectra, Ookawa et al. observed the octahedral configuration as the preferred disposition of the $Fe^{3+}$ ions in both Imo-Fe and Imo/Fe routes. Additionally, based on k3-weithed EXAFS function and its Fourier transformation (FT), Ookawa et al. argues the formation of different Fe ion states between the routes Imo-Fe and Imo/Fe. However, the electronic effect of the substitution of $Al^{3+}$ by the $Fe^{3+}$ in the imogolite was not totally elucidated. The inclusion of ferric species like the ferric chloride hexahydride ($FeCl_3 \cdot 6H_2O$) to the imogolite structure is not limited to the case of adsortion process in the imogolite, it is also being applied in the preparation of ppy-imogolite (polypyrrole (ppy)) hybrid materials where the iron acts as an oxidant in the mechanism of polymerization of (ppy).[21]

The aim of this work is to study from the point of view of first principles some of the electronic properties of the iron ion inclusion into and on the imogolite structure and its structural analogue germanate, $(OH)_3Al_2O_3GeOH$, to help in the electronic interpretation of the experimental phenomenology of the incorporation or segregation of $Fe^{3+}$ ions from and toward the imogolite nanotube.

## Methodology

To simulate the pseudocrystalline one-dimensional structure of the Fe-imogolite-like fibers, a sequence of models with periodic boundary conditions were built taking as structural imogolite the standard model reported by Wada et al.[2,22] and Farmer et al.[5] The skeleton of the iron containing imogolite-like structure is a model with ten gibbsite units, $N_u = 10$, considered as the most likely structure model of natural imogolite.[23] The dimensions of the rectangular simulation cells along the tubular axial direction (*c*, Figure 1) were chosen as 8.78 and 8.8 Å for the silicon and germanium containing structures respectively on the basis of previous calculations using the Γ point approximation.[24] The transversal cell lengths *a* and *b* were chosen to be 40 Å to inhibit the interaction of the imogolite images along these non axial directions because the periodic boundary conditions of the models, Figure 1. In our case the *c* cell length is slightly larger than the experimental X-ray diffraction value for aluminosilicate and the aluminagermanate imogolite-like structures, that is, 8.51 Å. This overestimation has been reported as a consequence of the semilocal density functional approximation in the case of aluminosilicates.[25−28]

Experimental reports of McBride et al.[17] show the fibrous structure are preserved only in the cases where the ratio Al/Fe ≤ 9. Therefore; the iron containing in our models was restricted to values above this critical threshold. As regards to the iron ions localization, they were located in three different qualitative imogolite places. The first iron localization corresponds to the isomorphic substitution of $Al^{3+}$ by $Fe^{3+}$ ions in the imogolite atomic layer denoted as $Al_5$ in Figure 2a and b; this configuration plays the role of iron absorbed on the imogolite structure in a similar way to the Imo-Fe structure reported by Ookawa et al.[13] This kind of configuration was in this work simulated by four models with isomorphic substitution of one to four $Al^{3+}$ atoms by $Fe^{3+}$ which corresponds to $x = (0.025, 0.05, 0.075, 0.1)$ with a general notation M-Imo-Fe, see Table 1 and Supporting Information.

The two remaining iron configurations corresponds to the case where the iron atom is adsorbed on the inner or the

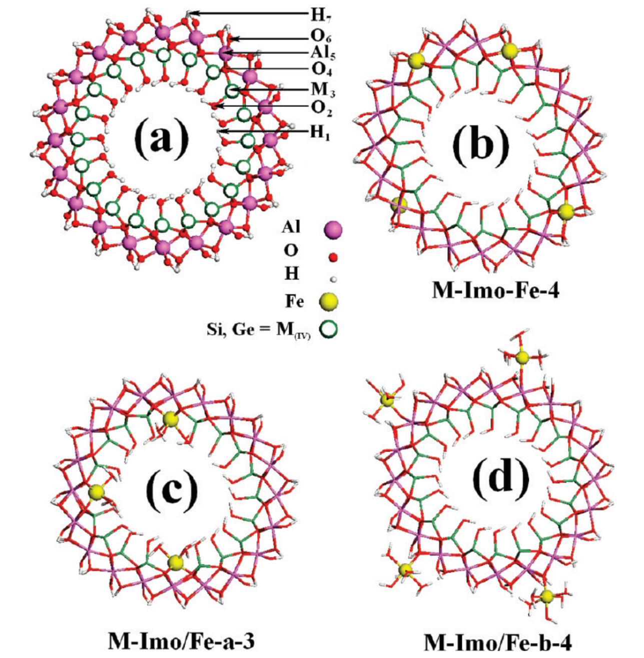


***Figure 2.*** (a) Transversal section of the imogolite structure where is indicated the notation of the atomic layers in the structure. (b) Schematic representation of octahedral isomorphic substitution of $Al^{+3}$ for $Fe^{+3}$. (c−d) Adsorption of octahedral $Fe^{+3}$ ions on the inner and outer imogolite surface. In all the cases the M symbol denotes both Si or Ge containing structures.

***Table 1.*** Notation Use in the Fe-Containing Aluminosilicate and Aluminogermanate Nanotubes Analyzed in This Work[a]

| compound | short name | $x$ | Al/Fe | class |
|---|---|---|---|---|
| $H_{80}O_{140}Al_{39}Fe_1M_{20}$ | M-Imo-Fe-1 | 0.025 | 39.0 | subst |
| $H_{80}O_{140}Al_{38}Fe_2M_{20}$ | M-Imo-Fe-2 | 0.050 | 10.0 | subst |
| $H_{80}O_{140}Al_{37}Fe_3M_{20}$ | M-Imo-Fe-3 | 0.075 | 12.3 | subst |
| $H_{80}O_{140}Al_{36}Fe_4M_{20}$ | M-Imo-Fe-4 | 0.100 | 9.0 | subst |
| $H_{85}O_{142}Al_{40}Fe_1M_{19}$ | M-Imo/Fe-a-1 | 0.024 | 40.0 | inner/abs |
| $H_{90}O_{144}Al_{40}Fe_2M_{18}$ | M-Imo/Fe-a-2 | 0.048 | 20.0 | inner/abs |
| $H_{95}O_{146}Al_{40}Fe_3M_{17}$ | M-Imo/Fe-a-3 | 0.070 | 13.3 | inner/abs |
| $H_{100}O_{148}Al_{40}Fe_4M_{16}$ | M-Imo/Fe-a-4 | 0.090 | 10.0 | inner/abs |
| $H_{87}O_{145}Al_{40}Fe_1M_{20}$ | M-Imo/Fe-b-1 | 0.024 | 40.0 | outer/abs |
| $H_{94}O_{150}Al_{40}Fe_2M_{20}$ | M-Imo/Fe-b-2 | 0.048 | 20.0 | outer/abs |
| $H_{101}O_{155}Al_{40}Fe_3M_{20}$ | M-Imo/Fe-b-3 | 0.070 | 13.3 | outer/abs |
| $H_{108}O_{160}Al_{40}Fe_4M_{20}$ | M-Imo/Fe-b-4 | 0.090 | 10.0 | outer/abs |

[a] The M symbol denotes the Si or Ge aluminosilicate nanotube variation. On the other hand, the notation subst, inner/abs and outer/abs corresponds with the $Fe^{+3}$ in the imogolite-like structure as is shown in Figure 2a and Figures S1 and S2 of the supporting information.

outer the imogolite surfaces producing surface defects in the standard imogolite model being this kind of configuration similar to the Ookawa's Imo/Fe arrangement.[13] In the first Imo/Fe configuration denoted in general as M-Imo/Fe-a, see Table 1, the iron atom substitutes the element of the group IV (layer $M_3$ in Figure 2a) in the pore tubular region, being linked to three oxygen atoms of the $O_4$ layer, Figure 2a. Because the octahedral disposition is the preferred configuration of the $Fe^{3+}$ ions a hydrated sphere of three water molecules were added to the iron ion to preserve both the octahedral iron configuration and the oxidation 3+, Figure 2c.

The second Imo/Fe configuration denoted in general as M-Imo/Fe-b, see Table 1, corresponds to the adsorption of irons ion in the outer imogolite surface where the $Fe^{3+}$ ions are monolinked to an oxygen of the $O_6$ layer, taking a similar place to the $H_7$ hydrogen position in the imogolite outer surface, Figure 2a-2d. The oxidation state 3+ of the iron in the Imo/Fe-b structures was kept constant by surrounding the ion by three water molecules and two $OH^-$ molecules in octahedral configuration. In order to analyze the iron concentration effects on the electronic properties of the silicon and germanium imogolite-like structures, a sequence of 4 models with similar structural characteristics to the Imo-Fe, Imo/Fe-a and Imo/Fe-b were built varying the number of iron ions in the model from 1 to 4 ($0.05 \leq x \leq 0.1$; $x =$ Fe/(Fe + Al)) for each case. Table 1 and the Supporting Information display the notation of the 12 model structural variations considered in this work where Al/Fe ratio was kept above the experimental fibrous morphology and non-iron-aggregation threshold of 9 (Al/Fe = 9; $x = 0.1$) reported by McBride et al.[17]

The iron containing imogolite models were structural optimized carrying out geometry optimizations of the radial nanotube dimension keeping the cell lengths $a$, $b$, and $c$ constant during the geometry optimization. The non-dependence of the $c$ length with respect to the iron content arises because the iron concentration in the cell is small with respect to the entire structure content; however, this restriction should be considered just as a first approach. The ab initio optimizations were carried out using unrestricted all electron DFT calculations as is implemented in the $DMol^3$ code[29] with Double Numerical basis set (DN),[30] together with the Perdew−Wang exchange correlation functional, GGA-PW91.[31] The geometries were optimized until the maximum force and displacement on the system were 0.004 Ha/Å and 0.005 Å, respectively, to ensure a near ground-state configuration. Because of the high level computational cost of the unrestricted calculations, all the ab initio calculations were performed considering only the $\Gamma$ point in the reciprocal space, which provides a good qualitative description of the problem as it is shown and validated in previous work.[24]

## Electronic Structure Analysis

The electronic effects of the addition of iron ions into the imogolite systems are mainly displayed in the bandgap region giving rise to electronic contamination states.Figure 3a−d shows the electronic density of states (DOS), varying the iron content $x =$ Fe/(Fe + Al), of the Si-Imo-Fe (isomorphic $Al^{3+}$ substituted by $Fe^{3+}$) models compared with the DOS of the non iron content imogolite-like system. In the curves of Figure 3a−d the DOS was split such that energy of the maximum of the valence bands of both systems coincides. The contamination of the bandgap by the Fe produces electronic states localized mainly in three bandgap regions: (a) in the top of the valence band, (b) in the bottom of the conduction band, and (c) in the middle of the bandgap, see insets in Figure 3a−d.

In the case of the silicon containing imogolite-like structures (M = Si in Table 1), the small concentration of iron in the models does not alter significantly the shape of

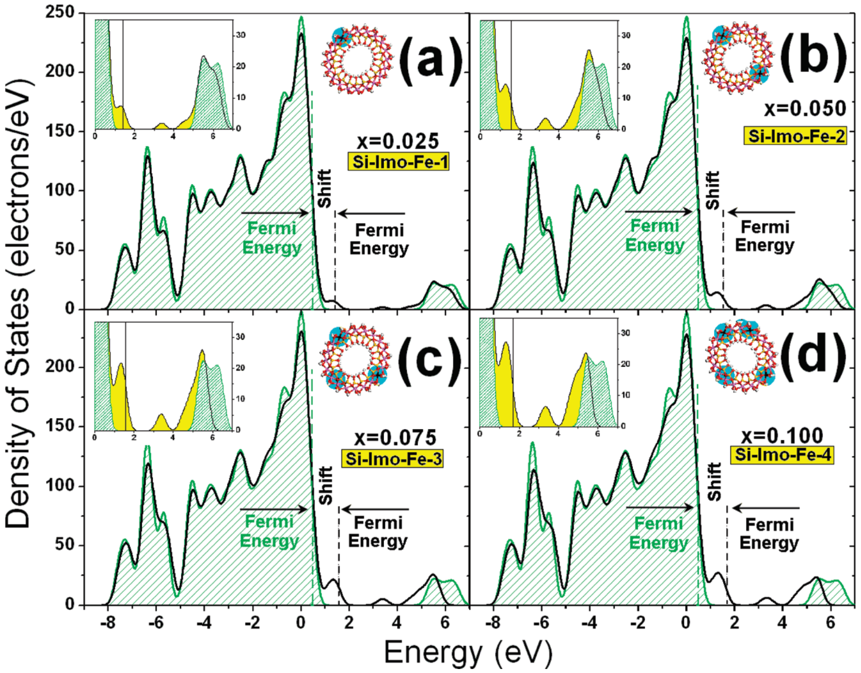


***Figure 3.*** (a−d) Total DOS of the Fe isomorphic substituted imogolite structures: Si-Imo-Fe. The DOS of the iron containing models were compared with DOS of the non-contenting models, green dashed line. The insets in a−d show the bandgap region where the new defect states were yellow highlighted. Finally, the dashed line indicates the Fermi energy for the both Fe containing and non containing imogolite-like models.

the valence band outside the bandgap region; however, the defects generated by the iron on the imogolite-like structures produce a shift of the largest occupied state, Fermi energy, with respect to the uncontaminated case, indicated as a dashed line in Figure 3a−d. The Fermi energy shift and the energy localization of defect states in the bandgap region are almost independent of the of the iron content in the models, generating a reduction of the bandgap value from ∼4.6 eV in the case of non-iron-containing model[24] to 1.49, 1.69, 1.67, and 1.97 eV for the isomorphic substituted Si-Imo-Fe-1, Si-Imo-Fe-2, Si-Imo-Fe-3, and Si-Imo-Fe-4 models, respectively. In the case of the sequence of silicon imogolite-like models denoted as (Si/Imo-Fe-a and Si/Imo-Fe-b), the localization of the defect electronic states in the DOS displays similar characteristics and bandgap values to those analyzed in the Si-Imo-Fe models, see Figures S3−S4 in the Supporting Information.

On the other hand, the presence of germanium on the imogolite-like models, M = Ge in Table 1, seems not to affect drastically the behavior of the electronic states observed for the case of silicon imogolite models. Figure 4a−d displays the DOS for the Ge/Imo-Fe-b models where, like in the case of Si-Imo-Fe models, the DOS curves of the non-iron-containing and iron-containing Ge/Imo-Fe-b models were split in such away that the energy of the maximum of the valence bands of both systems coincide. The insets of Figure 4a−d show the difference in the curves between the non-iron-containing and iron-containing Ge/Imo-Fe-b models in the bandgap region observing a shift of the Fermi energy together with the presence of defect states in the top, middle, and bottom of the bandgap region like in the case of the silicon imogolite-like models. The bandgap values on germanium containing imogolite-like structures, the DOS, display a larger iron dependence than in the silicon case having values of 2.6, 1.00, 0.94, and 0.9 eV for the Ge/Imo-Fe-b-1, Ge/Imo-Fe-b-2, Ge/Imo-Fe-b-3, and Ge/Imo-Fe-b-4 models, respectively. Similar iron concentration bandgap dependence is observed for the systems Ge-Imo-Fe and Ge-Imo/Fe, see Figures (S5−S6) in the Supporting Information.

With the aim of discerning the origin of the defect states in the DOS curves, an analysis of the correlation between the electronic states within the DOS and the geometric radial position in the nanotube structure through the local density of states (LDOS) was carried out in the Si/Imo-Fe-b-4 and Ge/Imo-Fe-a-4 models whose results are displayed in Figures 5 and 6. The top of the valence band, in both Si/Imo-Fe-b-4 and Ge/Imo-Fe-a-4 models, has the main contributions from the oxygen layers of the imogolite-like structure ($O_2$, $O_4$, $O_6$). These oxygen weights are similar to previous works in non-iron-containing aluminosilicate and aluminogermanate nanotube models.[24] However, in iron content cases there are an additional contribution coming from the hydratation sphere of the iron ion states, see Figures 5a−c, 6a−c, and S7−S12 in the Supporting Information. On the other hand, the main contributions to the valence band shoulder generated by the shift of the Fermi energy come from states linked to the iron, hydration, and $O_6$ atoms in the case of the Si/Imo-Fe-b-4 model, whereas in almost all the analyzed cases the main contributions to the shoulder of the top valence band are connected to the iron ion and to the $O_4$ imogolite-like layer, Figures 5, 6, and S7−S12 in the Supporting Information.

In all the cases, the middle bandgap electronic states are almost exclusively associated to the iron ions, whose importance grows when the iron content is increased. An example of these contributions is the composition of the DOS peak at the top of the conduction bandgap region which has

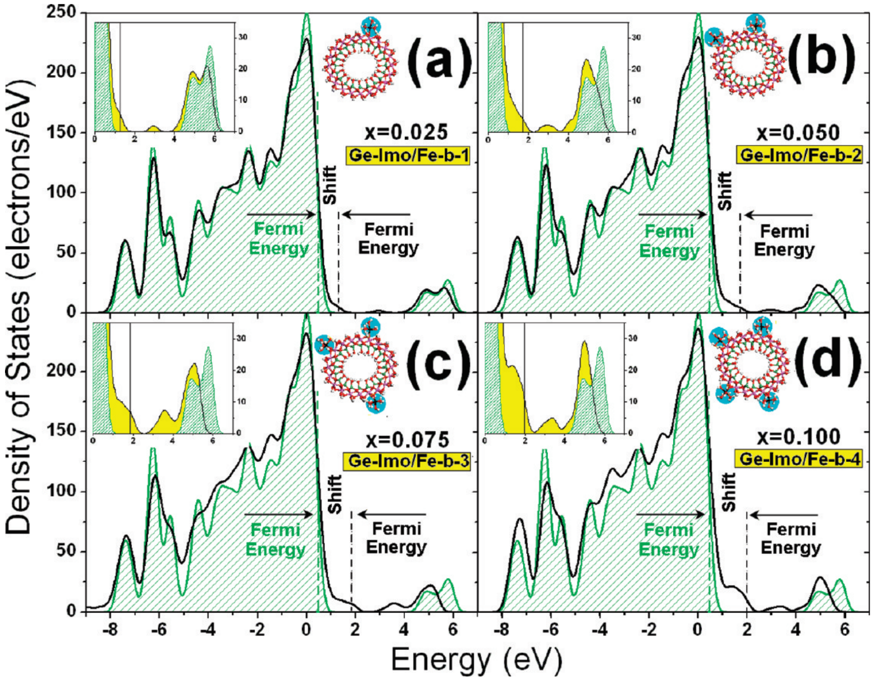


***Figure 4.*** (a−d) Total DOS of the Fe ion adsorbed on the outer imogolite-like surface models: Ge-Imo/Fe-b. The DOS of the iron-containing models were compared with DOS of the non-iron-containin models, green dashed line. The insets in a−d shows the bandgap region where the new defect states were yellow highlighted. Finally, the dashed line indicates the Fermi energy for the both Fe containing and non containing imogolite-like models.

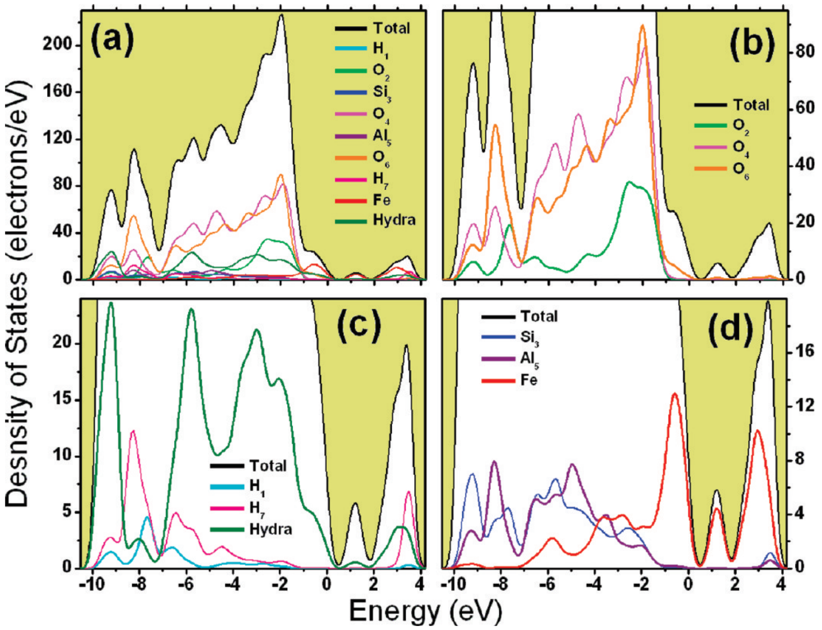


***Figure 5.*** (a) Overview of the local density of states contribution to the total DOS for the Si-Imo/Fe-b-4, the notation indicates the atomic layer whereas hydra is the notation for the hydratation sphere of Fe. (b) Zoom of the oxygen contribution to DOS. (c) Zoom to the hydrogen and hydratation contribution to DOS. (d) Zoom of the $Si_3$, $Al_5$, and Fe to the DOS.

a high contribution of Fe states with smaller contributions coming from the hydratation iron sphere. Additional contributions to this peak come from the $H_7$ and $O_6$ in the case of Si/Imo-Fe-b-4 model, whereas for the Ge/Imo-Fe-a-4 model the $H_7$, $O_6$ and $Ge_3$ atomic layers also contribute to this peak. Similar aspects in the DOS partial contributions are shown by the rest structures, Figures S5−S12 in the Supporting Information. Because of the previous analysis, it is concluded that the inclusion of iron in the imogolite-like structure adds new electronic states to the DOS associated to the Fe ion and their hydratation sphere which are overlapped and shifted from the original imogolite states.

The distribution of the electronic states around the Fermi energy displayed in Figures 5, 6, and S7−12 in the Supporting Information clearly shows a high localization around the ion atoms. This localization phenomenon is also present when the spin density of states (spin-DOS) are analyzed, Figures 7 and S13−18 in the Supporting Information. In particular, Figure

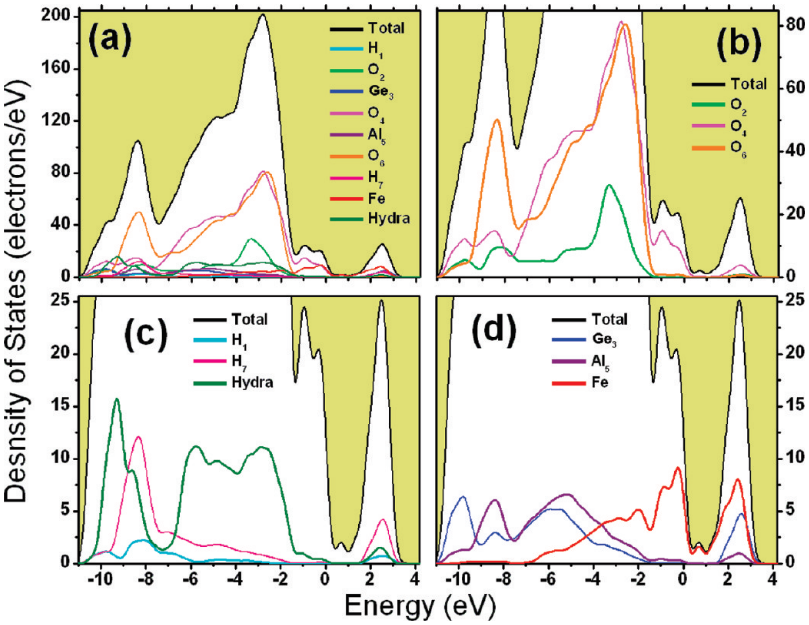


*Figure 6.* (a) Overview of the local density of states contribution to the total DOS for the Ge-Imo/Fe-a-4, the notation indicates the atomic layer whereas hydra is the notation for the hydratation sphere of Fe. (b) Zoom of the oxygen contribution to DOS. (c) Zoom to the hydrogen and hydratation contribution to DOS. (d) Zoom of the $Ge_3$, $Al_5$ and Fe to the DOS.

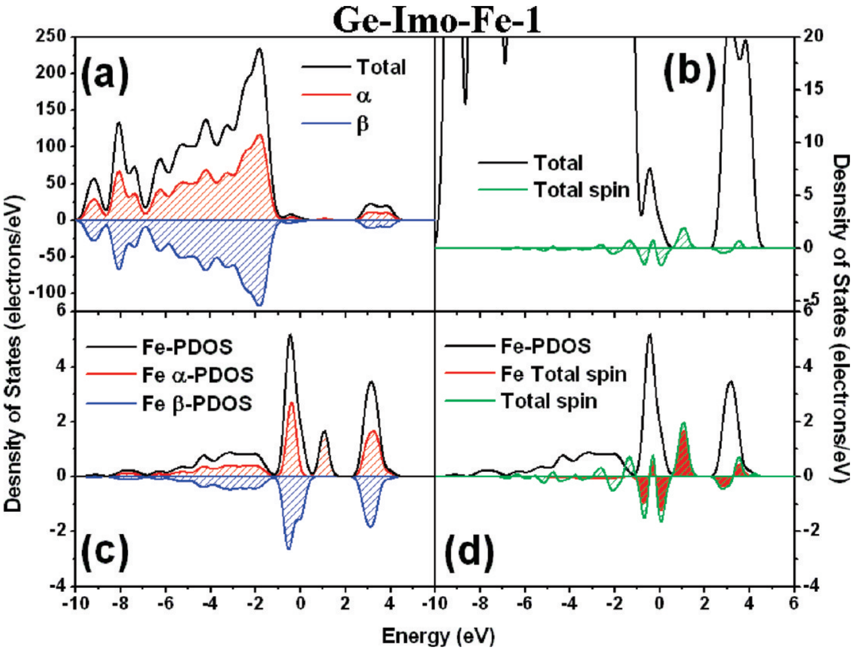


*Figure 7.* Spin-DOS of the system Ge-Imo-Fe-1. (a) α and $\beta$ spin DOS contribution to Total DOS (b) Relative position and weight of the total spin DOS with respect to the Total DOS respect to the Total DOS. (c) Fe PDOS and its relative α and $\beta$ spin PDOS. (d) Relative positions and weights of the Fe-PDOS, Fe Total spin, and the total spin of the model.

7a depicts the spin-DOS for the case Ge-Imo-Fe1 where the α and $\beta$ spin-DOS shows similar shape and weights outside the band gap region. The spin difference between the α and $\beta$ DOS displays, clearly, the localization of the total spin densities around the bandgap region, showing additionally an oscillatory behavior of the total spin-DOS in this region, Figure 7b. A closer analysis of the nature of the spin-DOS displays that the main contribution to the spin-DOS in the bandgap region is linked to iron ions incorporated in the structures being this phenomena common to all the analyzed structures, as is depicted in Figures 7c−d and S13−S18 in the Supporting Information.

As regards the final spin of the configurations, the imogolite-like structures displays in most cases a doublet configuration (spin = 1/2), when the number of iron ions is even, and singlet spin configuration (spin = 0) when number of iron ions is odd. The exceptions to this behavior are the following cases: Si-Imo/Fe-a-2, Si-Imo-Fe-4, Si-Imo/Fe-b-4,Ge-Imo-Fe-2, Si-Imo/Fe-b-4, and Si-Imo/Fe-b-4 with triplet spin configuration ($s$ = 1).

The localization of the Fermi energy orbitals around the iron atoms is ratified by the geometric distribution of the HOMO and LUMO orbitals. In particular, Figure 8 shows

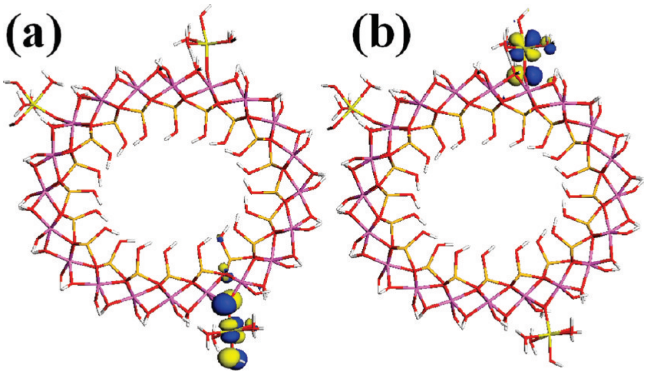


***Figure 8.*** (a) Iso-surface with value 0.01 of the HOMO orbital for the structure Si-Imo/Fe-b-3. (b) Iso-surface with value 0.01 of the LUMO orbital for the structure Si/Imo-Fe-b-3. Clearly the both orbitals are localized around the Fe ion.

the spatial distribution of the isosurface with a value 0.01 for the Si/Imo-Fe-b-3 model where the localization of the HOMO and LUMO orbitals, around the position of two of the Fe atoms, is displayed. Note that the HOMO orbital is localized only in one of the three Fe atoms whereas the LUMO orbital is localized in a different Fe atom. The localized orbital linked to the third Fe ion is energetically situated in the neighborhood of the Fermi energy.

## Reach and Limitations of the Bandgap Calculation

With the aim test the reach and limitations of the GGA-PW91 the band gap values in the case of Fe-aluminosilicate and Fe-aluminogermanate systems with respect to other DFT-functional approaches, a sequence of calculations were carried out using 3 types of functionals: GGA-PW91,[31] the revised Perdew−Burke−Ernzerhof GGA-RPBE,[32] and the hybrid B3LYP.[33,34] As test system, we have chosen a set of iron-containing flat imogolite-like structures with cell dimensions $a = 8.46$ Å, $b = 9.8$ Å, which corresponds to two gibbsite-like units. To ensure and empty atom gap between the imogolite-like layers the $c$ length was selected as 20 Å. The iron ions were placed in configurations similar to those used on the tubular configuration, subst, inner/abs, and outer/abs, as is depicted in Figures 9 and S19−S23 in the Supporting Information. Like in the case of tubular structures, the electronic DOS calculations were carried out using the Γ point approximation and spin-unrestricted methodology. For both GGA-PW91 and GGA-RPBE, the calculations were carried out using the orbital localized code DMol[3] with all electron approach where the iron imogolite-like surface was geometry optimized using the GGA-PW91 and the GGA-RPBE functionals respectively. On the other hand the hybrid-B3LYP the DOS calculations were carried out using the pseudopotential plane wave CASTEP code.[35] where the iron ion-imogolite-like system was geometry optimized using the GGA-PW91.

The results of the DOS calculations in flat iron-containing models show, in general, similar DOS shape and bandgap values independently of the DFT approach. In the case of the hybrid-B3LYP functional the middle bandgap states have a lower height with respect to the DOS than the GGA-PW91 and GGA-RPBE functionals; however, the middle bandgap states that define the bandgap value are still present in the hybrid-B3LYP DOS calculations with gap values comparable to those found in the flat and tubular imogolite-like structures.

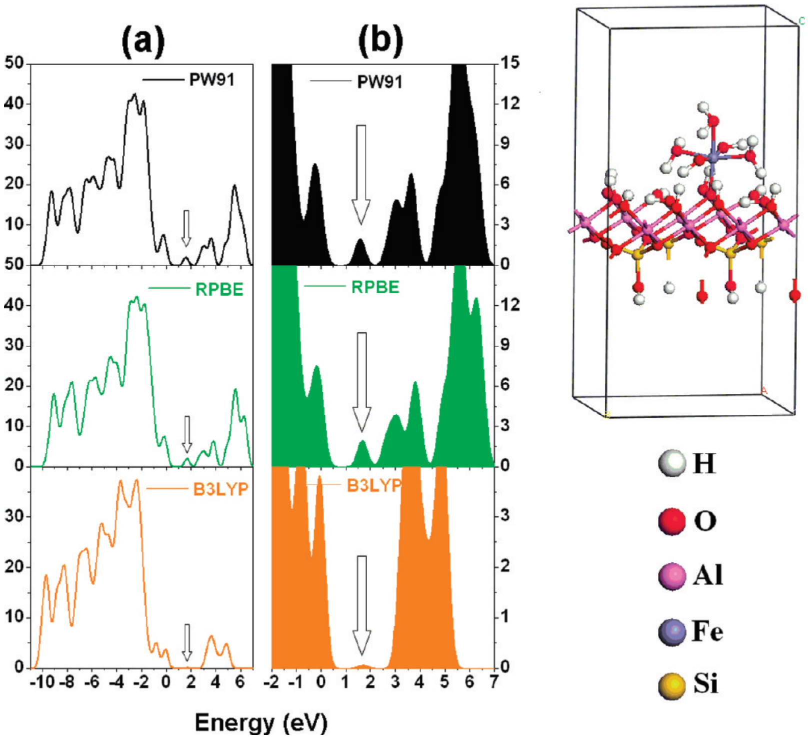


***Figure 9.*** (a) DOS of silicon containing flat imogolite structure using the functionals PW91, RPBE, and B3LYP. This system is similar to the configurations outer/abs: $H_{87}O_{145}Al_{40}Fe_1Si_{20}$, $H_{94}O_{150}Al_{40}Fe_2Si_{20}$, $H_{101}O_{155}Al_{40}Fe_3Si_{20}$, and $H_{108}O_{160}Al_{40}Fe_4Si_{20}$ of the Table 1. (b) Zoom of the DOS depicted where is displayed the middle bandgap states.

This phenomenon is most clearly displayed by the subst. and outer/abs configurations, Figures 9 and S19−S21 in the Supporting Information. On the other hand the flat inner/abs configurations shows larger discrepancies between the GGA-PW91 and GGA-RPBE and the hybrid B3LYP calculations in the bandgap region due to the larger water versus Si−OH and Ge−OH interaction in this configurations, Figures S22−S23 in the Supporting Information. On the basis of these calculations, some qualitative and quantitative aspects of the DOS of iron containing imogolite-like systems and the order or magnitude of the bandgap found in the tubular imogolite-like structures are validated, at least in terms of the DFT.

## Conclusions

The electronic properties of isomorphic substituted and adsorbed iron aluminosilicate and aluminogermanate nanotube structures were analyzed based on all electrons unrestricted ab initio calculations. The optimized nanotube models have iron content in the interval $0.05 \leq x \leq 0.1$ and varying iron ion positions in three different octhahedral configurations: inner, outer imogolite surface and isomorphic substituted in the Al gibbsite layer. In all the configurations and content cases the electronic Fe-imogolite-like silicate/gemanate DOS displays large changes because of the contamination by Fe electronic states of the bandgap region, generating the reduction of the gap values from ~4.7 to [2.0−1.4 eV] and from ~4.2 to [2.6−1.0 eV] for the Fe-silicon and Fe-germanium imogolite-like nanotubes. The Fe inclusion into the imogolite-like structures produces a shift of the Fermi energy and the overlapping of Fe electronic states to the original imogolite-like states at the top, the middle, and the bottom of the gap region.

**Supporting Information Available:** Figures showing snapshots of optimized iron-containing aluminosilicates and aluminogermanate, DOS of the imogolite−iron dimer, adsorbed Fe−imogolite over the inner imogolite surface, isomorphic substituted Al by Fe in germanium-containing nanotube, adsorbed imogolite over the inner imogolite surface, germanium-containing flat imogolite structure, andsilicon-containing flat imogolite structure, sequence of total DOS and PDOS, and sequence of spin-DOS. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Yoshinaga, N.; Aomine, S. *Soil Sci. Plant Nutr. (Tokyo)* **1962**, *8*, 22–29.

(2) Wada, K.; Yoshinaga, N. *Am. Mineral.* **1969**, *54*, 50–71.

(3) Wada, S.-I; Eto, A.; Wada, K. J. *J. Soil Sci.* **1979**, *30*, 347–355.

(4) Wada K. In *Minerals in Soil Environments*, 2nd ed.; Soil Science Society of America: Madison, WI, 1989.

(5) Farmer, V. C.; Fraser, A. R.; Tait, J. M. *J. Chem. Soc., Chem. Commun* **1977**, *13*, 462–463.

(6) Mukherjee, S.; Kim, K.; Nair, S. *J. Am. Chem. Soc.* **2007**, *129*, 6820–6826.

(7) Levard, C.; Rose, J.; Masion, A.; Doelsch, E.; Borschneck, D.; Olivi, L.; Dominici, C.; Grauby, O.; Woicik, J. C.; Bottero, J.-Y. *J. Am. Chem. Soc.* **2008**, *130*, 5862–5863.

(8) Yang, H.; Wang, C.; Su, Z. *Chem. Mater.* **2008**, *20*, 4484–4486.

(9) Denaix, L.; Lamy, I.; Bottero, J. Y. *Colloids Surf., A* **1999**, *158*, 315–325.

(10) Ackerman, W. C.; Smith, D. M.; Huling, J. C.; Kim, Y.-W.; Bailey, J. K.; Brinker, C. J. *Langmuir* **1993**, *9*, 1051–1057.

(11) Masaya, S.; Shin, S.; Masaki, M.; Shinji, T.; Tadato, M. *J. Ceram. Soc. Jpn.* **2001**, *109*, 874–881.

(12) Imamura, S.; Kokubu, T.; Yamashita, T.; Okamoto, Y.; Kajiwara, K.; Kanai, H. *J. Catal.* **1996**, *160*, 137–139.

(13) Ookawa, M.; Inoue, Y.; Watanabe, M.; Suzuki, M.; Yamaguchi, T *Clay Sci.* **2006**, *12*, 280–284.

(14) Nordstrand R. A., U.S. Patent 4394253, 1984.

(15) Hirohisa, Y.; Michalik, J.; Sadlo, J.; Perlinska, J.; Takenouchi, S.; Shimomura, S.; Uchida, Y. Appl. *Clay Sci.* **2001**, *19*, 173–178.

(16) Nakagaki, S.; Wypych, F. *J. Colloid Interface Sci.* **2007**, *315*, 142–157.

(17) MacBride, M. B.; Farmer, V. C.; Russell, J. D.; Tait, J. M.; Goodman, B. A. *Clay Miner.* **1984**, *19*, 1–8.

(18) Ugolini, F. C.; Dahlgren, R. A. *Soil Sci. Soc. Am. J.* **1991**, *55*, 1166–1171.

(19) Taylor, R. M.; Raupach, M.; Chartres, C. J. *Clay Miner.* **1990**, *25*, 375–389.

(20) Imamura, S.; Hayashi, Y.; Kajiwara, K.; Hoshino, H.; Kaito, C. *Ind. Eng. Chem. Res.* **1993**, *32*, 600–603.

(21) Lee, Y.; Kim, B.; Yi, W.; Takahara, A.; Sohn, D. *Bull. Korean Chem. Soc.* **2006**, *27*, 1815–1818.

(22) Wada, S.; Wada, K. *Clays Clay Miner.* **1982**, *30*, 123–128.

(23) Cradwick, C. P. G.; Farmer, V. C.; Russell, J. D.; Masson, C. R.; Wada, K.; Yoshinaga, N. *Nat. Phys. Sci.* **1972**, *240*, 187–189.

(24) Alvarez-Ramírez, F. *Phys. Rev. B* **2007**, *76*, 125421−1-125421−14.

(25) Li, H.; Mahanti, S. D.; Pinnavaia, T. J. *J. Phys. Chem. B* **2005**, *109*, 2679–2685.

(26) Gale, J. D.; Rohl, A. L.; Milman, V.; Warren, M. C. *J. Phys. Chem. B* **2001**, *105*, 10236–10242.

(27) Refson, K.; Park, S.-H.; Sposito, G. *J. Phys. Chem. B* **2003**, *107*, 13376–13383.

(28) Teobaldi, G; Beglitis, N. S.; Fisher, A. J.; Zerbetto, F; Hofer, W. A. *J. Phys.: Condens. Matter* **2009**, *21*, 95301–95309.

(29) S. D. I. DMol3. User Guide, release 4.0.

(30) Delley, B. J. *J. Chem. Phys.* **2000**, *113*, 7756–7764.

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

(32) Hammer, B.; Hansen, L. B.; Nørskov, J. K. *Phys. Rev B.* **1999**, *59*, 7413–7421.

(33) Kim, K.; Jordan, K. D. *J. Phys. Chem.* **1994**, *98*, 10089–10094.

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

(35) Clark, S. J.; Segall, M. D; Pickard, C. J.; Hasnip, P. J; Probert, M. J; Refson, K.; Payne, M. C. *Z. Kristallogr.* **2005**, *220*, 567–570.

CT9004992

JCTC Journal of Chemical Theory and Computation

# Mixed Resolution Modeling of Interactions in Condensed-Phase Systems

Sergei Izvekov and Gregory A. Voth*

*Center for Biophysical Modeling and Simulation and Department of Chemistry, University of Utah, 315 South 1400 E., Room 2020, Salt Lake City, Utah 84112-0850*

Received August 7, 2009

**Abstract:** A new mixed resolution method is developed for modeling molecular interactions that employs a distance-dependent coupling of atomistic and coarse-grained force fields. In the mixed resolution interaction (MRI) method, detailed atomistic structure is maintained over the whole system. However, the atomistic force field is used for close interparticle separations (called the atomistic zone), while at large separations the coarse-grained forces are "unfolded" into atomistic interactions in a way that reduces the cost of the simulation compared to standard long-range approximations or cutoff schemes. Several variations of the unfolding scheme are described. The method is applied to develop MRI models of bulk TIP3P water, based on one-site multiscale coarse-grained (MS-CG) water potentials located at the molecular centers of either mass or geometry. With a sufficiently large atomistic zone (>0.7 nm), the MRI models provide excellent simulations of the bulk water phase. MRI modeling is further illustrated for liquid methanol with both one- and two-site coarse graining. The MRI water models are then used to simulate aqueous solutions, where the solutes are treated at the atomistic level. It is shown that the MRI treatment significantly alters solute association dynamics if it relies on the MS-CG force fields obtained solely from the bulk phase. Possible modifications of the MRI procedure to improve the transferability of water potentials to heterogeneous systems are, therefore, discussed. The best result is obtained if water molecules within a preselected cutoff distance from the solute are described using only atomistic potentials. As a final example, the MRI method is applied to model a solvated phospholipid bilayer.

## 1. Introduction

The structure and dynamics of complex condensed-phase systems are often influenced by multiscale phenomena. It is possible to adequately describe the behavior of such systems on mesoscopic or macroscopic scales, although such models are necessarily less complete than those described by the full set of atomistic variables and the laws describing their interactions. The limited success and transferability of many empirical models, tailored to study condensed-phase systems on particular scales, testifies to this fact. The first and most widely used strategy for empirical modeling, which may be recognized as a top-down approach, is based on extracting dynamic variables and interactions directly from the system's mesoscopic properties. In contrast, the more recent bottom-up philosophy in condensed-phase modeling assumes that a hierarchy of relevant scales exists in condensed-phase systems, beginning with the molecular scale. The bottom-up approach first identifies the most important degrees-of-freedom on mesoscopic scales and then projects the atomistic coordinates and interactions onto the phase space spanned by the relevant coordinates. Such a strategy holds certain advantages over top-down strategies since explicit molecular-scale force information is systematically propagated upward in scale to the mesoscopic level. However, it is also technically more complex and, in some cases, may become infeasible because atomistic simulations are computationally expensive.

A simpler but potentially less accurate approach than the fully atomistic description is particle-based coarse graining,

* Corresponding author. Telephone: 801-581-7272. Fax: 801-581-4353. E-mail: voth@hec.utah.edu.

10.1021/ct900414p CCC: $40.75
Published on Web 11/03/2009

where groups of atoms are replaced by single interaction sites.[1] Particle-based coarse-grained (CG) simulations have become popular because they provide controllable, quasi-atomistic resolution and require significantly less computational cost than that of a fully atomistic treatment. While most particle-based CG methods rely on underlying atomistic structures to partition the system into interaction centers, strategies for coarse graining the interactions themselves have mainly been developed in the top-down spirit using thermodynamic information.[2−6] However, in recent years an efficient and systematic bottom-up strategy for constructing effective pairwise CG interactions has been developed that provides a mapping of the molecular interactions onto a pairwise decomposable many-body potential of mean force (PMF) in the CG coordinate space by matching the effective CG forces to the forces along fully atomistic simulations.[7−15] The method is called multiscale coarse graining (MS-CG) and has been applied to a variety of complex condensed-phase and biomolecular systems.[1]

In many instances, however, CG models are not sufficiently accurate so that atomistic or dynamics become necessary for some key part of the system. This fact has prompted the development of mixed resolution methods where the system is partitioned into different domains. In quantum mechanics/molecular mechanics (QM/MM) approaches, for example, a small subset of the system is treated quantum mechanically, while the rest is treated by a classical empirical potential. A common simplification in QM/MM, and in many mixed-resolution methods, is the prevention of particle exchange between domains. Such a restriction clearly limits the scope of their applications, for example by excluding systems with strong fluctuations and diffusive properties. Realistic modeling of particle flow across boundaries in a mixed-resolution representation requires the ability to adjust a particle's resolution as it moves across the predefined resolution boundaries. Several adaptive resolution (AdRes) schemes have recently emerged in response. The so-called "hybrid particle" in such schemes is represented by a weighted superposition of its various representations and interactions.[1,16−22] (The particle adopts its hybrid form only inside transition regions, which separate regions of different resolution.) The major feature distinguishing the various AdRes schemes is how interactions between domains of different resolutions are coupled in the transition region. In the potential-based AdRes method, effective forces are determined from a smoothly varying potential across the resolution transition region. In the force-based scheme, the reverse approach is adopted.

The present paper describes a significantly different strategy to couple CG and atomistic descriptions, which we call mixed resolution interaction (MRI) modeling. The general idea is to implement different resolution models depending only on the radial distance between the interacting particles. MRI modeling is based on the intuitively obvious fact that the interactions between sufficiently distant separated particles, belonging to the so-called "CG distance zone", should be well described by a lower resolution CG force field. As the particles move closer together, the CG force field should be smoothly replaced by a fully resolved atomistic force calculation. The overall MRI approach assumes that the CG forces can be mapped (or "unfolded") onto atomistic forces as two particles move out of the atomistic distance zone. It should be stressed that in the MRI method both resolutions (atomistic and CG) "live" throughout the entire simulation system. There are no resolution boundaries in real space.

The MRI force unfolding algorithm addresses the question of how a CG force, normally applied to the center-of-mass of an atomistic group (or more generally to the location of a CG site), should be distributed over individual atoms of the group. The solution to this problem is not unique, a fact that introduces some (unavoidable) ambiguity into the MRI method. The nonuniqueness of the unfolding scheme is associated with the inevitable loss of information incurred upon the coarse graining of a system.[1] However, the advantage of MRI modeling over the alternative of a pure CG simulation implemented at all length scales is that the MRI simulation limits the loss of information to primarily the long-range interactions. Fully atomistic resolution in the structure is preserved at short range, and these atomistic short-range interactions are maintained in every part of the system, also in contrast to the AdRes schemes.

In practice, the computational efficiency of the MRI method is defined by two factors: the size of the atomistic zone and the simplicity of the force unfolding algorithm. The atomistic zone should encompass all parts of the free energy surface that are "molecularly rough". Conversely, the atomistic interactions may be safely replaced by CG interactions in regions where the characteristic length scale of variations in the effective free energy surface (many-body PMF) begins to exceed the linear dimension of the CG groups. One possible criterion for the latter might be gleaned from the level of agreement between MS-CG potentials developed for two different choices of CG sites within the groups (e.g., center-of-mass vs geometrical center). Agreement between the two MS-CG effective potentials at some interparticle distance would indicate that the CG forces could be unfolded into atomistic ones at that distance in the MRI scheme. Note that because MRI potentials are pairwise, MRI modeling can be combined with accurate atomistic interactions in a selected subsystem, in principle allowing for an efficient way to simulate heterogeneous systems.

The following sections of this article are structured as follows. Section 2 begins with a summary of the MS-CG method and then presents an overview of the MRI method. In Section 3.1, the MRI one-site water model is described, and MRI simulation of TIP3P water in the bulk phase is presented. In Section 3.2, the same is done for liquid methanol. In Section 4, the MRI TIP3P water potential's transferability to heterogeneous systems is analyzed by simulating an aqueous sodium ion solution. Also in Section 4, the MRI method is applied to simulate a phospholipid bilayer. The paper closes with conclusions in Section 5.

## 2. Mixed Resolution Interaction Method

In essence, the MRI approach is a distance-dependent coupling of the atomistic and CG force fields, which must be chosen to be as consistent as possible. The best result can be achieved if the CG interactions represent an ensemble average of the exact atomistic interactions projected upon the coarse-grained degrees-of-freedom. In this case, the configurationally averaged atomistic and CG interactions between distant molecules are close to identical. The MS-CG method, which has been extensively described elsewhere,[1,7,8,13−15] derives the effective CG interactions through a statistical mechanically consistent projection of the atomistic forces onto the CG space. It, thus, satisfies the above requirement.

**2.1. MS-CG Interactions.** Here we present a brief outline of the MS-CG method relevant to MRI modeling.[14,15] In the MS-CG approach, coarse graining of an $n$-particle Cartesian phase space ($\mathbf{r}^n$, $\mathbf{p}^n$) with a Hamiltonian $h(\mathbf{r}^n, \mathbf{p}^n)$ corresponds to mapping the coordinates of the $n$ particles into $N$ CG groups. This process is represented as a canonical coordinate transformation to a new phase space ($\mathbf{R}^N$, $\mathbf{P}^N$). The latter is spanned by the intragroup locations of the $N$ atomic groupings:

$$\mathbf{R}_I(\mathbf{r}^n) = \sum_{i=1,s_I} c_{Ii}\mathbf{r}_i \tag{1}$$

$$\mathbf{P}_I(\mathbf{p}^n) = \sum_{i=1,s_I} \mathbf{p}_i$$

where $s_I$ is the number of atoms in the $I$th grouping, and the coefficients $c_{Ii}$ satisfy:

$$\sum_i c_{Ii} = 1 \quad \text{for } I = 1, ..., N \tag{2}$$

Equation 2 ensures that the ($\mathbf{R}^N$, $\mathbf{P}^N$) coordinates are canonical. $\mathbf{R}_I$ and $\mathbf{P}_I$ will be the position and the momentum of the center-of-mass (CM) of the $I$th group if $c_{Ii} = m_{Ii}/M_I$, where $m_{Ii}$ is the mass of the $i^{\text{th}}$ atom and $M_I = \sum_{i=1,s_I} m_{Ii}$ is the total mass of group $I$. Alternatively, the choice $c_{Ii} = 1/s_I$ is equivalent to the assumption that all atoms in the group have equal mass. In this latter representation, $\mathbf{R}_I$ is the geometrical center of the CG group.

The effective pairwise interaction Hamiltonian, which describes the canonical distribution in the CG phase space ($\mathbf{R}^N$, $\mathbf{P}^N$), is constructed to represent the all-coordinate exact CG many-body potential of mean force $U(\mathbf{R}^N)$. This can be achieved by introducing a pairwise approximation to $U(\mathbf{R}^N)$ into the CG Hamiltonian. A pairwise form ensures that the CG effective potential term can be expressed as the integral of a pairwise and, thus, the conservative approximate CG force field as

$$\mathbf{G}_{\mathbf{R}_I}(\mathbf{R}^N,\Omega) = \sum_{J \neq I} g(R_{IJ},\Omega)\mathbf{n}_{IJ} \tag{3}$$

whose free parameters $\Omega$ are adjusted to approximate (in the least-squares sense) the all-coordinate CG many-body potential of mean force field:[1,15]

$$\mathbf{F}_{\mathbf{R}_I}(\mathbf{R}^N) = -\nabla_{\mathbf{R}_I} U(\mathbf{R}^N) \tag{4}$$

In eq 3, $R_{IJ}$ is the distance between CG groups $I$ and $J$ located at $\mathbf{R}_I$ and $\mathbf{R}_J$, respectively, and $\mathbf{n}_{IJ} = \mathbf{R}_{IJ}/R_{IJ}$ is the unit vector pointing from $J$ to $I$. The pairwise MS-CG force field $\mathbf{g}_{IJ}(R_{IJ}, \Omega) = g(R_{IJ}, \Omega)\mathbf{n}_{IJ}$ is, thus, central and radially symmetric. As shown elsewhere,[14,15] if $\mathbf{G}_{\mathbf{R}_I}(\mathbf{R}^N, \Omega)$ [or the $\mathbf{g}_{IJ}(R_{IJ}, \Omega)$] is linear in $\Omega$ (i.e., if it can be expanded into a set of basis functions whose coefficients are the parameters $\Omega$), then a least-squares fit to the mean net forces $\mathbf{F}_{\mathbf{R}_I}(\mathbf{R}^N)$ is equivalent to a least-squares fit to their instantaneous magnitudes $\mathscr{F}_{\mathbf{R}_I}(\mathbf{R}^N)$. The latter can be evaluated from an atomistic simulation as $\sum_{i=1,s_I}\mathbf{f}_i$, where the $\mathbf{f}_i$ are instantaneous net forces on individual atoms of the group. The least-squares problem is thereby reduced to a set of linear equations,[8,15] which can be solved using a block-averaging scheme.[7−9] A linear basis set can be conveniently constructed using piecewise functions, for example, a spline representation.[7−9,15]

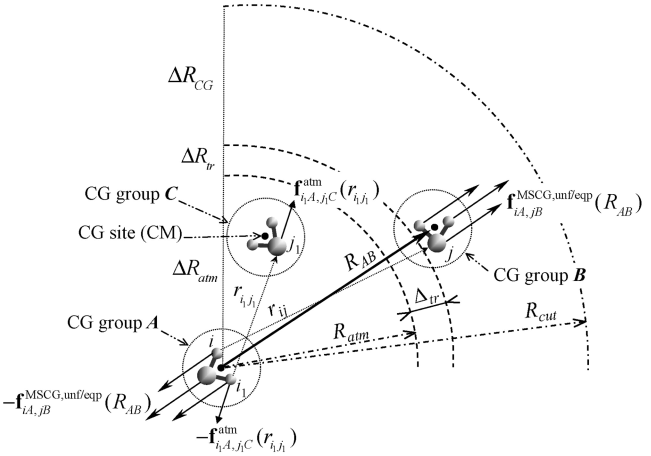

$\Delta R_{CG}$
$\Delta R_{tr}$
CG group ***C***
CG site (CM)
$\Delta R_{atm}$
CG group ***A***
$\mathbf{f}^{\text{atm}}_{i_1A,j_1C}(r_{i_1j_1})$
$\mathbf{f}^{\text{MSCG,unf/eqp}}_{iA,jB}(R_{AB})$
$j_1$
$R_{AB}$
$j$
CG group ***B***
$r_{i_1j_1}$
$r_{ij}$
$\Delta_{tr}$
$R_{cut}$
$R_{atm}$
$i$
$i_1$
$-\mathbf{f}^{\text{MSCG,unf/eqp}}_{iA,jB}(R_{AB})$
$-\mathbf{f}^{\text{atm}}_{i_1A,j_1C}(r_{i_1j_1})$

***Figure 1.*** Schematic of the equipartition approach for evaluating pairwise atomistic contributions given the total MRI force acting between two CG groups (eqs 6, 9−11).

**2.2. MRI-Coupling Atomistic and MS-CG Interactions.** This section describes the mixed resolution force-based scheme employed to couple atomistic and MS-CG interactions. At the same time, it suggests a potential-based scheme, whose relation to the force-based scheme is similar to that seen in the AdRes methods.[22,23] While the force-based scheme is generally easier to implement, it does have a drawback related to the difficulty of defining a potential for the mixed (coupled) force field.

The location chosen for a CG site within the group is assumed here to be to as its CM, thus, assuming a proper distribution of atomic masses. The interval between two CMs is partitioned into three regions, as shown in Figure 1: the atomistic zone $\Delta R_{\text{atm}}$; $R \le R_{\text{atm}}$, the transition region $\Delta R_{\text{tr}}$; $R_{\text{atm}} < R \le R_{\text{atm}} + \Delta_{\text{tr}}$, and the CG zone $\Delta R_{\text{CG}}$; $R_{\text{atm}} + \Delta_{\text{tr}} < R \le R_{\text{cut}}$. The distance $R_{\text{atm}}$ is the cutoff for the atomistic zone, while $\Delta_{\text{tr}}$ is the width of the transition region. The distance $R_{\text{cut}}$, the cutoff for the CG zone, is taken to be equal to the cutoff used for the MS-CG interactions.[15] The purpose of the transition region is to join smoothly the forces in the atomistic and CG regions.

The total force acting on the $i^{\text{th}}$ atom in the $k^{\text{th}}$ CG group of type $\alpha$, with the latter two indices grouped as $A = (\alpha k)$, can be written as

$$\mathscr{F}^{\text{MRI}}_{iA}(r) = \sum_{j,B} \mathbf{f}^{\text{MRI}}_{iA,jB}(r, R_{AB}) \tag{5}$$

The summation runs over all groups $B = (\beta l)$, $B \neq A$, whose atoms $(jB) = p$ act on the atom $(iA) = s$ with nonzero forces $\mathbf{f}_{sp}^{\mathrm{MRI}}$, given that the atoms $s$ and $p$ are separated by a distance $r$ and the groups' CMs are $R_{AB}$ apart. In the force-based scheme, the force between atoms $s$ and $p$ depends on which zone the intergroup separation $R = R_{AB}$ falls into

$$\mathbf{f}_{sp}^{\mathrm{MRI}}(r,R) = \begin{cases} \mathbf{f}_{sp}^{\mathrm{atm}}(r); & R \in \Delta R_{\mathrm{atm}} \\ \mathbf{f}_{sp}^{\mathrm{atm}}(r)w(R) + \mathbf{f}_{sp}^{\mathrm{MSCG,unf}}(R)(1 - w(R)); & R \in \Delta R_{\mathrm{tr}} \\ \mathbf{f}_{sp}^{\mathrm{MSCG,unf}}(R); & R \in \Delta R_{\mathrm{CG}} \end{cases} \tag{6}$$

In eq 6, $\mathbf{f}_{sp}^{\mathrm{atm}}(r)$ denotes the fully resolved atomistic force:

$$\mathbf{f}_{sp}^{\mathrm{atm}}(r) = \mathbf{f}_{sp}^{\mathrm{sr}}(r) + \mathbf{f}_{sp}^{\mathrm{coul}}(r), \tag{7}$$

where $\mathbf{f}^{\mathrm{sr}}$ is a short-ranged (e.g., Lennard-Jones) contribution, and $\mathbf{f}^{\mathrm{coul}}$ is the electrostatic (Coulomb) component. The symbol $\mathbf{f}_{sp}^{\mathrm{MSCG,unf}}(R)$ in eq 6 represents a pairwise atomistic interaction derived from the CG interaction between groups containing the atoms $s$ and $p$, in accordance with the preselected unfolding algorithm for the pairwise CG force $\mathbf{f}_{AB}^{\mathrm{MSCG}} = \mathbf{f}_{\alpha\beta}^{\mathrm{MSCG}}(R_{AB})$ (the indices $\alpha$ and $\beta$ denote kinds of groups). To make the $\mathbf{f}_{sp}^{\mathrm{MRI}}(r, R)$ force continuous everywhere, a smooth weighting function $w(R)$ is defined in the transition region with boundary conditions $w(R_{\mathrm{atm}}) = 1$, $w(R_{\mathrm{atm}} + \Delta_{\mathrm{tr}}) = 0$. We have adopted here the weighting function of the form $\cos^2(\pi(R - R_{\mathrm{atm}})/2\Delta_{\mathrm{tr}})$, which is similar to that used in ref 16. Such a weighting function implements a particularly simple way to ensure an interpolation between $w = 0$ and $w = 1$ that is monotonic, continuous, and differentiable and has a zero slope at the boundaries of the atomistic and (more importantly) the coarse-grained regions. Apart from these requirements the precise functional form is not especially relevant.

**2.3. Unfolding MS-CG Interactions.** In unfolding the MS-CG force $\mathbf{f}_{AB}^{\mathrm{MSCG}} = \mathbf{f}_{\alpha\beta}^{\mathrm{MSCG}}(R_{AB})$ into atom−atom contributions $\mathbf{f}_{iA,jB}^{\mathrm{MSCG,unf}}(R_{AB})$, where each contribution represents the force on atom $(iA)$ due to atom $(jB)$ at distance $r_{ij}$, the following constraint must first be satisfied:

$$\mathbf{f}_{AB}^{\mathrm{MSCG}} = \sum_{i,j} \mathbf{f}_{iA,jB}^{\mathrm{MSCG,unf}}(R_{AB}) \tag{8}$$

Of course, eq 8 alone is not enough. Additional assumptions regarding the directions and the amplitudes of the unfolded atomistic forces are required. First we will assume that:

$$\mathbf{f}_{iA,jB}^{\mathrm{MSCG,unf}} = f_{iA,jB}^{\mathrm{MSCG,unf}} \mathbf{e}_{ij}^{AB} \tag{9}$$

where $\{\mathbf{e}_{ij}^{AB}\}$ is a preselected set of unit vectors defining the possible directions of the unfolded forces. Newton's third law states that $\mathbf{e}_{ji}^{BA} = -\mathbf{e}_{ij}^{AB}$ and $f_{jB,\,iA}^{\mathrm{MSCG,unf}} = f_{iA,\,jB}^{\mathrm{MSCG,unf}}$. The latter two rules ensure that the Newton's third law is fulfilled all the time in the MRI simulation. The most natural choices of $\mathbf{e}_{ij}^{AB}$ point back and forth along the vector $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ connecting the two atoms. However, such a choice would require an evaluation of the unique basis vector $\mathbf{e}_{ij}^{AB}$ for each pair of atoms and then inversion of the matrix $||(\mathbf{e}_{ij}^{AB} \times \mathbf{e}_{i'j'}^{AB})||$ to project the CG force $\mathbf{f}_{AB}^{\mathrm{MSCG}}$ onto the full basis $\{\mathbf{e}_{ij}^{AB}\}$. This approach would be computationally expensive. A simpler choice is to assume that all vectors in the basis have the same direction along the radius vector $\mathbf{R}_{AB} = \mathbf{R}_A - \mathbf{R}_B$, connecting the CMs of the two groups

$$\mathbf{e}_{ij}^{AB} = \mathbf{e}^{AB} = \mathbf{R}_{AB}/R_{AB} \tag{10}$$

A rule for how to project the $\mathbf{f}_{AB}^{\mathrm{MSCG}}$ force onto the vectors $\mathbf{e}^{AB}$ is still needed, as the latter is not linearly independent. An "equipartition" rule is probably the simplest choice, which distributes the MS-CG force equally over all atomic pairs:

$$f_{iA,jB}^{\mathrm{MSCG,unf/eqp}} = f_{AB}^{\mathrm{MSCG}}/(N_\alpha N_\beta) \tag{11}$$

where $N_\alpha$ and $N_\beta$ are the numbers of atoms in the groups of type $\alpha$ and $\beta$, respectively, and $f_{AB}^{\mathrm{MSCG}}$ is the modulus of $\mathbf{f}_{AB}^{\mathrm{MSCG}}$. This rule is depicted in Figure 1. Calculating the force between the two CG groups by eqs 8−11 is, thus, $\sim N_\alpha N_\beta/2$ times faster than evaluating the short-ranged interactions for each pair $(iA, jB)$.

The simplicity of the equipartition approach comes at a cost, however. First of all, it ignores the structure of the groups. For example, if any of the CG group atoms are charged, then the forces will not be consistent with the partial charge distribution over the group because interactions in the CG zone are mostly of an electrostatic origin. Second, the direction of the $\mathbf{f}_{ij}^{\mathrm{MSCG,unf}}$ force is always the same no matter how the groups are mutually oriented. Third, the unfolded forces can be noncentral, as the $\mathbf{R}_{AB}$ and $\mathbf{r}_{ij}$ vectors may be not collinear. The noncentrality of unfolded forces may lead to difficulties in introducing a corresponding potential energy, as will be discussed shortly.

Let us, thus, consider possible enhancements or alternatives to the equipartition algorithm, in particular approaches that better account for atomic partial charges and orientations, or those that yield central forces. One straightforward modification addressing the problem of atomic partial charge dependence would be to select the $\mathbf{e}_{ij}^{AB}$ vector, whose sign (direction) agrees with the product of the respective partial atomic charges $q_iq_j$. This modification to eq 11 leads to the unfolding rule:

$$\mathbf{e}_{ij}^{AB} = \mathbf{R}_{AB}/R_{AB}sgn(q_iq_j) \tag{12}$$

$$f_{iA,jB}^{\mathrm{MSCG,unf/coul}} = f_{AB}^{\mathrm{MSCG}} q_iq_j/(Q_\alpha Q_\beta)$$

where $Q_\alpha$ is the net charge of a group of type $\alpha$. This change seems well justified in situations where the direction of the force between atoms at large separations is determined by the sign of their partial charge product $q_iq_j$.

The equipartition approach can also be made to depend on orientation by redefining the forces in the CG zone. We write $\mathbf{f}_{AB}^{\mathrm{MSCG}}(R_{AB}, \mathbf{d}_A, \mathbf{d}_B)$, where $\mathbf{d}_{A(B)}$ is a vector representing the orientation (e.g., dipole moment) of group $A(B)$. To be consistent with the generic MS-CG force field $\mathbf{f}_{\alpha\beta}^{\mathrm{MSCG}}(R_{AB})$, this function has to satisfy the condition $<\mathbf{f}_{AB}^{\mathrm{MSCG}}(R_{AB}, \mathbf{d}_A, \mathbf{d}_B)>_{\mathbf{d}} = \mathbf{f}_{\alpha\beta}^{\mathrm{MSCG}}(R_{AB})$. In this context, the brackets represent a configurational average over group orientations. The following form meets this condition:

$$\mathbf{f}_{iA,jB}^{\mathrm{MSCG}}(R_{AB}, \mathbf{d}_A, \mathbf{d}_B) = f_{AB}^{\mathrm{MSCG}}/(N_\alpha N_\beta) D(R_{AB}, \mathbf{d}_A, \mathbf{d}_B) \qquad (13)$$

where

$$D(R_{AB},\mathbf{d}_A,\mathbf{d}_B) = 2\delta d_{\mathbf{d}_A,\mathbf{d}_B}(R_{AB}) + 1 \qquad (14)$$

In eq 14, the term:

$$\delta d_{\mathbf{d}_A,\mathbf{d}_B}(R_{AB}) = d_{\mathbf{d}_A,\mathbf{d}_B} - \langle d_{\mathbf{d}_A,\mathbf{d}_B}\rangle \qquad (15)$$

represents the instantaneous deviation of the group orientation from its ensemble average. The first term is defined in terms of inner products of vectors as follows:

$$d_{\mathbf{d}_A,\mathbf{d}_B} = (\mathbf{d}_A,\mathbf{d}_B) - 3(\mathbf{d}_A,\mathbf{n}_{AB})(\mathbf{d}_B,\mathbf{n}_{AB}) \qquad (16)$$

where the dipoles are separated by a distance $R_{AB}$ and $\mathbf{n}_{AB}$ is a unit vector along the radius vector $\mathbf{R}_{AB}$. Indeed, eq 13 produces the desired configurational average because

$$\langle D(R_{AB},\mathbf{d}_A,\mathbf{d}_B)\rangle_{\mathbf{d}} = 1 \qquad (17)$$

A factor of 2 was introduced in the definition of $D(R_{AB}, \mathbf{d}_A, \mathbf{d}_B)$ [eq 14] so that $D$ can adopt positive or negative values. The $\langle d_{\mathbf{d}_A,\mathbf{d}_B}\rangle$ term can be calculated in advance, for example, by taking an ensemble average over the same atomistic trajectories and mesh that were used to obtain the MS-CG potentials. More generally, a number of different functions $D(R_{AB}, \mathbf{d}_A, \mathbf{d}_B)$ might be constructed for eq 13 which satisfy the condition in eq 17. This scheme is referred to as the dipole-disorder, or more simply, the dipole algorithm.

The last drawback of the equipartition algorithm, noncentrality of the unfolded forces, can be circumvented by assigning the force of eq 11 to only one pair of atoms: the $i$, $j$ (i.e., $N_\alpha$, $N_\beta = 1$ and $\mathbf{e}_{ij}^{AB} = \mathbf{r}_{ij}/r_{ij}$) that lie closest to the CMs of their respective atomic groups. It is then assumed $f^{\mathrm{MSCG,unf}} = 0$ for the rest of the atom pairs. For example, in the case of water, the unfolding forces can be assumed to be nonzero only for the O−O pair. Such a scheme may be sufficiently accurate for systems with smaller CG groups, such as water molecules.

To summarize, we note that the MRI model described in this section can be viewed as a mixed resolution generalization of cutoff treatments where the long-ranged interactions depend on the charge (or neutrality) of the groups. The underlying idea is similar in that interactions between atoms are switched to CG interactions when the separation between two groups becomes larger than a predefined cutoff. They differ because the resolution of the interaction changes in the present MRI approach.

## 3. MRI Models of Bulk Polar Solvents

**3.1. Water.** The accuracy of MRI modeling will first be explored for TIP3P bulk water under ambient conditions ($T$ = 300 K and $p$ = 1 bar). The interactions in the CG zone are represented by one-site MS-CG potentials similar to those presented in ref 8. The reference system used to obtain the MS-CG interactions consisted of 512 molecules in a periodic, cubic volume simulated under constant *NVT* conditions and equilibrium density (1014.6 kg/m$^3$). Electrostatic interactions were treated using the Ewald summation method, and Lennard-Jones interactions were cut off at 1 nm. The configurations of this simulation were sampled each 0.1 ps, over a total of 100 ps. Two effective CG interactions were derived from the same atomistic trajectory/force data: one where the CG sites were located at the group CM, and another where they were located at the group geometrical centers (GC). MS-CG forces were determined using the block-averaging method on a linear mesh covering distances up to 1 nm, with a bin size of 0.005 nm. The result was insensitive to block size.

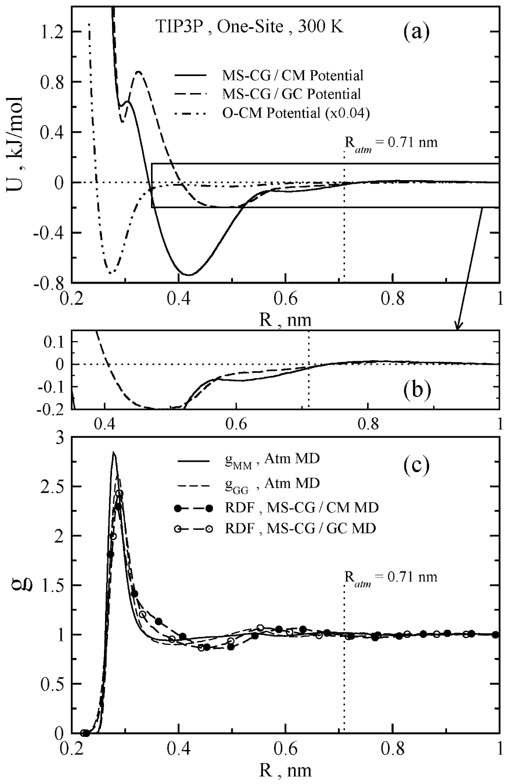


***Figure 2.*** (a): One-site effective CG interactions associated with CM (solid), GC (dashed) sites for bulk TIP3P at ambient conditions and configurationally averaged interaction between an oxygen atom (O) and a water molecule scaled by a factor of 4 × 10$^{-2}$ (dot-dashed). (b): Magnified tail region of potentials from (a). (c): CM and GC RDFs, $g_{\mathrm{MM}}$, $g_{\mathrm{GG}}$ from the reference atomistic (solid and dashed, respectively), MS-CG/CM (filled circles), and MS-CG/GC (empty circles) simulations.

These two CG potentials, hereafter referred to as the CM and GC potentials $U_{\mathrm{CM}}^{\mathrm{MSCG}}$ and $U_{\mathrm{GC}}^{\mathrm{MSCG}}$, respectively, are shown in Figure 2. The same plot shows CM and GC radial distribution functions (RDFs) (denoted by $g_{\mathrm{MM}}$ and $g_{\mathrm{GG}}$, respectively) derived from the reference atomistic MD and MS-CG simulations (the latter being lines labeled as "MS-CG/CM(GC)" in the legend). The RDFs resulting from the CG simulations are rather different from the structures seen in the reference atomistic MD simulation. The MS-CG/GC and atomistic $g_{\mathrm{GG}}$ RDFs exhibit a reasonably close match − the positions of the first and second solvation shells are well captured in the MS-CG simulation. The CM and CG potentials also nearly coincide at short separations ($r < 0.285$ nm), that is, up to the position of the first peak in $g_{MM}$ or $g_{GG}$. Their correspondence indicates that hydrogen bonding in a well-defined arrangement of interacting molecules is the dominant contributor to MS-CG potentials at these separations. Beyond the first peak, the molecular orientations become increasingly random causing the two potentials to differ. The CM and GC potentials also coincide at large ($r > 0.71$ nm, vertical-dotted lines in Figure 2) separations, that is, roughly the outer edge of the second solvation shell. The CM potential is substantially more attractive at intermediate distances, with a depth of almost − 0.75 kJ/mol compared to − 0.25 kJ/mol for the GC potential. The overall neutrality and small size of the water molecule implies a fast decay of the effective one-site interaction and also explains its insensitivity to the location chosen for CG sites at relatively short separations. In Figure 2a, we have plotted an effective interaction between an oxygen atom and a water molecule as a function of O−CM distance. The effective potential was obtained by averaging the O−$H_2O$ potential energy over all configurations along the reference atomistic trajectories.

The CM and GC potentials start showing similarities at 0.71 nm, suggesting that this distance should be the size of the atomistic zone $\Delta R_{atm}$. Due to agreement of the CM and GC MS-CG potentials at distances larger than 0.71 nm, one might expect that convergence of the properties in the MRI simulations can be reached with an atomistic zone radius $R_{atm}$ larger than that distance. As will be discussed later, this is indeed the case. We, therefore, carried out MRI simulations using both MS-CG/CM and MS-CG/GC potentials and cutoff to the atomistic zone at $R_{atm} = 0.05$, 0.35, 0.5, 0.7, and 0.75 nm (the models are labeled MRI/CM(GC)/$R_{atm}$). All models used a transition zone of width $\Delta_{tr} = 0.1$ nm. We note that standard deviation of temperature in the *NVT* simulation for $R_{atm} = 0.7$ nm was similar to that found in the Ewald atomistic simulation (7.7 K). For $R_{atm} = 0.35$ nm, the standard deviation of temperature increased to 8.6 K.

The MRI/CM and MRI/GC models with almost no atomistic zone ($R_{atm} = 0.05$ nm), so that all interactions are unfolded from the MS-CG potential, performed very differently. The MRI/CM/0.05 simulation exhibited glassy dynamics, with overstructured RDFs (the first peak in $g_{OO}$ is at 6.0), and a diffusion coefficient several orders of magnitude lower than the atomistic simulations predict. By contrast, the MRI/GC/0.05 simulation yielded accelerated dynamics, understructured RDFs (max $g_{OO} = 1.4$) and diffusions almost twice as fast as the fully CG MD simulation, or 10 times faster than the atomistic diffusion. The internal pressure was large and positive in both cases. Clearly, these results demonstrate that it is essential in the MRI method to have the atomistic zone of a sufficient size in order to achieve accurate results.

Figure 3 compares the structural properties (O−O RDFs) of TIP3P bulk water obtained from MRI simulations using other values of $R_{atm}$ as well as the reference atomistic and the fully MS-CG data. The MRI simulations were run in the constant *NPT* ensemble for $R_{atm} = 0.7$ and 0.75 nm and run in the constant *NVT* ensemble for smaller values of $R_{atm}$. MRI/CM simulations produced a slightly worse liquid structure for $R_{atm} \leq 0.5$ nm, a direct consequence of the less-accurate structure produced by the MS-CG MD simulation with the bare CM potential (see Figure 2). Diffusion coefficients and some other thermodynamic properties from the MRI/CM and MRI/GC simulations are summarized in Table 1. All MRI models accurately modeled the first solvation shell, but the structure, dynamics, and thermodynamic properties of the liquid only became good as $R_{atm}$ approached 0.71 nm (i.e., the distance beyond which the CM and CG potentials begin to coincide, cf. Figure 2b). As can be seen from the data shown in the Supporting Information (Table 1S), simulations using the Coulomb and dipole unfolding schemes for $R_{atm} > 0.7$ nm performed similar to the equipartition unfolding. At smaller values of $R_{atm}$, the dipole unfolding scheme in the constant *NVT* simulation yielded a slightly better structure and dynamics compared to those of the equipartition and Coulomb schemes. Unfortunately, including orientation-dependent interactions in the CG zone via dipole unfolding did not improve the pressure-related properties of the liquid.

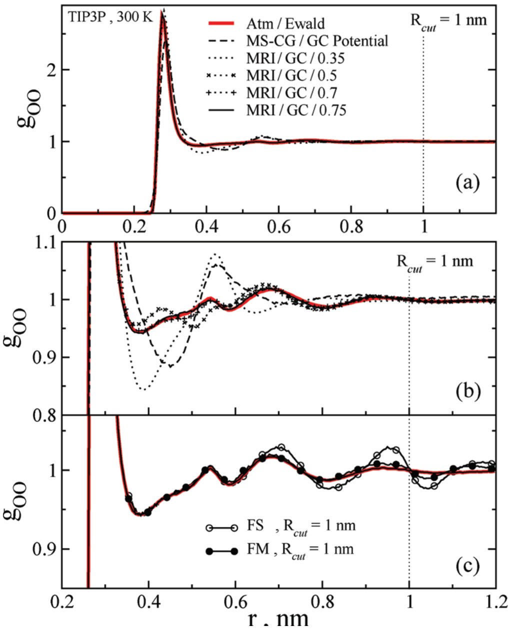


***Figure 3.*** Comparison of the oxygen−oxygen structure in TIP3P water at ambient conditions, simulated with the atomistic Ewald (red), MS-CG/GC (dashed), and MRI/GC interactions. In the latter cases, various sizes of the atomistic zone are also shown: $R_{atm}$ = 0.35 (dotted), 0.5 (crosses), 0.7 (pluses), and 0.75 nm (solid). (c) Atomistic Ewald and MRI structures are compared to an atomistic structure obtained using the FS (empty circles) and FM electrostatic descriptions (filled circles).

Figure 4 shows the distribution of the three-body orientational order parameter:

$$q = 1 - \frac{3}{8}\sum_{j=1,3}\sum_{k=j+1,4}\left(\cos \mathrm{O}_j\mathrm{O}_i\mathrm{O}_k + \frac{1}{3}\right)^2 \qquad (18)$$

where $O_jO_iO_k$ is the angle formed by the oxygens of a central molecule and two of its four closest neighbors.[24] The average $\langle q \rangle$ is a measure of tetrahedral order, where a value of 1 indicates a perfect tetrahedral network and a value of 0 indicates uncorrelated particles. The MRI models whose

***Table 1.*** Properties of Bulk TIP3P Water at Ambient Conditions[a,b]

| model, $R_{atm}$(nm) | atm | 0.35 | 0.50 | 0.70 | 0.75 |
|---|---|---|---|---|---|
| $N_{sol}$/min $g_{OO}$ | 7.2/0.377 | 6.4/0.363<br>7.9/0.388 | 6.8/0.368<br>6.53/0.368 | 7.2/0.377<br>7.2/0.377 | 7.2/0.377<br>7.2/0.377 |
| $\rho(\delta\rho)$ [kg/m$^3$]/$P$ (bar) | 1014.6(11.8)/3.7 | .../−1658.4<br>.../+641.0 | .../−915.9<br>.../−687.3 | 1019.1(12.0)/3.7<br>1020.0(12.0)/3.7 | 1015.6(11.8)/3.7<br>1015.1(11.8)/3.7 |
| $D_s$ ($10^{-9}$m$^2$/s) | 5.3 | 6.3<br>8.1 | 5.7<br>5.5 | 5.4<br>5.4 | 5.4<br>5.4 |
| $U^{conf}(\delta U^{conf})$ [kJ/mol] | −41.1(0.3) | −44.8(0.4)<br>−43.0(0.4) | −45.2(0.4)<br>−45.2(0.4) | −45.9(0.4)<br>−46.0(0.4) | −45.9(0.4)<br>−46.0(0.4) |
| $\kappa_T$ ($10^{-5}$ bar$^{-1}$) | 4.9 | −<br>− | −<br>− | 5.0<br>5.0 | 4.9<br>4.9 |

[a] As determined from atomistic MD (atm) with Ewald summation and MRI simulations for various atomistic zone sizes $R_{atm}$. [b] Among the MRI models (columns three through six), the first line is for center-of-mass CG sites (MRI/CM), and the second is for geometrical center CG sites (MRI/GC). The properties shown are $N_{sol}$/min $g_{OO}$, [number of molecules in first solvation shell]/[first minimum in $g_{OO}$]; $\rho(\delta\rho)/P$, [density]/[equilibrium pressure], where the standard deviation of density is shown in parentheses if the constant *NPT* ensemble is used, and the symbol "..." denotes that the constant *NPT* was not used. Additional tabulated results are: $D_s$, the self-diffusion coefficient; $U^{conf}(\delta U^{conf})$, the average configuration energy per particle (and its standard deviation in parentheses); and $\kappa_T$, the isothermal compressibility.

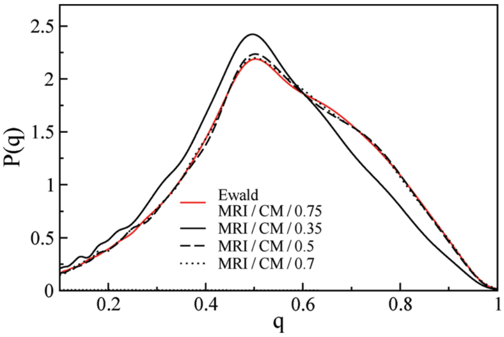


***Figure 4.*** Distribution of the order parameter $q$, $P(q)$ from eq 18, from the atomistic Ewald (red) and MRI/CM simulations with $R_{atm}$ = 0.35 (solid), 0.5 (dashed), 0.7 (dotted), and 0.75 nm (also red). The last is essentially identical to the exact Ewald simulation.

atomistic zones extend beyond the radius of the first solvation shell accurately reproduce the atomistic $q$ distribution (giving values of $\langle q \rangle = 0.63$ for both the atomistic Ewald and MRI/0.75 simulations). In the simulation with $R_{atm} = 0.35$ nm, a distance which just reaches the first solvation shell, the distribution of $q$ shifted to smaller values ($\langle q \rangle = 0.52$). This result suggests a decreased probability of finding nearest neighbors arranged tetrahedrally and, thus, suggests a softer hydrogen bonding. As reported in one of our previous papers,[25] adding a cutoff radius to the electrostatic interaction with force-shifted potentials does not affect the distribution of $q$ for TIP3P water. However, a short-ranged model developed using the inverse MC method[26] appears to be much less accurate than MRI models with $R_{atm} > 0.7$ nm.

Another informative quantity is the deviation between the MRI and atomistic force fields, in terms of the total force acting on each atom in a configuration. Figure 5a shows the time evolution of the ratio between the Euclidean norm of this deviation and the atomistic force:

$$\Delta F = \frac{(\sum_{I \in \text{conf}} \|F_I^{\text{cutoff}} - F_I^{\text{Ewald}}\|)^{1/2}}{(\sum_{I \in \text{conf}} \|F_I^{\text{Ewald}}\|)^{1/2}} \tag{19}$$

where the sum is over all atoms in the configuration. On the same plot, the ratio $\Delta F$ is evaluated for simulations using two cutoff electrostatic potentials:[25] one using a force-shifted (FS) and the other using a force-matched (FM) correction with $R_{cut}$ = 1.0 nm. (The FM correction is a short-ranged representation of Ewald electrostatics, obtained through force-matching Ewald trajectories in the bulk SPC/E water under ambient conditions.)[25] Even for the MRI potential with an atomistic zone as large as 0.75 nm, the average error in the MRI forces is 3.5 times larger than that of the error produced with FS electrostatics. Figure 5b shows the relative contribution of the unfolded to the total forces in MRI models with various $R_{atm}$ values and unfolding schemes. The CG contribution is consistently small: for the equipartition scheme, it is $8 \times 10^{-3}$ at $R_{atm} = 0.3$ nm and less than $1 \times 10^{-3}$ at $R_{atm} > 0.7$ nm. The constant *NPT* simulation with a neutral group-based cutoff, however, produces significantly denser water (by about 2−3%) than the Ewald simulation. This error noticeably and adversely affects the liquid structure, but surprisingly the diffusion rate is close to the atomistic value. In the constant *NVT* simulation at the Ewald equilibrium density, diffusion slowed to $4.7 \times 10^{-9}$ m$^2$/s. The MRI potential with $R_{atm} = 0.7$ nm, however, reproduced the target density to within 0.4−0.5%, and the rate of self-diffusion to within 2%. The results further improved with the MRI/0.75 model, as seen in Table 1. These observations highlight the importance of a repulsive component to the effective one-site MS-CG potentials at distances $R > 0.7$ nm (see Figure 2b), which is absent from the effective interactions that ignore many-body correlations, such as group-based cutoffs.

**3.2. Methanol.** In this section, we consider one-site and two-site MS-CG models of liquid methanol ($CH_3OH$), similar to those reported in ref 8. A one-site MS-CG model with the CG site assigned to the molecular CM was previously shown to capture the structure of liquid methanol far better than that of a similar CG treatment of water.[8] A two-site model mapping the OH and $CH_3$ groups to CG sites also yielded good structural properties.

The reference atomistic simulation consisted of 125 MeOH molecules at $T$ = 300 K in the constant *NVT* ensemble, contained in a box having 2.048 nm side lengths, and equilibrated for 2 ns. The interactions were modeled using Ewald electrostatics and the OPLS-AA force field,[27] which

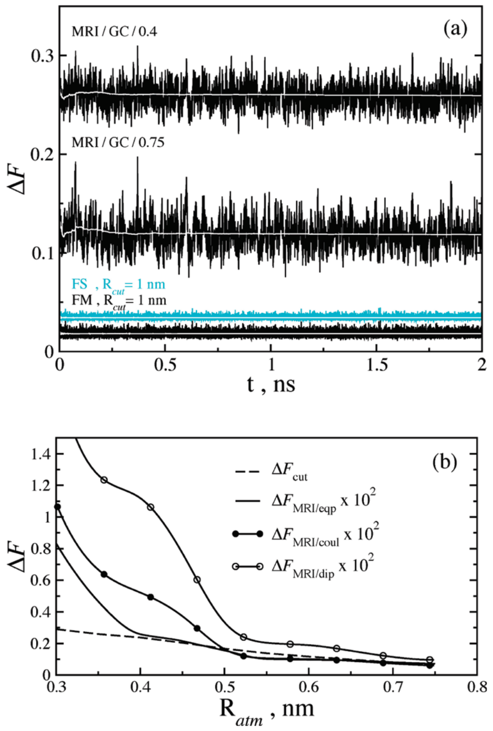


***Figure 5.*** (a): Time evolution of the relative error Δ$F$ in the Euclidean norm of the net forces (eq 19), resulting from MRI interactions with various atomistic zone cutoffs, compared to fully atomistic interactions. The net forces on each atom in the simulation at a given time step of the simulation are summed in calculating this error, and the atom trajectories are determined using a fully atomistic scheme. Data are also shown for reference atomistic simulations using the FS and FM description of electrostatics. The white lines show running time averages of each quantity. (b): Relative average contributions of the unfolded MS-CG/GC forces to the total forces using equipartition (solid), Coulomb (filled circles), and dipole (open circles) unfolding schemes.

is six site and flexible. The MS-CG procedure was essentially the same as for the water system. Figure 6a compares the effective MS-CG one-site interactions between molecular CMs and between GCs. The MS-CG/CM and MS-CG/GC force fields produced virtually identical RDFs to the reference atomistic RDFs ($g_{MM}$ and $g_{GG}$, in arbitrary units, are presented in the same figure).

As the MeOH molecule is polar, one might expect that the effective one-site MS-CG interaction is insensitive to location of CG site at sufficiently large separations. As seen in Figure 6a, this is indeed the case. The MS-CG/CM and MS-CG/GC potentials both decay rapidly out to about 0.5 nm, exhibit substantial differences in the intermediate region, and then become indistinguishable (within the tolerance of the MS-CG procedure) at $R = 0.71$ nm. This threshold is surprisingly close to that of the water one-site model. By contrast, the MS-CG CM and GC interactions in a two-site representation, shown in Figure 6b, differ at all distances within 1.0 nm because the OH and $CH_3$ groups are charged and because the interaction between them decays less rapidly with distance.

For the one-site potentials, the structures generated by the MRI/CM and MRI/GC potentials improved progressively for larger values of $R_{atm}$, as demonstrated in Figure 7. The reason for the convergence of the properties is the agreement between the CM and GC one-site MS-CG potentials at distances larger than 0.71 nm, as discussed earlier for the case of water. For $R_{atm} = 0.75$ nm, the structure was virtually identical to the atomistic case. In contrast with water, however, the one-site methanol MRI potentials failed to reproduce the reference diffusion and internal pressure (and, therefore, density) as shown in Table 2. Diffusion in the MRI/CM/0.75 simulation was about 1.5 times faster, and the density was about 3.7% higher (if an isothermal compressibility of 12.2 × $10^{-5}$ bar$^{-1}$ is assumed). However, the equilibrium pressure was positive in the MRI/CM/0.5 simulation and negative in the MRI/CM/0.75 simulation. One possible explanation for the difficulty with pressure reproduction in the MRI simulation is a large size of the MeOH molecules, as compared to water. Thus, an intermediate value of $R_{atm}$ would reproduce the correct density in the MRI simulation.

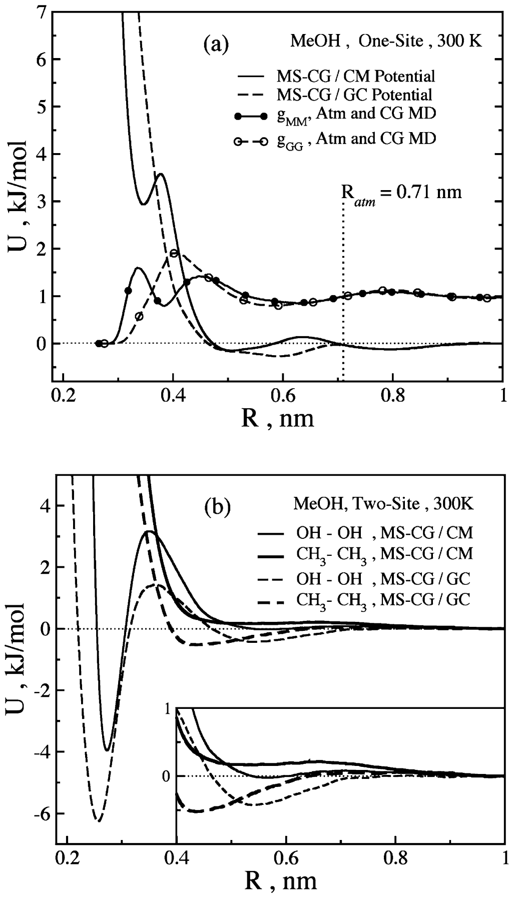


***Figure 6.*** One-site (a) and two-site (b) MS-CG effective interactions associated with the CM (solid) and GC (dashed) sites for bulk MeOH at ambient conditions. The CM and GC RDFs from the reference atomistic and MS-CG (filled and empty circles) simulations are depicted in arbitrary units for one-site coarse graining.

## 4. Atomistic/MRI Modeling: Solutes in MRI Water

In terms of geometry and charge distribution, the MRI effective potentials are expected to be short ranged for small, neutral groups. In principle, to simulate a system containing

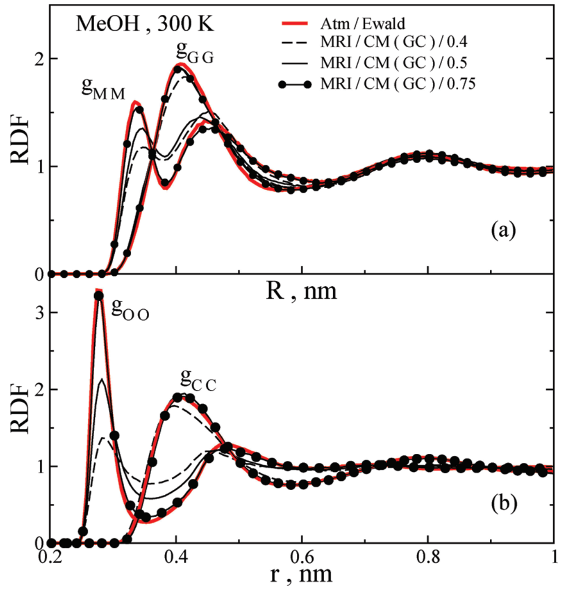


***Figure 7.*** Comparison of the CM−CM structure (a) and atom−atom structure (b) in the liquid MeOH at ambient conditions, simulated with the atomistic Ewald (red) and MRI models with various sizes of the atomistic zone: $R_{atm}$ = 0.4 (dashed), 0.5 (thin solid), and 0.75 nm (filled circles).

***Table 2.*** Properties of Bulk Liquid Methanol at Ambient Conditions[a,b]

| model, $R_{atm}$ (nm) | atm | 0.40 | 0.50 | 0.75 |
|---|---|---|---|---|
| $P$ (bar) | 6 | 5 386 | 2 587 | −302 |
| $D_s$ ($10^{-9}$ m$^2$/s) | 2.31 | 9.35 | 5.05 | 3.36 |

[a] As determined from atomistic MD (atm) with Ewald summation and one-site MRI/CM simulations with various atomistic zone sizes $R_{atm}$. [b] The atomistic simulations were in the constant *NPT* ensemble, while the MRI simulations were in the constant *NVT* ensemble at equilibrium atomistic density (773.7 kg/m$^3$). The properties listed here are $P$, the equilibrium pressure, and $D_s$, the self-diffusion coefficient.

charged species/groups, the MRI description may be introduced for neutral species, while the rest are treated on the atomistic level. In water solutions, for instance, the water−water interaction can be modeled using MRI, while the water-solute and solute−solute interactions are described atomistically. Such a description would be especially justified for bulky or charged solutes in solvents, such as water with low molecular weight. In many systems, such as biomolecular simulations, solvent−solvent interactions are the most computationally intensive aspect of the model. In such cases, MRI modeling of water may substantially increase simulation efficiency.

As the water and methanol examples have shown, MRI modeling is more efficient in scenarios where the effective intergroup interactions decay faster. Rapid decay is indicative of dominant high-order terms in the multipole expansion. The two lowest order multipole potentials, those between point charges and between a point charge and a dipole, are both long range. In the periodic simulation geometry, they even contain contributions from infinitely remote particles via the Ewald summation. The MRI formalism may, therefore, be less accurate in describing low-order interactions. This issue will now be explored.

**4.1. Aqueous Ionic Solutions.** As an example of such a mixed description, we now consider an aqueous ionic solution. The water−water interaction can be efficiently treated using the MRI formalism with an adequately large atomistic zone, as discussed in Section 3.1. Each ion is also a simple group but charged, so water−ion and ion−ion interactions have to be treated atomistically. Use of the MRI water−water interaction implies that ion−water electrostatics should also be evaluated using an approximate scheme, such as cutoff electrostatics. The latter requirement holds further for the ion−ion interaction, as the ionic subsystem is charged. In the mixed atomistic/MRI models reported here, we, thus, used an electrostatic potential with a FM correction[25] of $R_{cut}$ = 0.9 nm. The FM correction has been shown to be transferable to different water models, thermodynamic conditions, and aqueous solutions, outperforming conventional cutoff treatments, such as (damped) force-shifted potentials.[25]

Two ionic solutions with approximately the same concentration were simulated. The first consisted of 4 sodium ions in 512 TIP3P water molecules, and the second had 12 ions in 1500 TIP3P molecules. The simulation volumes were periodic cubes of edge length 2.476 and 3.543 nm, respectively. In the reference atomistic simulations, carried out under constant *NVT* conditions, electrostatics were evaluated for the whole system using the Ewald method. In high-dielectric solvents, such as water, ionic charging free energies calculated by Ewald summation are largely invariant to the system size.[28] The use of counterions to maintain the charge neutrality of the system is, therefore, not necessary. In the atomistic/MRI model, the water−water interaction was described using various MRI/GC potentials with atomistic zones up to $R_{atm}$ = 0.75 nm in size. Note that the bulk properties of water are well reproduced by the $R_{atm}$ = 0.75 nm model, as discussed earlier.

The water solvation structure of the ion was reproduced accurately by the mixed MRI/atomistic simulation, as can be seen in the Supporting Information (Figure 1S). The only difference was a slightly overstructured first solvation shell: max [$g_{Na^+-O}$] = 7.2 in the MRI model compared to 6.9 in the fully atomistic Ewald simulation. This small difference is likely a consequence of our use of cutoff electrostatics in the atomistic interaction set for the mixed simulation. Despite the good structure of the MRI water, the structure of ionic association was drastically altered. Figure 8a compares the ion−ion RDFs of the reference atomistic and MRI/GC water simulations. In the latter, the ions show an excessive tendency to aggregate, which grew progressively stronger for smaller values of $R_{atm}$. Figure 8b shows the effective solvent-free ion−ion potentials, obtained using the MS-CG method. The results are similar to the all-site MS-CG potentials and, like them, can be interpreted as an approximate pairwise decomposition of the all-ion PMF. For complexes of two ions, the solvent-free ionic MS-CG potential is simply a pairwise PMF.

In the MRI water simulation with $R_{atm}$ = 0.4 nm, the ions organized into a close-packed structure with a binding energy of −17 kJ/mol. After rapid formation of the ionic complex, they remained confined within a distance of about $R_{atm}$ for the rest of the simulation. Even in the MRI water simulation

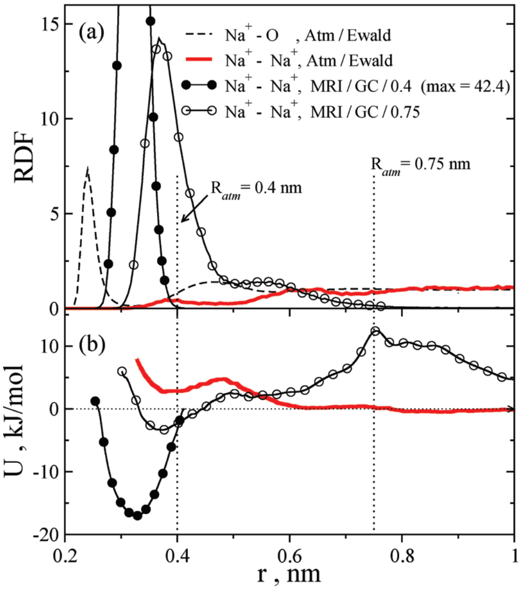
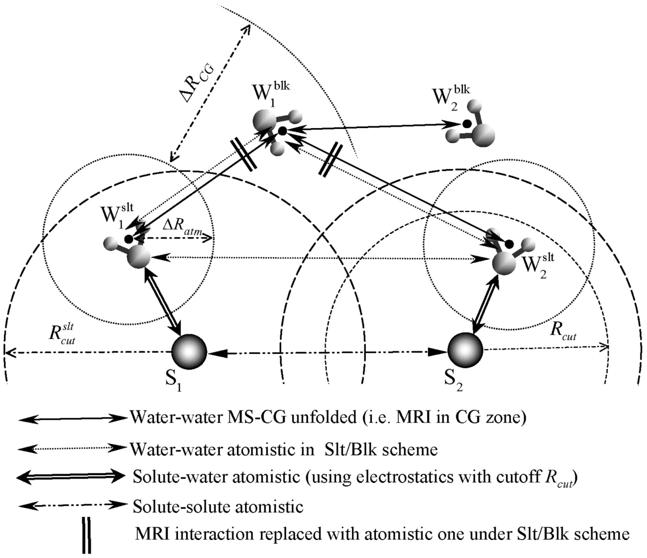


***Figure 8.*** Simulations of ions in water. Results are shown for a simulation with 4 $Na^+$ ions and 512 $H_2O$ molecules. (a): Ion−ion RDFs from atomistic Ewald MD (red), MRI/GC water with $R_{atm}$ = 0.4 nm (filled circles), and $R_{atm}$ = 0.75 nm (empty circles) simulations. The atomistic oxygen structure around the ions is also plotted (dashed line) as a reference. (b): Corresponding many-body MS-CG ion−ion effective interaction potentials.



Water-water MS-CG unfolded (i.e. MRI in CG zone)
Water-water atomistic in Slt/Blk scheme
Solute-water atomistic (using electrostatics with cutoff $R_{cut}$)
Solute-solute atomistic
MRI interaction replaced with atomistic one under Slt/Blk scheme

***Figure 9.*** Schematic description of the slt/blk scheme for a treatment of water−water interactions in the MRI water ionic solution.

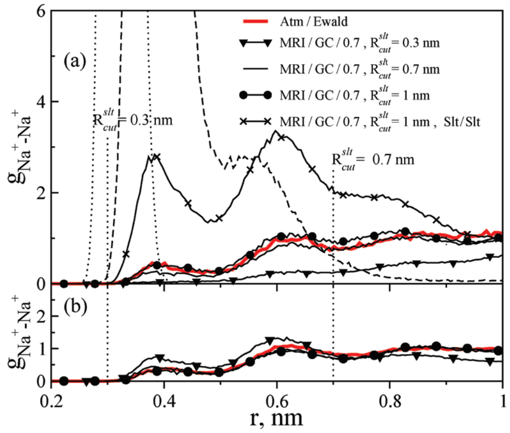


***Figure 10.*** Ion−ion structure for the same system (a), shown in Figure 8, and for a simulation with 12 ions in 1 500 $H_2O$ molecules (b). Both systems use the MRI/GC/0.75 model and the solute/bulk (slt/blk) scheme for water−water interactions (see Figure 9), with $R_{cut}^{slt}$ = 0.3 (triangles), 0.7 (thin solid), or 1.0 nm (circles). The solid line with crosses is for an MRI simulation using the solute/solute (slt/slt) scheme, with $R_{cut}^{slt}$ = 1.0 nm. The dotted and dashed lines, which correspond to the purely MRI modeling results shown as filled circle and empty circle lines from Figure 8, respectively, are shown for a comparison.

with $R_{atm}$ = 0.75 nm, the effective ion−ion interaction was significantly more attractive than that of fully atomistic data, with a globally stable contact minimum of −3.3 kJ/mol. The position of the minimum (0.38 nm), however, is close to the metastable contact minimum of the atomistic effective potential. The ion−ion effective interaction in the MRI/GC/0.75 water solution exhibits a high barrier with a maximum at about $R_{atm}$. A similar feature is probably present in the MRI/GC/0.4 model, but it could not be verified as the region outside $R_{atm}$ was undersampled. In the reference atomistic simulation, the ion−ion effective interaction almost vanishes at distances of about 0.6 nm.

A possible origin of such behavior is explained in Figure 9. The unfolded CG interaction between water molecules is an approximation to the atomistic water interaction in bulk (between molecules $W_1^{blk}$ and $W_2^{blk}$ in Figure 9). Meanwhile, structural correlations between water molecules inside and outside the neighborhood of a solute (i.e., the region within $R_{cut}$ of a solute) will be strongly perturbed by the presence of a charge (for example, between $W_{1(2)}^{slt}$ and $W_1^{blk}$). Thus, interactions in the perturbed zone of the water−water environment may be not adequately represented by the unfolded CG bulk interaction. This may lead to a more complex solvation structure and, therefore, different solute association dynamics.

The simplest remedy for the spurious, excessive segregation behavior is to describe the interactions between water molecules from the $R_{cut}^{slt}$ neighborhood (denoted $W_{1(2)}^{slt}$ in Figure 9) and between all other molecules within $R_{cut}$, including those from the “bulk” (e.g., $W_{1(2)}^{slt}$−$W_1^{blk}$ interactions), using the atomistic force field. The interaction between solvent molecules, if both are outside the $R_{cut}^{slt}$ solute region (e.g., between the $W_1^{blk}$ and $W_2^{blk}$ molecules), can be safely treated using the MRI formalism. Since the computational efficiency of such a treatment increases with the ratio of: (a) the number of water pairs formed by the bulk water molecules to (b) the number of pairs in which at least one molecule is within $R_{cut}^{slt}$ of the solute species, the suggested approach is suitable for systems with a large bulk water subsystem (e.g., biological molecules in water). This idea is referred to as the solute/bulk (slt/blk) scheme. Figure 10 shows the ion−ion structures of MRI simulations running under the slt/blk scheme. For $R_{cut}^{slt}$ = 0.7 nm, the ionic association dynamics were essentially identical to the fully atomistic. In the 12-ion slt/blk MRI simulation with $R_{cut}^{slt}$ = 0.3 nm, the ion−ion structure was reproduced with almost

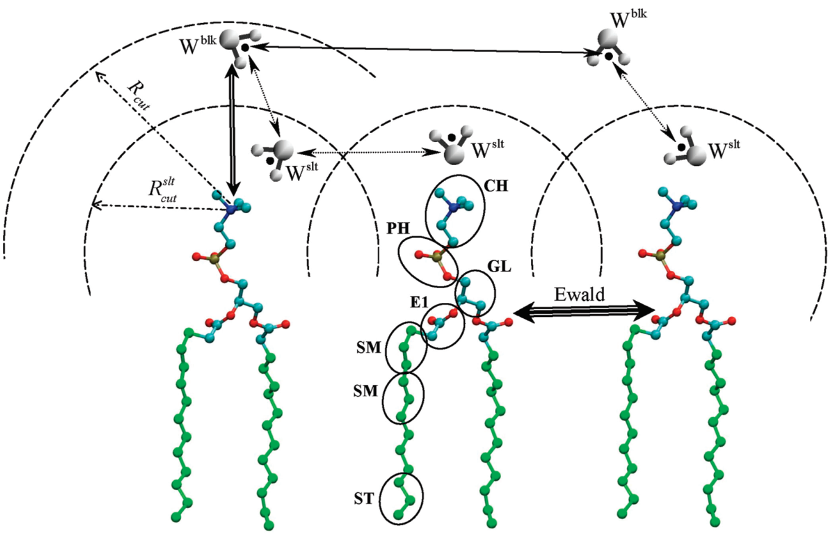


***Figure 11.*** A schematic model of the interactions used in the simulating a DMPC bilayer in MRI water, with the slt/blk treatment of water molecules at the interface. Line legend is same as in Figure 9.

atomistic quality, while in the 4-ion system with $R^{\mathrm{slt}}_{\mathrm{cut}} = 0.3$ nm, the structure was somewhat worse. In the 4-ion simulations under the slt/blk scheme, ionic diffusion was 1.05(0.01) $\times$ $10^{-9}$ m$^2$/s for $R^{\mathrm{slt}}_{\mathrm{cut}} = 0.7$ nm but attained 1.18 $\times$ $10^{-9}$ m$^2$/s for $R^{\mathrm{slt}}_{\mathrm{cut}} = 1.0$ nm. The latter almost exactly matches the atomistic MD value of 1.20 $\times$ $10^{-9}$ m$^2$/s. The same tendency was observed in the 12-ion simulations.

It could be argued that the slt/blk scheme goes too far in its atomistic treatment of water–water interactions, as the association dynamics of solutes are governed mainly by interactions between water molecules from the ionic solvation shells (note that it might be better to say "within the $R^{\mathrm{slt}}_{\mathrm{cut}}$ interaction shells"). However, it could be sufficient to accurately evaluate only interactions between pairs of water molecules that both lie within $R^{\mathrm{slt}}_{\mathrm{cut}}$ of the solute. This scheme, referred to as solute/solute (slt/slt), is certainly computationally cheaper. Unfortunately, as demonstrated in Figure 10a, the ion–ion structures are different under the slt/blk and slt/slt schemes. A simulation with slt/slt enhancement of the MRI water–water interactions yielded better solute dynamics than one without, but the structure was still significantly different from the fully atomistic results. In particular, excessive contact (first peak in the RDF) and solvent-separated (second peak) pairs were still apparent in the ionic solute RDFs. The best results were, therefore, produced by the slt/blk approach.

**4.2. Lipid Bilayer System.** In this section, the MRI water model coupled with the slt/blk algorithm is applied to a phospholipid dimyristoylphosphatidylcholine (DMPC) bilayer. The bilayer was represented by 64 DMPC molecules and by 1 312 water molecules, corresponding to a hydration of 21 $H_2O$ per DMPC. The DMPC molecules were modeled using a united atom atomistic force field.[29] This simulation used the rigid TIP3P water model with Lennard-Jones interactions for the hydrogen atoms set to zero. Electrostatic interactions were calculated via the particle-mesh Ewald (PME) summation.[30] The initial configuration was taken from ref 31 and then equilibrated for 20 ns in the constant *NPT* ensemble.

The bilayer geometry represents an infinite polarized interface between lipids and water, so the electrostatic contribution to configurational energy within the monolayer scales as $1/R_{\mathrm{cut}}$ at distances greater than the cutoff $R_{\mathrm{cut}}$. The effective FM electrostatic interaction in disordered water solutions decays much faster.[25] As discussed in the literature,[32] introducing a cutoff in the electrostatics may affect interactions between the normal components of headgroup dipoles. This issue could be a major factor, potentially affecting bilayer properties, such as surface equilibrium area and structure. It is, thus, best to treat the electrostatic interactions between lipids as accurately as possible, using the Ewald or PME methods. On the other hand, it has been noted that the effect of a cutoff on the bilayer properties is *less* important if the bilayer is properly hydrated (probably >27–28 molecules per lipid), and that, with proper hydration, the simulations of bilayers with a neutral group implementation do not suffer greatly from the omission of long-range electrostatics.[32] This observation suggests that our MRI water model could be accurate enough for the bilayer simulation.

An appropriate scheme for simulating the bilayer in MRI water is sketched in Figure 11. More specifically, water–water interactions were treated using the MRI/GC/0.7 model and the slt/blk scheme with $R^{\mathrm{slt}}_{\mathrm{cut}} = 0.6$ nm, as described in Section 4.1. The MS-CG/GC one-site water potential was recalculated for this simulation, using a fit to the water subsystem of the atomistic bilayer, and is, therefore, slightly different from the MS-CG/GC bulk potential studied in Section 3.1. One important difference is that the TIP3P potential in the bilayer simulation sets Lennard-Jones interactions for hydrogen to zero. The polarized (more ordered) water structure near the bilayer is a second major factor. Similar to the MRI modeling of ionic solutions, lipid–water interactions in the MRI modeling here were described atomistically with cutoff electrostatics using the FM correction, with $R_{\mathrm{cut}} = 0.9$ nm (see Figure 11). In this particular simulation geometry, about 30% of water–water pairs interacted through the MRI potential. The Ewald method was used to evaluate the electrostatic contribution to lipid–lipid interactions and can be considered accurate due to the overall neutrality of the lipid subsystem.

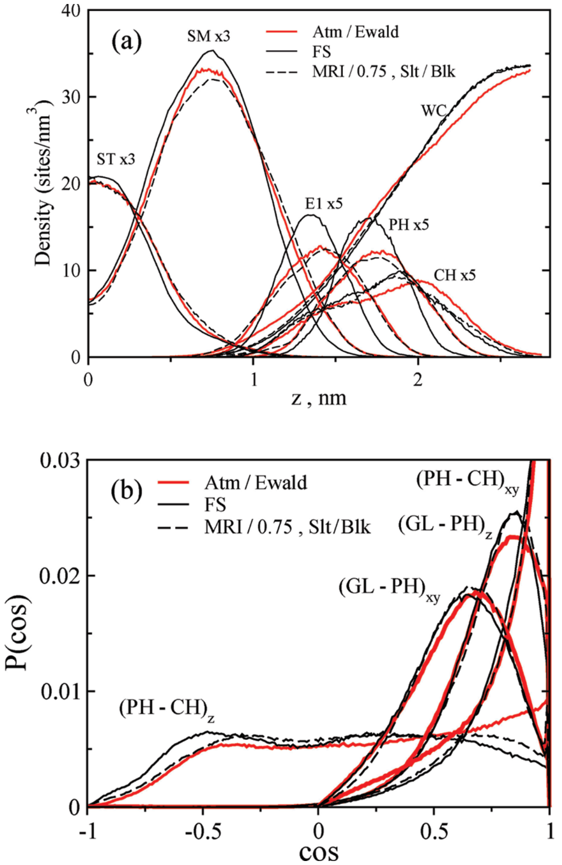


***Figure 12.*** Structure of the DMPC bilayer determined using atomistic Ewald (red), FS (solid thin), and MRI/0.75 with slt/blk water (dashed) simulations. (a): Bilayer normal density profiles of lipid groups and water (CM). (b): Distributions of the normal (*z*) and in-plane (*xy*) components of vectors connecting headgroup sites, PH−CH and GL−PH. The partitioning of a DMPC molecule into CG groups is shown in Figure 11. The symbols are: CH, choline moiety; PH, phosphate group; GL, glycerol backbone; E1, ester group at *sn* − 1 chain; and SM/ST, three-carbon groups of the acyl chains.

A comparison of the reference atomistic simulation using the Ewald method, the MRI simulation, and an atomistic MD simulation using the FS cutoff ($R_{\text{cut}} = 1.0$ nm) is shown in Figure 12. In Figure 12a, we compare the density profiles of water and lipid groups, in accordance with the partitioning adopted in ref 7 and shown in Figure 11. Figure 12b compares the distributions of normal and in-plane group orientations obtained by the three simulations. Note that the FS cutoff and Ewald simulations produced very different normal density distributions. In the cutoff simulation, the DMPC density of headgroup sites was more localized, and the water also penetrated less into the headgroup region, resulting in a lower water density outside the center of the water region. This water density profile is consistent with a positive lateral pressure at the center of the water region.[33] The lower hydration level of headgroups in the cutoff simulation (compared to the Ewald simulation) affected bilayer properties, such as equilibrium area per lipid. The MRI water density follows virtually the same pattern. Thus, it is fair to say that the loss of medium- and long-ranged lipid−water and water−water interactions causes weaker hydration of hydrophilic lipid sites in both the cutoff and MRI simulations.

A surprising result is that the MRI simulation reproduced the normal density of headgroup sites much better than that of the FS cutoff simulation. This advantage is likely a result of using the Ewald method to treat electrostatics within the lipid subsystem. The headgroup orientations were also slightly more accurate in the MRI simulation than in the FS cutoff simulation. For example, the distribution of the normal component of PH−CH was in better agreement with the Ewald results. This result is important, as the normal component of the interaction between headgroup dipoles is strong.

## 5. Conclusions

In this paper a novel mixed resolution interaction (MRI) method has been proposed that combines atomistic and coarse-grained (CG) descriptions of molecular interactions in condensed-phase systems. The method defines the mixed resolution interactions only along the pairwise interactions between particles and not at space-fixed boundaries in the simulation cell like other mixed resolution approaches. The total forces acting on individual atoms in the MRI framework include both accurate atomistic contributions arising from nearby molecules and from "unfolded" forces derived from the molecular MS-CG forces between pairs of distant molecules. We have also introduced a transition region between the atomistic and CG zone, which smoothes the forces. The MS-CG force fields, described fully elsewhere, are based on a preselected partitioning of the system into CG units and have been previously shown to properly incorporate many-body effects.

Implementation of the MRI algorithm involves a CG group-based cutoff treatment of interactions, similar to the neutral group implementation used in cutoff electrostatics. A group-based cutoff formalism is computationally more demanding than a simple cutoff treatment. However, as atomistic interactions outside the relative small atomistic zone are still derived from the MS-CG forces through the computationally cheap unfolding scheme mentioned above, the MRI simulations will lead to significantly improved computational performance, especially for very large systems.

The MRI methodology was applied to both liquid water and methanol, both of which are important solvents. With a sufficiently large atomistic zone, it was possible to achieve a description of water properties superior to that provided by simple cutoff methods. The MRI description of liquid methanol resulted in good liquid structure, but thermodynamic and diffusion properties were reproduced less accurately. This deficiency may be attributable to the bulkier methanol molecules or to the inaccuracy of the one-site CG methanol model. For both water and methanol, the MRI modeling is likely to be improved, if the underlying CG model is made to be more thermodynamically and structurally accurate.

The transferability of the MRI water potentials to inhomogeneous environments, however, appears to be more challenging. For an aqueous ion solution, the MRI treatment of water−water interactions led to unnatural stability in the ionic complexes. This artifact can be attributed to the fact that water−water interactions are effectively different in the environment perturbed by the ionic solutes. To improve the dynamics of the ionic solute association, we added an atomistic description of the interactions among water molecules within a preselected radius of the solutes. This approach, denoted the solute/bulk (slt/blk) scheme, proved to be effective and is expected to perform even better for systems with a larger water subsystem (for example, solvated proteins). We also tested a computationally less-expensive algorithm for correcting the interactions in perturbed water, where atomistic potentials are used only if both water molecules belong to solute neighborhoods. However, this approach proved insufficient to correctly simulate solute association dynamics. Finally, the MRI description of water−water interactions with the solute/bulk enhancement was applied with reasonable success to simulate a phospholipid bilayer, where the electrostatic interactions within the bilayer subsystem were evaluated using accurate Ewald summation.

The advance described in this paper on mixed resolution modeling must be considered preliminary. Clearly, the MRI method does not always perform well for heterogeneous systems, though likely it is better than purely coarse-grained modeling of similar systems. Moreover, the fully efficient and optimized computational implementation of the MRI method remains for future work. Nevertheless, the overarching concept of the MRI approach is a novel one that changes resolution of the interactions as a function of only the interparticle separation and not in terms of regions fixed in space.

**Acknowledgment.** The research was supported by the National Science Foundation (CHE-0719522) and the Office of Naval Research.

**Supporting Information Available:** Properties of bulk TIP3P water at ambient conditions, as determined from MRI/GC simulations with dipole (first line of data for each property) and Coulomb (second line) unfolding schemes for various atomistic zone sizes $R_{atm}$, and an Ion-water structure for a 12 ion system in 1 500 water molecules. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Noid, W. G.; Ayton, G. S.; Izvekov, S.; Voth, G. A. The Multiscale Coarse-Graining Method: A Systematic Approach to Coarse Graining. In *Coarse-graining of condensed-phase and biomolecular systems*; Voth, G. A., Ed.; CRC Press/Taylor and Francis Group: Boca Raton, FL, 2009; pp 21.

(2) Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Klein, M. L. *J. Phys. Chem. B* **2001**, *105*, 4464.

(3) Marrink, S. J.; Mark, A. E. *J. Am. Chem. Soc.* **2003**, *125*, 15233.

(4) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750.

(5) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144.

(6) Lyubartsev, A. P.; Laaksonen, A. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1995**, *52*, 3730.

(7) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 2469.

(8) Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2005**, *123*, 134105.

(9) Izvekov, S.; Voth, G. A. *J. Chem. Theory Comput.* **2006**, *2*, 637.

(10) Ayton, G. S.; Noid, W. G.; Voth, G. A. *Mat. Res. Bull.* **2007**, *32*, 929.

(11) Zhou, J.; Thorpe, I. F.; Izvekov, S.; Voth, G. A. *Biophys. J.* **2007**, *92*, 4289.

(12) Liu, P.; Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 11566.

(13) Noid, W. G.; Chu, J. W.; Ayton, G. S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 4116.

(14) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. *J. Chem. Phys.* **2008**, *128*, 244114.

(15) Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. *J. Chem. Phys.* **2008**, *128*, 244115.

(16) Praprotnik, M.; Site, L. D.; Kremer, K. *J. Chem. Phys.* **2005**, *123*, 224106.

(17) Christen, M.; van Gunsteren, W. F. *J. Chem. Phys.* **2006**, *124*, 154106.

(18) Praprotnik, M.; Site, L. D.; Kremer, K. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2006**, *73*, 066701.

(19) Praprotnik, M.; Kremer, K.; Site, L. D. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2007**, *75*, 017701.

(20) Praprotnik, M.; Site, L. D.; Kremer, K. *J. Chem. Phys.* **2007**, *126*, 134902.

(21) Praprotnik, M.; Kremer, K.; Site, L. D. *J. Phys. A: Math. Theor.* **2007**, 40.

(22) Ensing, B.; Nielsen, S. O.; Moore, P. B.; Klein, M. L.; Parrinello, M. *J. Chem. Theory Comput.* **2007**, *3*, 1100.

(23) Delle Site, L. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2007**, *76*, 047701.

(24) Errington, J. R.; Debenedetti, P. G. *Nature* **2001**, *409*, 318.

(25) Izvekov, S.; Swanson, J. M. J.; Voth, G. A. *J. Phys. Chem. B* **2008**, *112*, 4711.

(26) Matysiak, S.; Clementi, C.; Praprotnik, M.; Kremer, K.; Site, L. D. *J. Chem. Phys.* **2008**, *128*, 24503.

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

(28) Hummer, G.; Pratt, L. R.; Garcia, A. E. *J. Phys. Chem.* **1996**, *100*, 1206.

(29) Smondyrev, A. M.; Berkowitz, M. L. *J. Comput. Chem.* **1999**, *20*, 531.

(30) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.

(31) Ayton, G.; Smondyrev, A. M.; Bardenhagen, S. G.; McMurtry, P.; Voth, G. A. *Biophys. J.* **2002**, *83*, 1026.

(32) Wohlert, J.; Edholm, O. *Biophys. J.* **2004**, *87*, 2433.

(33) Lindahl, E.; Edholm, O. *J. Chem. Phys.* **2000**, *113*, 3882.

CT900414P

JCTC Journal of Chemical Theory and Computation

# Electrostatic Interactions in Dissipative Particle Dynamics: Toward a Mesoscale Modeling of the Polyelectrolyte Brushes

Cyrille Ibergay,[†] Patrice Malfreyt,[*,†] and Dominic J. Tildesley[‡]

*Laboratoire de Thermodynamique et Interactions Moléculaires, FRE CNRS 3099,*
*Université Blaise Pascal, 63177 Aubière Cedex, France, and Unilever Research,*
*Port Sunlight, Bebington, Wirral CH63 3JW, U.K.*

Received June 9, 2009

**Abstract:** We report mesoscopic simulations of bulk electrolytes and polyelectrolyte brushes using the dissipative particle dynamics (DPD) method. The calculation of the electrostatic interactions is carried out using both the Ewald summation method and the particle−particle particle-mesh technique with charges distributed over the particles. The local components of the pressure tensor are calculated using the Irving and Kirkwood, and the method of planes and mechanical equilibrium is demonstrated. The profiles of the normal component of the pressure tensor are shown to be similar for both the Ewald and particle−particle particle-mesh methods for a single polyelectrolyte brush. We show that the PPPM method with the MOP technique is the appropriate choice for simulations of this type. The mesoscale modeling of a strongly stretched polylectrolyte brush formed by strong charged polymer chains at a high grafting density shows that the polyelectrolyte follows the nonlinear osmotic regime, as expected from the calculation of the Gouy−Chapman length and the dimensionless Manning ratio.

## 1. Introduction

Molecular simulations of polymers and surfactant solutions demands modeling on a hierarchy of length and time scales spanning several orders of magnitude. For example, in a polymer brush, the size of the atomic constituents is on the order of 1 Å, and the fastest motions are at times on the order of $10^{-14}$ s, whereas collective relaxation in the system can occur on a scale of micrometers and at times that exceed 1 ms. Molecular simulations of brushes require that the equations of motion be solved with a time scale of a femtosecond and a length scale of angströms, and this might not be the most efficient approach for studying mesoscale phenomena such as the friction between brushes.

The development of simulation techniques capable of accessing mesoscopic length and time scales is an area of active research.[1−4] One approach is the dissipative particle dynamics (DPD) method, initially proposed by Hoogerbrugge and Koelman.[1] This method consists of reducing the complexity of the atomistic description of the system through the use of a coarse-grained model. In this method, a number of atoms are combined into particles that interact with each other through soft conservative and pairwise dissipative and random forces. The dissipative and random forces are related through the fluctuation−dissipation theorem, leading to a local conservation of the momentum, which is required for a correct description of hydrodynamics.[3] DPD has been successfully applied to investigate a variety of soft-matter problems such as the microphase separation of block copolymers,[5,6] polymer surfactants in solution,[7] and the structure and rheology of biological membranes.[8]

We have used the DPD approach to study the interaction between two solid surfaces coated with grafted polymer chains. At a relatively high surface coverage under good solvent conditions, the polymer chains are strongly stretched in the direction perpendicular to the surface; this leads to a structure called a polymer brush. End-grafted polymer chains give rise to a wide range of important industrial applications

[*] Corresponding author e-mail: Patrice.MALFREYT@univ-bpclermont.fr.

[†] TIM, FRE CNRS 3099.

[‡] Unilever Research.

10.1021/ct900296s CCC: $40.75 © 2009 American Chemical Society
Published on Web 10/29/2009

in the stabilization of colloidal suspensions, adhesion, lubrication, friction, and wear. We have also adapted the standard DPD method[9] to model the friction between two polymer brushes as a function of the quality of the solvent[10] and the separation between the surfaces.[11,12]

In this article, we aim to simulate charged polymers grafted to surfaces. Recent experiments have shown that polyelectrolyte brushes are better lubricants than neutral brushes.[13] However, the inclusion of the long-range electrostatic interactions in DPD is required to model such effects accurately. Electrostatic interactions were recently introduced into DPD by Groot[14] and by Alejandre and co-workers.[15] Both proposed replacing the point charge at the center of the DPD particle with a charge distribution smeared across the particle. Groot proposed a method in which the electrostatic field is calculated locally using a grid technique, whereas Alejandre et al. used a modification of the standard Ewald sum method.[16] These two methods have been applied to study a bulk electrolyte and polyelectrolyte−surfactant solutions.[14,15] Good agreement was found between the two methods for the radial distribution functions of charged particles in bulk electrolytes and polyelectrolyte−surfactant solutions. In this article, we compare these methods for systems with reduced periodicity.

In a brush system, the calculation of the friction coefficient requires the calculation of the normal and tangential components of the pressure tensor. When the system is nonperiodic in one dimension, it is important to calculate profiles of the pressure tensor along this axis. This establishes the mechanical equilibrium in the system and allows an accurate average value of the friction to be obtained from the profile.[10,12,17] In this work, we consider the most efficient way of calculating the profiles of the pressure tensor when electrostatic interactions are included in the DPD approach.

Different methods can be used to calculate the local pressure components along a specific direction. The potential term in the pressure tensor introduces arbitrariness because there is a choice of the contour joining the two particles. Several choices have been developed to calculate the potential component of the pressure tensor, including those of Irving and Kirkwood (IK)[18] and Harasima.[19] The IK definition is applicable only for pairwise-additive interactions. Contributions such as the reciprocal part of the electrostatic interactions treated with the Ewald sum are not pairwise-additive, so the definition of Harasima can be used.[20,21] An alternative approach consists of using the method of planes (MOP) formalism,[22] which avoids the heuristic notions of the force across a unit area. This method can be applied to both pairwise-additive and non-pairwise-additive interactions. In this work, we compare these approaches.

Additionally, in some important cases, the modeling of a surface requires a system that is nonperiodic in the third dimension. The conventional Ewald summation method can be applied by elongating the primary cell in the direction of the surface by adding a sufficiently large vacuum between the periodic images. The aim is to dampen out the interslab interactions. This methodology is often referred to as the supercell approximation, and it has been applied successfully.[23−25] To remove the forces due to the net dipole of the periodically repeating slabs, a correction dipole term must be added.[23,24] This methodology is known as the EW3DC method.

There are some quasi-periodic Ewald methods, such those due to Hautman and Klein[26] and Lekner,[27] that can be applied in these geometries. These often result in series expansions of the electrostatic interactions where the convergence depends on the particular distribution of particles. In addition, these methods can be extremely expensive in terms of computational resources. However, the MMM technique[28] adapted by Strebel and Sperb[29] for slab geometries[30,31] maintains a reasonable computational cost with an $O(N^{5/3} \log N)$ behavior, where $N$ is the number of charged particles. Therefore, we do not consider these methods further in the work. Instead, we employ the Groot particle−particle particle-mesh (PPPM) method in the case of a slab geometry, as well as the EW3DC approach.

To explore the different methods for accounting for the electrostatic interactions in the calculation of the local pressure components, we focus on three model systems: a three-dimensional bulk electrolyte, a bulk electrolyte embedded between two parallel surfaces, and a system of charged polyelectrolyte brushes. We then use these systems to validate the methodology, and we complete this work with a preliminary study of grafted polyelectrolytes, which represent an interesting topic with many unresolved problems for both experiment and theory. The system studied presently is a model system formed with a high surface coverage and a relatively strong charge fraction.

In section 2, we present the conventional forces used in the DPD model. In section 3, we describe the two techniques used for the calculation of the electrostatic forces. The different definitions for the calculation of the local components of the pressure tensor are described in section 4. The results for the different model systems are given in section 5. Finally, we conclude in section 6 by providing a brief summary of our main results.

## 2. Dissipative Particle Dynamics (DPD) Model

**2.1. Standard DPD Forces.** In the DPD approach, solvent particles are coarse-grained into soft beads that interact with the pairwise-additive force $\mathbf{f}_i$ defined as the sum of three contributions

$$\mathbf{f}_i = \sum_{j \neq i} (\mathbf{f}_{ij}^{\mathrm{C}} + \mathbf{f}_{ij}^{\mathrm{R}} + \mathbf{f}_{ij}^{\mathrm{D}}) \tag{1}$$

where $\mathbf{f}_{ij}^{\mathrm{C}}$, $\mathbf{f}_{ij}^{\mathrm{R}}$, and $\mathbf{f}_{ij}^{\mathrm{D}}$ are the conservative, random, and dissipative forces, respectively. The conservative repulsive force, $\mathbf{f}_{ij}^{\mathrm{C}}$, derives from a soft interaction potential and is expressed as

$$\mathbf{f}_{ij}^{\mathrm{C}} = \begin{cases} a_{ij}\omega^{\mathrm{C}}(r_{ij})\hat{\mathbf{r}}_{ij} & (r_{ij} < r_{\mathrm{c}}) \\ 0 & (r_{ij} \geq r_{\mathrm{c}}) \end{cases} \tag{2}$$

where $a_{ij}$ is the maximum repulsion parameter between particles $i$ and $j$, $r_{ij}$ is the relative displacement of particles $i$ and $j$, and $\hat{\mathbf{r}}_{ij}$ is the corresponding unit vector. The weight function $\omega^{\mathrm{C}}(r_{ij})$ is equal to $1 - r_{ij}/r_{\mathrm{c}}$ for $r_{ij} \leq r_{\mathrm{c}}$ and vanishes for $r_{ij} \geq r_c$. The dissipative and random forces are given by

$$\mathbf{f}_{ij}^{\mathrm{D}} = -\gamma\omega^{\mathrm{D}}(r_{ij})(\hat{\mathbf{r}}_{ij}\cdot\mathbf{v}_{ij})\hat{\mathbf{r}}_{ij} \qquad (3)$$

$$\mathbf{f}_{ij}^{\mathrm{R}} = \sigma\omega^{\mathrm{R}}(r_{ij})\theta_{ij}\frac{1}{\sqrt{\delta t}}\hat{\mathbf{r}}_{ij} \qquad (4)$$

where $\delta t$ is the time step. $\mathbf{v}_{ij} = \mathbf{v}_i - \mathbf{v}_j$ is the relative velocity, $\sigma$ is the amplitude of the noise, $\theta_{ij}$ is a random Gaussian number with zero mean and unit variance. $\gamma$ and $\sigma$ are the dissipation strength and noise strength, respectively. The terms $\omega^{\mathrm{D}}(r_{ij})$ and $\omega^{\mathrm{R}}(r_{ij})$ are dimensionless weighting functions. Espanol and Warren[3] showed that the system will sample the canonical ensemble and obey the fluctuation−dissipation theorem if the following conditions are satisfied:

$$\gamma = \frac{\sigma^2}{2k_{\mathrm{B}}T} \quad \text{and} \quad \omega^{\mathrm{D}}(r_{ij}) = [\omega^{\mathrm{R}}(r_{ij})]^2 \qquad (5)$$

where $k_{\mathrm{B}}$ is the Boltzmann constant and $T$ is the temperature. The weighting function $\omega^{\mathrm{R}}(r_{ij})$ is chosen to be similar to $\omega^{\mathrm{C}}(r_{ij})$.

The equations of motion are integrated using a modified version of the velocity-Verlet algorithm.[14] The force is updated once per iteration, and because the force depends on the velocities, the velocity in the next time step has to be estimated by a predictor algorithm. The velocity is then corrected in the last step. The reduced time step $\delta t$ was taken as 0.02 for all of the simulations, except for those involving polyelectrolyte brushes, for which it was equal to 0.06.

When fully flexible polymer chains are considered in solvent, the integrity of the polymer chain is ensured by an additional spring force between neighboring beads given by

$$\mathbf{f}_{ij}^{\mathrm{S}} = -k_{\mathrm{s}}(\mathbf{r}_{ij} - r_0)\hat{\mathbf{r}}_{ij} \qquad (6)$$

where the equilibrium bond distance $r_0$ is 0 and the spring constant $k_{\mathrm{s}} = 4.0$. This pairwise force is then added to the sum of the DPD conservative force in eq 1.

## 3. Electrostatic Interactions

In the following sections, we present two techniques to take into account the electrostatic interactions at a mesoscopic level. The first one consists of solving the electrostatic field on a grid. This is referred to as the particle−particle particle-mesh (PPPM) algorithm,[32−34] although, in the original version of the PPPM algorithm, the far field was solved using a fast Fourier transform.[14,32,35] The second consists of adapting the standard Ewald method to DPD particles.

**3.1. Particle−Particle Particle-Mesh (PPPM) Method.** Electrostatic interactions were incorporated into the DPD model by Groot.[14] The first step consists of finding a model for the density distribution adapted to a charged bead. The use of a soft potential in DPD allows for the overlap between DPD beads. When charged DPD beads are modeled, this can lead to the formation of artificial ion pairs and cause the divergence of the electrostatic potential. To avoid this problem, Groot chose to spread out the charges using the distribution

$$f(r) = \frac{3}{\pi r_{\mathrm{e}}^{3}}(1 - r/r_{\mathrm{e}}) \quad \text{for} \quad r < r_{\mathrm{e}} \qquad (7)$$

where $r_{\mathrm{e}}$ is the electrostatic smearing radius and $f(r) = 0$ when $r$ is greater than $r_{\mathrm{e}}$. The expression for the potential between two of these charge clouds is given in Appendix A, and the representation of the potential and its corresponding force with respect to the distance is shown in Figure 1. The electrostatic field is then solved on a lattice according to the method of Beckers et al.[32] The charges are assigned to the lattice nodes within the cell, and the long-range part of the interaction potential is calculated by solving the Poisson equation on the grid. Details of the charge assignment can be found in Groot's[14] original article. Whereas, in the original PPPM method, the far field was solved using a Fourier transform, the method developed by Groot used real-space successive overdamped relaxations. This makes the Groot method close to the multigrid method of Sagui and Darden.[34] However, for convenience, this method will be referred to as the PPPM method in the present work.

The electrostatic force $\mathbf{f}_i^{\mathrm{E}}$ on a charged bead $i$ is calculated from

$$\mathbf{f}_i^{\mathrm{E}}(\mathbf{r}_i) = -q_i\sum_j f_j(\mathbf{r}_i)\nabla\psi(\mathbf{r}_j) \qquad (8)$$

where $\mathbf{r}_i$ is the position of the charged bead $i$ and $q_i$ is the number of unit charges on bead $i$. $f_j(\mathbf{r}_i)$ is defined as

$$f_j(\mathbf{r}_i) = \frac{1 - |\mathbf{r}_j - \mathbf{r}_i|/r_{\mathrm{e}}}{\sum_n (1 - |\mathbf{r}_n - \mathbf{r}_i|/r_{\mathrm{e}})} \qquad (9)$$

where $\mathbf{r}_j$ is the position of the node $j$ and the sum over $n$ runs over all nodes within a distance $r_{\mathrm{e}}$ from $\mathbf{r}_i$. This function means that a charge proportional to $f(r)$ in eq 7 is assigned to each node $i$ and normalized such that the sum of all of the charges within a distance $r_{\mathrm{e}}$ equals the charge on the bead $i$. $\psi(\mathbf{r}_j)$ is the local electrostatic field at lattice node $j$.

The field $\psi(\mathbf{r}_j)$ is calculated from the Poisson equation expressed in reduced DPD units as

$$\nabla^*\cdot[p(r)\nabla^*\psi] = -\Gamma\rho^* \qquad (10)$$

where $\rho^*$ is the concentration of cations minus the concentration of anions per $r_c^3$, $\nabla^*$ is the gradient in reduced units, and $p(r)$ is the local polarizability relative to that of pure water.

The total momentum of the simulation cell is conserved by removing a possible residual force for each charge. This residual force is on the order of $5 \times 10^{-5}$ in reduced units and is expressed as $\sum_i^N q_i\mathbf{f}_i^{\mathrm{E}}/N$. The total force $\mathbf{f}_i$ of eq 1 is then modified by adding the $\mathbf{f}_i^{\mathrm{E}}$ contribution and the residual force.

**3.2. Ewald Summation (EW3DC) Method.** The method recently proposed by González-Melchor et al.[15] consists of combining the Ewald technique[16] and a charge distribution for particles. In the case of an electroneutral system formed by $N$ particles, with each particle $i$ carrying a point charge $q_i$ at position $r_i$ in a volume $V = L_xL_yL_z$, the long-

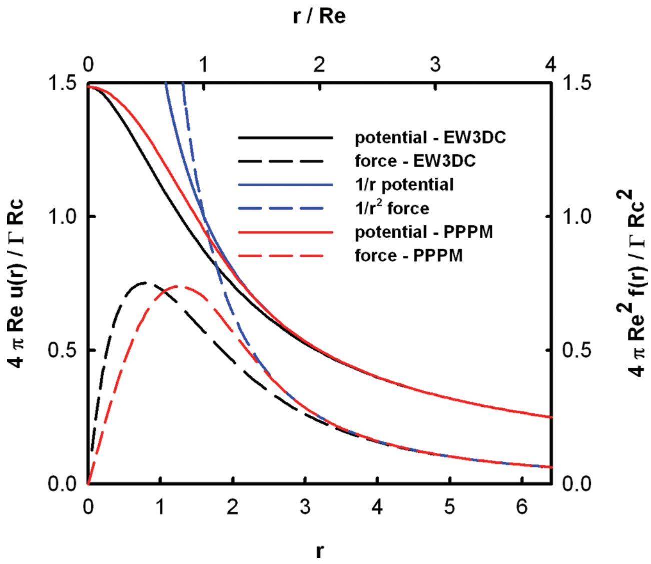


***Figure 1.*** Electrostatic potential and force calculated by the EW3DC and PPPM methods. For comparison, we include the Coulombic potential and force, which diverges at $r = 0$. The $x$ axis is expressed in standard DPD units ($r/r_c$), whereas the top axis gives the distance $r/r_e$. The potential and force expressions are plotted for two equal-sign charge distributions.

range electrostatic interactions are decomposed into contributions in real space and in reciprocal space

$$U(\mathbf{r}^N) = \frac{1}{4\pi\varepsilon_0\varepsilon_r}\left[\sum_i \sum_{j>i} q_i q_j \frac{\mathrm{erfc}(\alpha r_{ij})}{r_{ij}} + \frac{2\pi}{V}\sum_{k\neq 0}^{\infty} Q(h)\, S(\mathbf{h})\, S(-\mathbf{h}) - \frac{\alpha}{\sqrt{\pi}}\sum_i^N q_i^2\right] \quad (11)$$

where erfc($x$) is the complementary error function. α is chosen so that only pair interactions in the central cell need to be considered in evaluating the first term of eq 11. The functions $Q(h)$ and $S(\mathbf{h})$ are defined by the equations

$$Q(h) = \exp(-h^2/4\alpha^2)/h^2 \quad (12)$$

$$S(\mathbf{h}) = \sum_{i=1}^{N} q_i \exp(i\mathbf{h}\cdot\mathbf{r}_i) \quad (13)$$

where the components of the reciprocal vector **h** are defined as $2\pi(l/L_x, m/L_y, n/L_z)$ where $l$, $m$, and $n$ take values of 0, ±1, ±2, ..., ±∞.

To remove the divergency of the Coulombic potential at $r = 0$, Alejandre and co-workers[15] considered a Slater-type charge distribution on DPD particles of the form

$$f(r) = \frac{q}{\pi\lambda^3}\exp(-2r/\lambda) \quad (14)$$

where $\lambda$ is the decay length of the charge. The distribution is normalized to $q$.

The magnitude of the reduced force between two charge distribution is then given by the sum of a pairwise-additive contribution $\mathbf{f}_{ij}^{E,R}$ coming from the real-space term and a non-pairwise-additive contribution $\mathbf{f}_i^{E,K}$ from the reciprocal-space term. These two contributions are given by the expressions

$$\mathbf{f}_{ij}^{E,R} = \frac{\Gamma}{4\pi} q_i q_j \left[\frac{2}{\sqrt{\pi}}\exp(-\alpha^2 r_{ij}^2) + \mathrm{erfc}(\alpha r_{ij})\right] \times \{[1 - \exp(-2\beta r_{ij})][1 + 2\beta r_{ij}(1 + \beta r_{ij})]\}\frac{\mathbf{r}_{ij}}{r_{ij}^3} \quad (15)$$

$$\mathbf{f}_i^{E,K} = -\frac{\Gamma}{4\pi} q_i \left\{\frac{2\pi}{V}\sum_{\mathbf{h}\neq 0} Q(h)\mathbf{h} \times \mathrm{Im}[\exp(-i\mathbf{h}\cdot\mathbf{r}_i)\, S(\mathbf{h})]\right\} \quad (16)$$

where Im denotes the imaginary part of the complex variable.

To remove the net dipole moment of the simulation cell, a $z$-component force is added for each bead

$$\mathbf{f}_{i,z} = -\frac{\Gamma}{V} M_z \quad (17)$$

where $M_z$ is the net dipole moment of the simulation cell given by $\sum_i q_i \mathbf{r}_i$ and $V$ is the volume expressed in reduced units. This contribution is the correction term from Yeh and Berkowitz,[23] which results from the plane-wise summation method proposed by Smith.[36]

Within the EW3DC method, the force acting on the $i$th particle becomes

$$\mathbf{f}_i = \mathbf{f}_i^{E,K} + \mathbf{f}_{i,z} + \sum_{j\neq i}(\mathbf{f}_{ij}^{C} + \mathbf{f}_{ij}^{R} + \mathbf{f}_{ij}^{D} + \mathbf{f}_{ij}^{E,R}) \quad (18)$$

The pairwise $\mathbf{f}_{ij}^{E,R}$ force is then added to sum of the conservative, dissipative, and random pairwise forces, whereas $\mathbf{f}_i^{E,K}$ and $\mathbf{f}_{i,z}$, which are not pairwise-additive, are added to the force $\mathbf{f}_i$ acting on particle $i$. The real parts of the force and energy equations are shown in Figure 1. In this work, we compare the results from these two different techniques, and it is worth pointing out that the potentials and forces are slightly different, as can be seen in Figure 1. In Groot's method, an approximate potential is derived from the distribution of eq 7. In Alejandre et al.'s method, the potential is exactly defined from eq 14. Following Alejandre et al., $\lambda$, the decay length of the charge, is adjusted to bring the potential into the closest possible agreement.

## 4. Calculation of the Pressure Tensor

**4.1. Irving and Kirkwood (IK) Definition.** The method of Irving and Kirkwood[18] (IK) is based on the notion of the force across a unit area. The pressure tensor is then written as a sum of a kinetic term and a potential term resulting from the intermolecular forces. Whereas the first term is well-defined, the potential term is subjected to arbitrariness because there is no unique way to determine which intermolecular forces contribute to the stress across d$A$. There are many ways of choosing the contour joining two interacting particles. Irving and Kirkwood[18] chose as a contour the straight line between the two particles. Other choices are possible and result from the lack of uniqueness in the definition of the microscopic stress tensor. The components of the pressure tensor[37−39] in the Irving and Kirkwood definition are expressed by

$$p_{\alpha\beta}(z) = \frac{1}{V}\left\langle \sum_{i=1}^{N} H(z_i) m_i \mathbf{v}_{i,\alpha}\mathbf{v}_{i,\beta}\right\rangle + \frac{1}{A}\left\langle \sum_{i=1}^{N-1}\sum_{j>i}^{N} (\mathbf{r}_{ij})_{\alpha} (\mathbf{f}_{ij})_{\beta} \frac{1}{|z_{ij}|}\; \theta\left(\frac{z-z_i}{z_{ij}}\right)\theta\left(\frac{z_j - z}{z_{ij}}\right)\right\rangle \quad (19)$$

The first term on the right-hand side of this equation represents the kinetic part, and the second term is the configurational pressure calculated from the conservative potentials. $H(z_i)$ is a top-hat function. $\alpha$ and $\beta$ represent the $x$, $y$, and $z$ directions. $\theta(x)$ is the unit step function defined by $\theta(x) = 0$ when $x < 0$ and $\theta(x) = 1$ when $x \geq 0$. $A$ is the surface area normal to the $z$ axis. The distance $z_{ij}$ between two particles is divided into $N_s$ slabs of thickness $\delta z$. Following Irving and Kirkwood, particles $i$ and $j$ give a local contribution to the pressure tensor in a given slab if the line joining these crosses, starts in, or finishes in the slab. Each slab has $1/N_s$ of the total contribution from the $i-j$ interaction. The normal component $p_N(z_k)$ is equal to $p_{zz}(z_k)$. $\mathbf{f}_{ij}$ in eq 19 is the pairwise force between particles $i$ and $j$ and sums the conservative, dissipative, random, and bead−spring forces, as well as the real-space contribution of the electrostatic forces. The reciprocal contribution to the electrostatic force is taken into account using the Harasima definition[19] of the pressure. The expressions for the local contributions to the pressure tensor are given in Appendix B for completeness.

**4.2. Method of Planes (MOP) Definition.** The method of planes[22] (MOP), introduced by Todd, Evans, and Daivis, is designed to calculate average cross-sectional pressures. The total pressure sums the kinetic and potential contributions as

$$p_{\alpha z}(z) = \frac{1}{A}\sum_{i=1}^{N}\left\langle \frac{m_i \mathbf{v}_{\alpha,i}\,\mathrm{sgn}(v_{i,z})}{\delta t}\right\rangle + \frac{1}{2A}\left\langle \sum_{i=1}^{N}\mathbf{f}_{\alpha,i}\,\mathrm{sgn}(z_i - z)\right\rangle \quad (20)$$

where $\mathbf{v}_{\alpha,i}$ is the $\alpha$ component of the velocity of particle $i$ and $\mathbf{f}_{\alpha,i}$ is the $\alpha$ component of the total force on particle $i$. The kinetic part is due to the momentum of the molecules as they cross the area during $\Delta t$. If, between times $t$ and $t + \Delta t$, particle $i$ moves through planes, we use the sign of the $z$ component of the velocity to specify the direction of the crossing. This method allows for the use of the total force $\mathbf{f}_{\alpha,i}$ calculated either from the PPPM or EW3DC method.

When the total force $\mathbf{f}_{\alpha,i}$ can be decomposed into pairwise contributions $\mathbf{f}_{\alpha,ij}$, the second term of eq 20 can be written as

$$\frac{1}{A}\left\langle \sum_{i}^{N-1}\sum_{j=i+1}^{N} \mathbf{f}_{\alpha,ij}[\theta(z_i - z)\theta(z - z_j) - \theta(z_j - z)\theta(z - z_i)]\right\rangle \quad (21)$$

When the force cannot be decomposed into pairwise contributions (PPPM method) and the system is periodic in all three directions, the use of the MOP methodology for the calculation of the pressure components is not possible.

## 5. Results and Discussion

**5.1. Bulk Electrolyte.** As a first test to validate the calculation of the pressure tensor, we consider the simple electrolyte previously studied by Groot[14] and Alejandre and co-workers.[15] The system consisted of $N = 3000$ particles in a simulation cell of volume 10 × 10 × 10. The box contained 2804 neutral particles, 98 particles with a positive charge of $+e$ and 98 particles with a negative charge of $-e$. The total density number was $\rho = 3$. A typical configuration is shown in Figure 2a. Using the appropriate scaling given in Table C-1 of Appendix C, this system corresponds to a salt concentration of 0.6 M. The interaction parameters for the conservative, dissipative, and random forces were $a_{ij} = 25.0$, $\lambda = 4.5$, and $\sigma = 3.0$, respectively. The time step $\delta t$ was equal to 0.02. Each simulation for these systems consisted of an equilibration period of 100 000 steps, followed by an acquisition period of 300 000 steps. The length of the production phase is about 1 $\mu$s. The periodic boundary conditions were applied in all three directions. Because the calculation of the pressure cannot be performed with a non-pairwise-additive force (PPPM method) in a three dimensionally periodic system, we can study only the 3D bulk electrolyte with the EW3DC method. We performed these simulations as a reference for the works described in section 5.3, where the box simulations represent the homogeneous phase in the middle of the cell.

Figure 3a shows the kinetic term for the pressure calculated according to the IK definition along the $z$ direction in the bulk electrolyte. We observe that this profile is constant with a mean value of 3.11 ± 0.01. This corresponds to an average temperature of 0.99 ± 0.01, which agrees well with the input temperature of 1.0. This means that the incorporation of the electrostatic interactions allows for very accurate temperature control with the use of the velocity-Verlet algorithm and a time step of 0.02. Figure 3b represents the normal component of the configurational pressure tensor along the $z$ direction calculated according to the IK and MOP approaches. As expected from a system mechanical equilibrium, the profile is constant, and the average pressure of $p_{zz}$ calculated over the different $z$ positions is equal to 20.7 ± 0.1 for the two definitions. The profiles of the real and reciprocal contributions of the Ewald summation method shown in Figure 3c are identical within the statistical fluctuations for the IK and MOP methods. The electrostatic contributions of −0.013 DPD units are relatively small compared to the magnitude of the configurational pressure. The value of pressure is determined from the contributions between ions of the same charge (0.11) and between ions of opposed charges (−0.123). The electrostatic interactions are small because of the strong cancellation between different ion pairs.

**5.2. Electrolyte Embedded between Two Planar Surfaces.** We now simulate the electrolyte between two planar solid surfaces composed of three layers of DPD particles tethered by springs to lattice points in a regular array (see Figure 2b for a typical configuration). Each surface was composed of a 17 × 18 layer of wall particles. The cell dimensions and the features of the system are given in Table 1. The separation distance between the two walls was 11. This value was chosen to reproduce a local value of the density number in the middle of the box similar to that of the bulk electrolyte system described in section 4.2. The time step was fixed at 0.02, and the simulations consisted of an

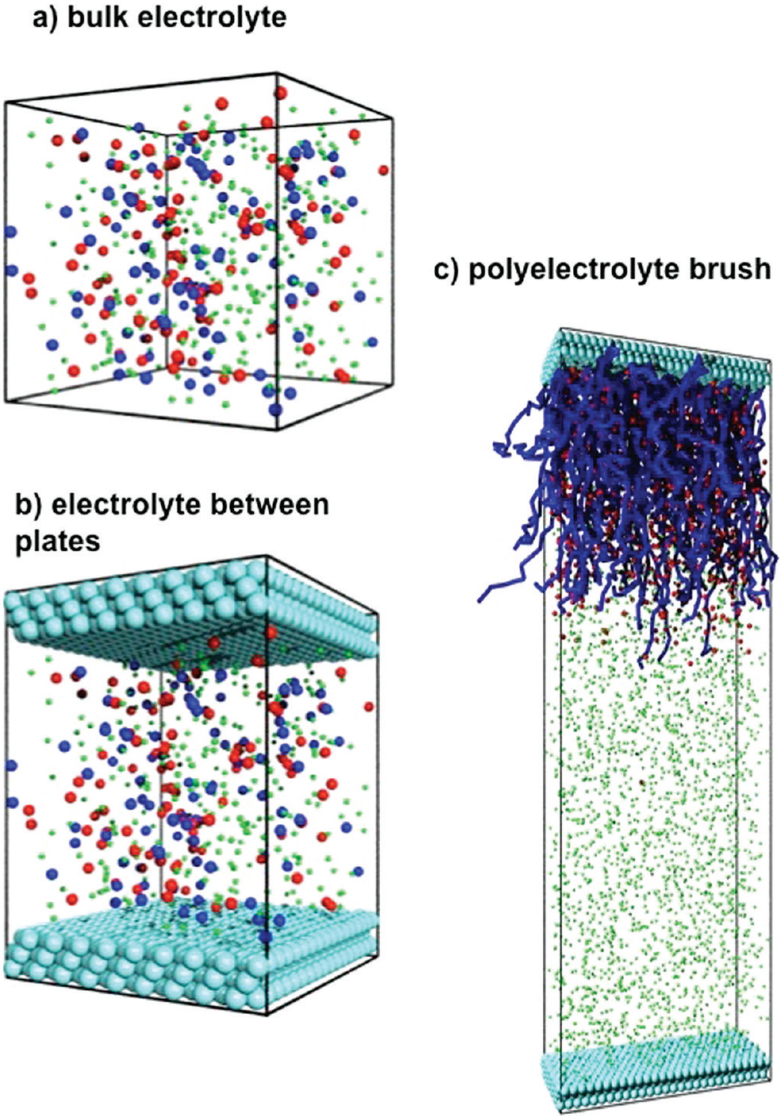


***Figure 2.*** Typical configurations of the three simulated systems. The solvent particles, the anions or polymer beads, and the counterions are represented in green, blue, and red, respectively.

equilibration period of 100 000 steps, followed by an acquisition period of 300 000 steps. Periodic boundary conditions were applied in all three directions. The simulation box dimension was elongated in the *z* direction by adding an empty space of at least twice the space of the fluid-occupied region. We added a correction term to remove the forces due to the net dipole of the periodically repeating slabs.[17,23,24] The reciprocal vector $\mathbf{h}_z^{\max}$ was increased (see Table C-1 in Appendix C) to allow for the elongation of the cell in the *z* direction.

When the electrostatic interactions are calculated with EW3DC, the average value of the normal component of the configurational pressure calculated in the middle of the pore from the profiles of Figure 4a is 20.7 ± 0.1 for the IK and MOP methodologies. This value agrees very well with that calculated in the bulk electrolyte. Figure 4b shows the profiles of the real and reciprocal parts of the normal pressure components calculated from EW3DC. We observe flat profiles in the middle of the pore in agreement with a homogeneous distribution of the ions along this direction. From these profiles, we deduce an average value of −0.012 ± 0.001 in the center of the box for the total electrostatic pressure. This value matches reasonably well with that calculated (−0.013) by EW3DC in the bulk electrolyte system. This result validates the use of a local definition for the pressure calculation within the supercell approximation. Additionally, we find that the sum of the kinetic, configurational, and electrostatic contributions to the pressure tensor lead to a completely flat profile (not shown here) as expected for a system at mechanical equilibrium.

In addition, the profiles of the configurational pressure in Figure 4c calculated from simulations using the PPPM method are also in line with those resulting from the EW3DC technique. The electrostatic pressure calculated according to the PPPM method is slightly different from that coming from the EW3DC method, with an average difference of 6%. The profiles of the electrostatic contributions to the normal pressure are shown in Figure 4d. This difference in the pressure values can be attributed to the fact that the electrostatic force used in EW3DC is slightly different from that used in the PPPM method, as shown in Figure 1. However, from these profiles, we can conclude that the

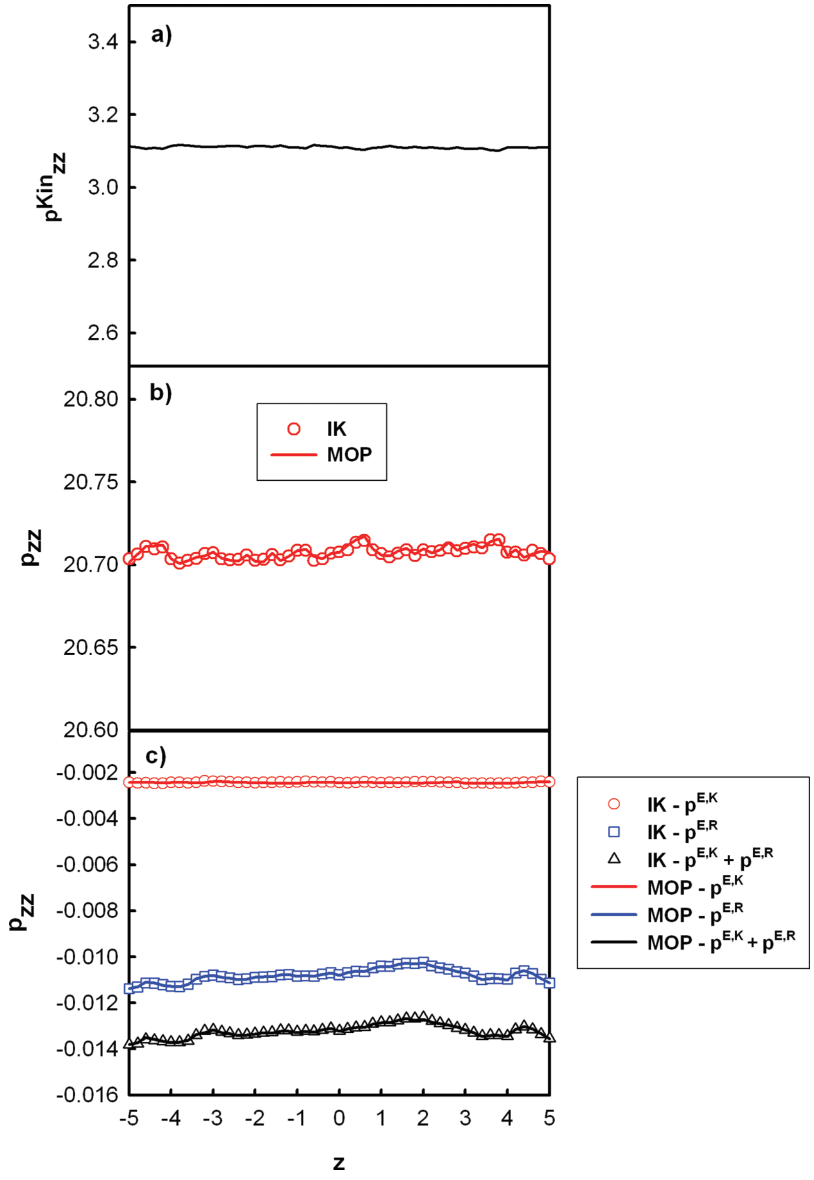


***Figure 3.*** (a) Kinetic and (b) configurational pressure components along the normal to the surface for the bulk electrolyte system. Contributions of the normal component of the pressure tensor from the real ($p^{E,K}$) and reciprocal ($p^{E,R}$) spaces.

***Table 1.*** Dimensions of the Box in Reduced Units and Number of Particles as a Function of the System

| system | $L_x/r_c$ | $L_y/r_c$ | $L_z/r_c$ | $h$ | $N_{solvent}$ | $N_+$[a] | $N_-$[a] |
|---|---|---|---|---|---|---|---|
| bulk electrolyte | 10 | 10 | 10 | | 2804 | 98 | 98 |
| bulk electrolyte between surfaces | 10.5 | 9.6 | 35 | 11 | 2804 | 98 | 98 |
| polymer brush[b] | | | | | | | |
| $f^c = 0$ | 16.7 | 6.4 | 152 | 50 | 13 863 | 0 | 0 |
| $f = 0.5$ | 16.7 | 6.4 | 152 | 50 | 12 783 | 1080 | 0 |
| $f = 1.0$ | 16.7 | 6.4 | 152 | 50 | 11 703 | 2160 | 0 |

[a] $N_+$ and $N_-$ represent the numbers of cations and anions, respectively. [b] In a polyelectrolyte brush, the number of counterions $N_+$ is equal to $fN_pN_b$. [c] Charge fraction.

***Table 2.*** Height of the Polymer Brush ($\langle z_m \rangle$), Height of the Counterion Layer ($\langle z_{ci} \rangle$), and Average Bond Length in the Polymer Chain ($\langle b \rangle$) Expressed in Reduced Units[a]

| fraction of charge ($f$) | method | $\langle z_m \rangle$ | $\langle z_{ci} \rangle$ | $\langle b \rangle$ |
|---|---|---|---|---|
| | Neutral Polymer Brush | | | |
| 0 | | 11.0 | | 0.97 |
| | Polyelectrolyte Brushes | | | |
| 0.5 | PPPM | 15.1 | 16.0 | 1.05 |
| 0.5 | EW3DC | 15.1 | 16.1 | 1.06 |
| 1.0 | PPPM | 18.9 | 18.9 | 1.14 |
| 1.0 | EW3DC | 19.0 | 19.1 | 1.15 |

[a] Heights calculated from the first moment of the density profiles.

mechanical properties of the electrolytes between the two plates are very similar when they are calculated by EW3DC or by PPPM.

Parts a and b of Figure 5 show the two-dimensional radial distribution functions between solvent−solvent and equal and unequal ion pairs from simulations performed using the EW3DC and PPPM methodologies. These two-dimensional distribution functions were calculated in the middle of the cell in a slab of width 0.3. For each case, we include for comparison the three-dimensional pair correlation functions calculated in the bulk electrolytes. First, we observe that the two- and three-dimensional radial distribution functions are well-matched, indicating that the structure in the middle of

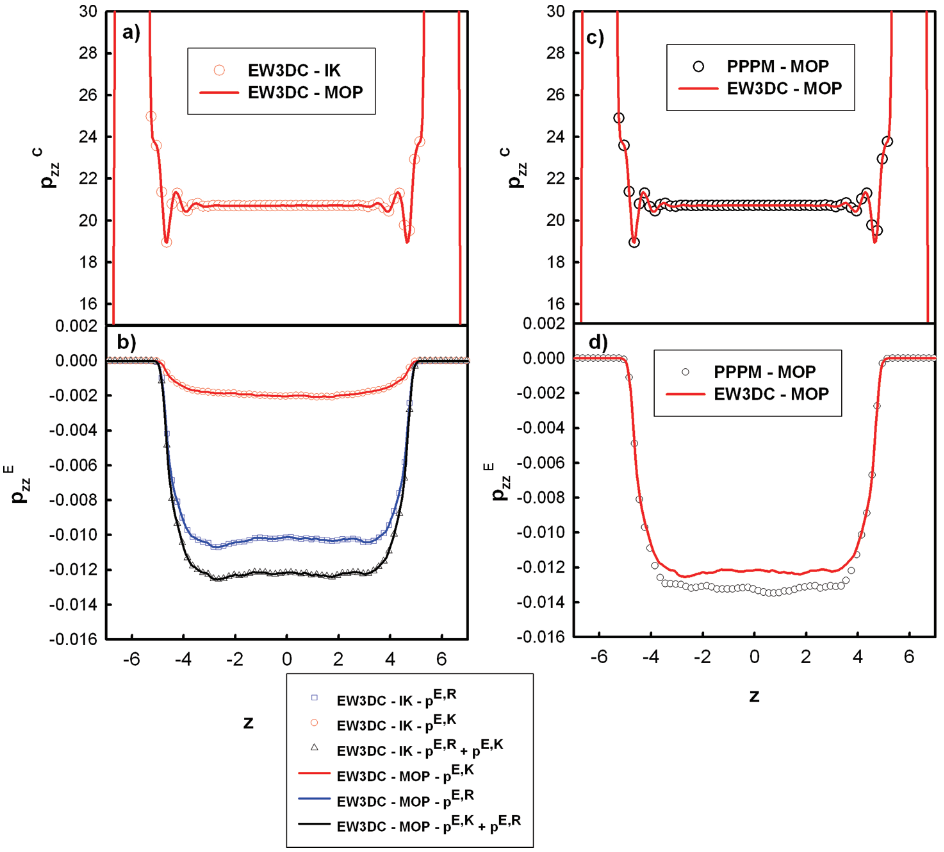


***Figure 4.*** (a) Normal component ($p_{zz}^{C}$) of the configurational part of the pressure calculated according to the IK and MOP definitions using the EW3DC method, (b) normal component of the contributions of the real ($p^{E,K}$) and reciprocal ($p^{E,R}$) spaces calculated with the IK and MOP definitions using the EW3Dc method, (c) normal component ($p_{zz}^{C}$) of the configurational part of the pressure calculated with MOP using both the PPPM and EW3DC methods, and (d) total normal component of the electrostatic contribution calculated using the MOP definition within the EW3DC and PPPM methods. The system consisted of an electrolyte between two surfaces.

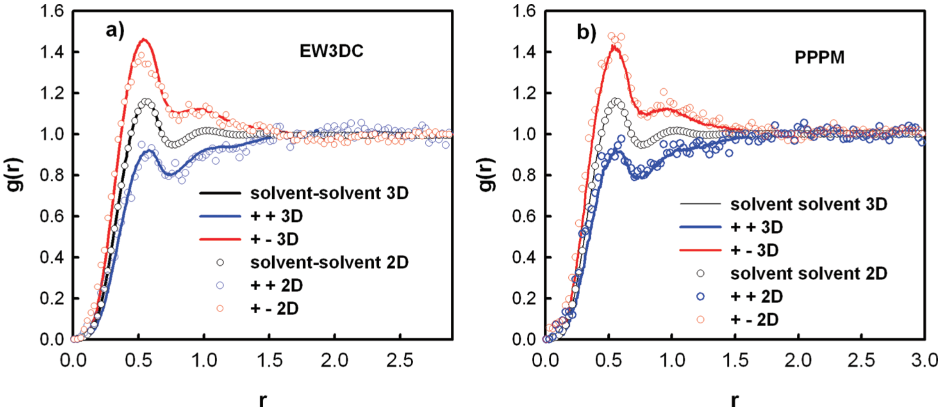


***Figure 5.*** Two-dimensional and three-dimensional radial distribution functions for different ion pairs. The two-dimensional distribution functions were calculated in the middle of the box in a slab with a width of of 0.3$r_c$, whereas the three-dimensional distribution functions were calculated in the bulk electrolyte.

the pore returns to that of the bulk. The two-dimensional curves are noisier because of the statistics. These distributions functions are also in line with those calculated in previous works by Alejandre and co-workers[15] and Groot.[14]

The calculation of the pressure components of the configurational and electrostatic parts in a system where a physical boundary prohibits passage of molecules in the third direction allows for the verification that the local definitions used for the pressure calculation (MOP and IK) are be relevant within the EW3DC and PPPM techniques. This was confirmed by comparing the structure and the mechanical properties of the central zone between the two surfaces with those of the three-dimensional bulk electrolytes.

This system allowed the relative speeds of the EW3DC and PPPM techniques to be compared. We simulated a confined polyelectrolyte of DPD particles between two walls. Initially, all of the particles were uncharged, and we performed a number of different simulations increasing the

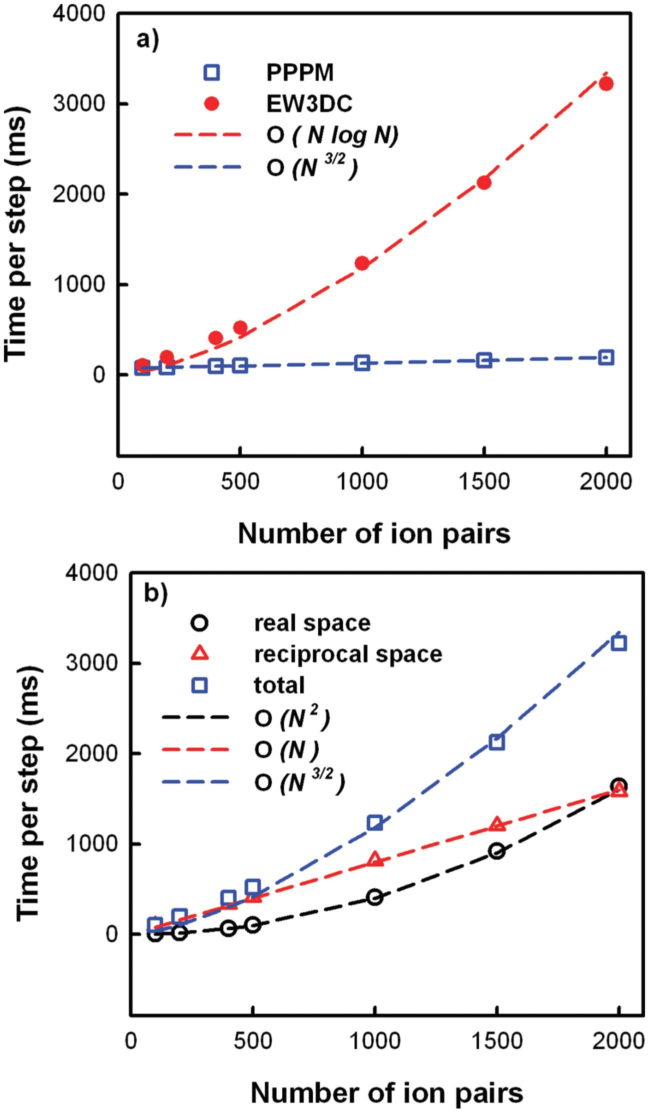


***Figure 6.*** Run times (in milliseconds) for one time step (a) in the calculation of the electrostatic force and (b) in the calculation of the real- and reciprocal-space terms of the electrostatic force. The average time was calculated from a DPD simulation carried out over 10 000 steps. The total number of particles was fixed at 16 000. The number of ion pairs increased from 0 to 2000, and the number of solvent particles decreased from 16 000 to 12 000. The system was an electrolyte between two plates.

number $N$ of ion pairs to 2000 while keeping the total number of particles fixed. Figure 6a shows the run time for one time step of a DPD simulation as a function of the number of ion pairs. The execution time for one time step includes only the calculation of the electrostatic force. This curve shows that the PPPM method becomes more efficient than EW3DC when the number of ion pairs is greater than 100. We verified that the EW3DC and PPPM methods scale as $O(N^{3/2})$[40] and $O(N \log N)$,[40] respectively. For the range of the number of ion pairs investigated here, the PPPM method performed faster than EW3DC for the study of polyelectrolytes with the DPD method. This is in line with Groot's observation[14] that the time used to distribute the charge, solve the field equation, and calculate the electrostatic forces is small compared to other elements of a time step. In Figure 6b, we show the times for the calculation of the reciprocal- and real-space terms of the electrostatic force in the EW3DC method. As expected from previous works,[40,41] the computation times for the calculation of the reciprocal- and real-space terms grow as $N^2$ and $N$, respectively. For the set of parameters (see Tables 1 and C-1) and numbers of ion pairs smaller than 1000, the computational cost of the calculation of the electrostatic forces with EW3DC comes from the calculation of the reciprocal-space term.

**5.3. Polyelectrolyte Brushes.** *5.3.1. Mechanical and Structural Properties.* Now that we have verified the consistency of the calculation of the local pressure within the supercell approximation in the EW3DC and PPPM methodologies, we focus on the calculation of the mechanical properties of a single polyelectrolyte brush. The system consisted of two planar solid surfaces composed of three layers of 324 DPD particles. The two surfaces were positioned at the top and bottom of the simulation cell. One of the two surfaces was coated with $N_p = 108$ polymer chains that were randomly grafted by a harmonic force acting between the end particles of the chains and the particles of the first layer of the wall. Each chain contained $N_b = 20$ polymer beads. The surface coverage was defined as $\rho_a = N_p/(L_xL_y)$. In our system, one-third of the wall particles of the third layer were connected to the first beads of the polymer chains. The different charge fractions $f$ were 1 (completely negatively charged), 0.5 (half fully charged), and zero (neutral). To preserve electroneutrality, there are $fN_pN_b$ counterions. The number of solvent particles was adjusted so that the overall reduced density between the two walls was close to 3. The cell dimensions are given in Table 1, and a representation of the simulation geometry is shown in Figure 2c. The $a_{ij}$ parameters were set to 25 for all interactions. The polymer brush was then modeled in athermal solvent conditions. To respect the supercell approximation, the simulation cell was elongated along the $z$ direction (see Table C-1 in Appendix C), and the reciprocal vector $\mathbf{h}_z^{\max} = 17$ was changed accordingly. The time-step was fixed at 0.06, and the simulations consisted of an equilibration period of 100 000 steps, followed by an acquisition period of 300 000 steps.

Figure 7a shows the normal components of the total pressure calculated along the $z$ direction according to the IK and MOP methods with the EW3DC methodology for the system with $f = 0.5$. Figure 7d compares the profiles of the total pressure calculated according to the MOP with EW3DC and PPPM. First, we found that the mechanical equilibrium is recovered for polyelectrolyte brushes with a very flat profile of the pressure across the simulation cell. This homogeneous profile of the normal component is independent of the method used for the pressure calculation and of the method used for the calculation of the long-range electrostatic interactions. Figure 7b shows the profiles of the normal component of the configurational (left axis) and electrostatic (right axis) pressure. Whereas the total pressure (left axis) exhibits a flat profile, one observes that the configurational part shows a positive contribution close to the grafted surface that is compensated by a symmetric negative value of the electrostatic pressure. One can observe

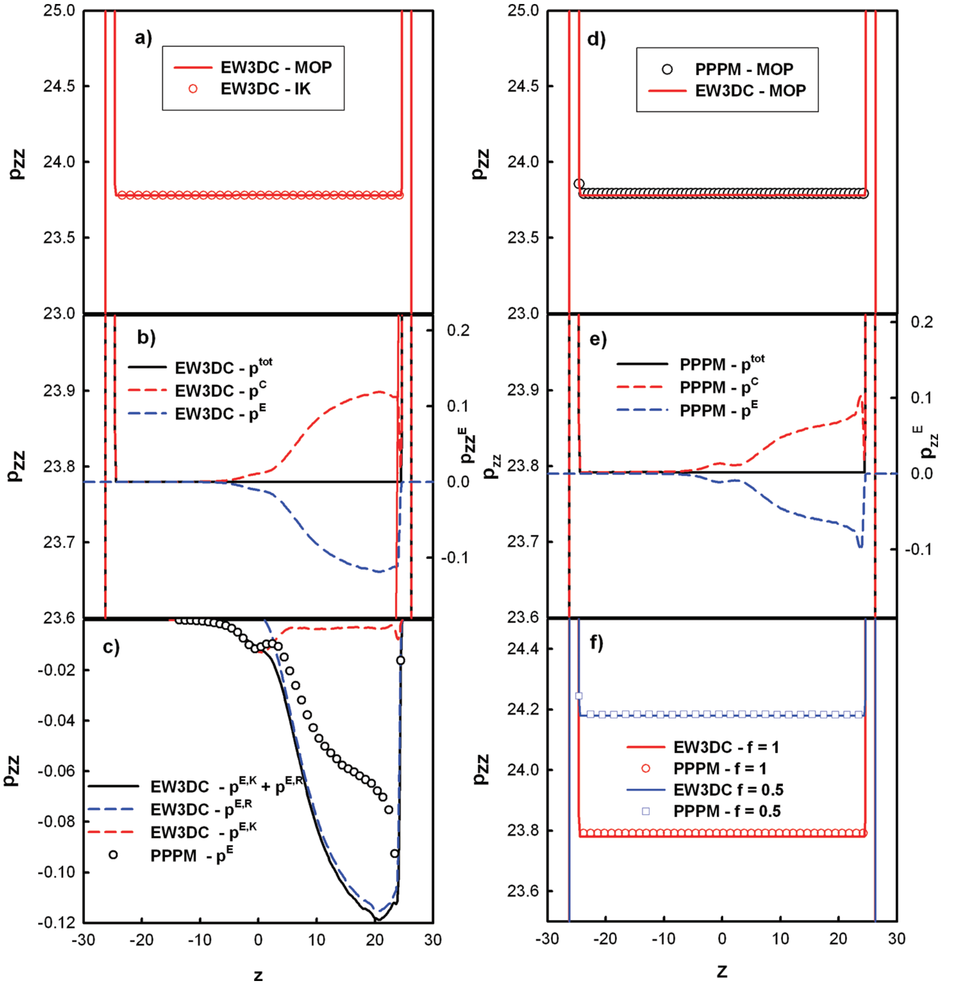


*Figure 7.* Normal component of the pressure ($p_{zz}$) as a function of the distance from the grafting surface calculated according to the MOP definition for (a) the total pressure calculated using EW3DC (included for comparison is the pressure profile calculated using IK); (b) the electrostatic ($p^{E}$), configurational ($p^{C}$), and total parts calculated using EW3DC; (c) the real-space ($p^{E,R}$) and reciprocal-space ($p^{E,K}$) contributions with the total electrostatic contribution calculated according to the EW3DC and PPPM methods; (d) the total pressure calculated using the EW3DC and PPPM methods; (e) the electrostatic ($p^{E}$), configurational ($p^{C}$), and total parts calculated using PPPM; and (f) the total pressure component for two different fractions of charges *f*. The configurational and total parts of the pressure in parts b and e are represented on the left axis, whereas the electrostatic pressure is represented on the right axis.

the same features with the PPPM methodology with slightly different profile shapes. The total pressures calculated by EW3DC and PPPM are identical. Figure 7c shows the profiles of the electrostatic pressure obtained by separating the real and reciprocal parts with the EW3DC method. The total electrostatic pressure is then compared to that calculated with the PPPM method. The slight differences between the pressure profiles obtained with EW3DC and PPPM are due to the fact that the force used with EW3DC is shifted to smaller distances compared to that used with PPPM (see Figure 1). Figure 7f shows that the total pressure decreases as the fraction of charges changes from 1 to 0.5.

Figure 8a shows the monomer density profiles $\rho_m(z)$ as a function of the distance from the grafting surface for three different fractions of charges (neutral, half-charged, and fully charged). First, we distinguish no difference in the density profiles obtained by the EW3DC and PPPM methods. We observe that, when the fraction of charge is increasing, the brush extends farther in the direction normal to the surface, although the profile remains parabolic. The electrostatic contribution to the pressure is relatively small compared to that of the configurational part, as the presence of the electrostatic interactions induces a strong stretching of the chains in the *z* direction. The repulsive electrostatic interactions between polymer chains and between beads within the chains tend to swell the brush and to straighten the polymer chain, respectively. These effects are lessened by the counterions that act to screen the charged monomer interactions. The decomposition of the electrostatic pressure into monomer−monomer, counterion−counterion, and monomer−counterion contributions yields values of 0.831, 0.836, and −1.691, respectively, for $f = 0.5$ and 2.717, 2.720, and −5.482, respectively, for $f = 1.0$. The resulting total electrostatic pressure within the brush is then equal to −0.024 for $f = 0.5$ and −0.045 for $f = 1.0$ in reduced units. This calculation shows the ability of the counterion to screen the charged monomer interactions. However, the reduction in entropy due to the presence of counterions in a small volume occupied by the polymer chains must be compensated by extending the chains against their elasticity. The competition

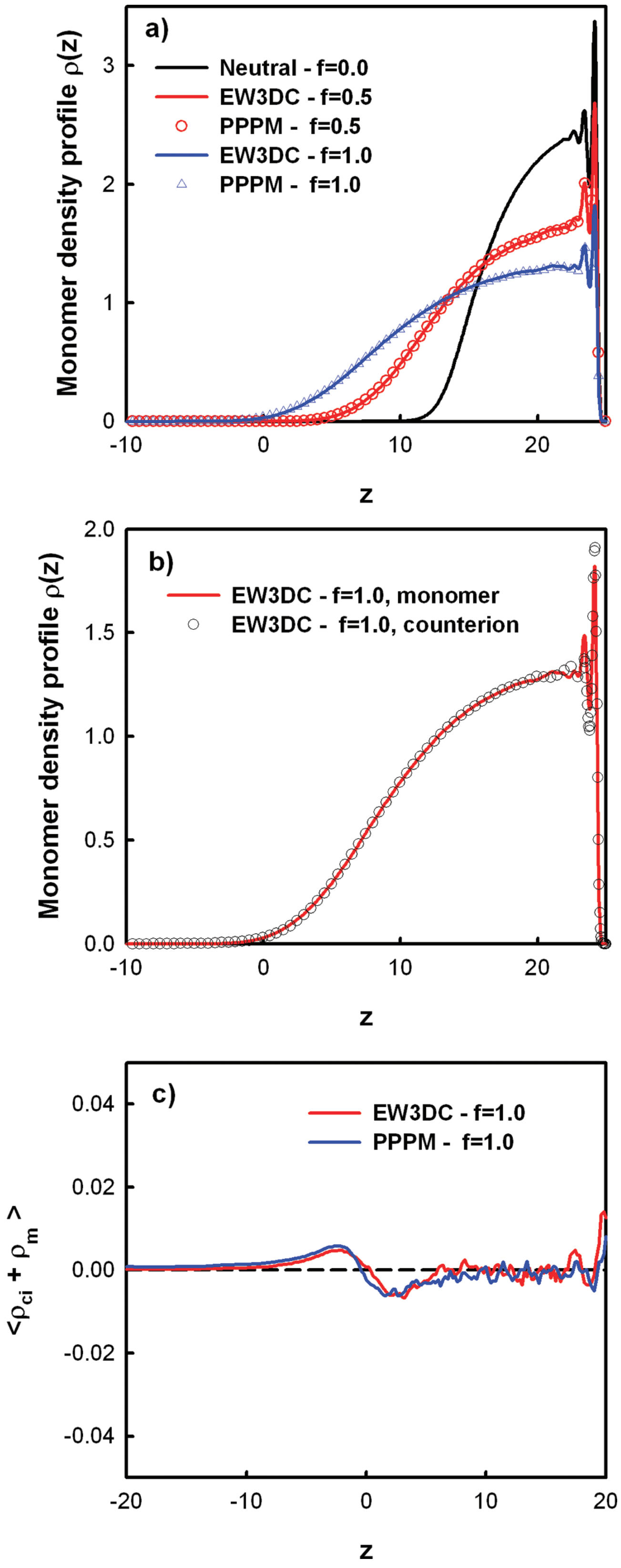


***Figure 8.*** Monomer density profiles of the single polyelectrolyte brush calculated with different degrees of charged monomers.

among the entropy, Coulomb interactions, and chain elasticity is complex and is difficult to consider analytically in the theoretical predictions of polylectrolytes properties. To measure the extension of the brush, we calculated the average heights of the brush $\langle z_m\rangle$ and of the counterion layer $\langle z_{ci}\rangle$ from the first moment of the density profiles of monomers and counterions as

$$\langle z_m\rangle = 2\frac{\int_0^\infty z\rho_m(z)\,dz}{\int_0^\infty \rho_m(z)\,dz} \quad \text{and} \quad \langle z_{ci}\rangle = 2\frac{\int_0^\infty z\rho_{ci}(z)\,dz}{\int_0^\infty \rho_{ci}(z)\,dz} \tag{22}$$

where $\rho_m(z)$ and $\rho_{ci}(z)$ are the density profiles of the monolayer and the counterions, respectively. The factor of 2 takes into account the fact that the brush height is twice the first moment when the monomer density profile is uniform inside the brush. The different brush heights are reported in Table 2, along with the average distance $\langle b\rangle$ between neighboring beads within the polymer chain. This table highlights that the brush height is increased by 40% and 70% with respect to that of a neutral brush when the fractions of charges are 0.5 and 1.0, respectively. This elongation of the chain is reflected in part in the increase of the bond length in the polymer chain. This is allowed in part because of the soft potential used for the bond between neighboring beads (see eq 6). For the fully charged polymer chains, a brush height of 19 indicates that the chains are stretched to about 86% of their contour length defined from the limiting value for $N_b\langle b\rangle$ of 22 for a fully extended chain structure.

Part b of Figure 8 shows the density profiles of the polymer beads and the counterions. Interestingly, this figure emphasizes that the profile of the polymer beads is coincident with that of the counterions. This indicates that the counterions are mostly confined in the brush layer. The thickness of the layer of the counterions calculated from the first moment of the density profiles is listed in Table 2. As expected from the density profiles, the height of the counterion layer is very close to the brush height. It then becomes important to check the local electroneutrality across the brush by plotting the sum of the charges of the counterions and the polymer beads along the $z$ direction. We note that the local eletroneutrality is satisfied in the direction perpendicular to the surface over almost the total brush height with two exceptions. The first relates to the local charge for $z > 20$ due to the layering of the grafted monomers close to the surface. The region close to the surface is not shown in part c of Figure 8. We also observe in the region close to the ends of the grafted chains a depletion of counterions with a negative local net charge, followed by a zone rich in counterions with a positive local charge. This leads to the formation of a local dipole. This has already been observed in the simulation of polyelectrolytes.[30,42] We also note that the electric properties in the brush layer are reproduced in the same way with the EW3DC and PPPM methodologies.

Figure 9a depicts the polyelectrolyte−ion pair distribution $p(r)$, where $r$ corresponds to the separation distance between the ion and the closest polyelectrolyte bead. The distributions are normalized according to $2\pi\int_0^\infty rp(r)\,dr = 1$. This figure shows that these distributions are centered around 0.6 and 0.7 when the degree of charge decreases from 1.0 to 0.5. When the polymer chains are fully charged, the electrostatic interactions are stronger, and the

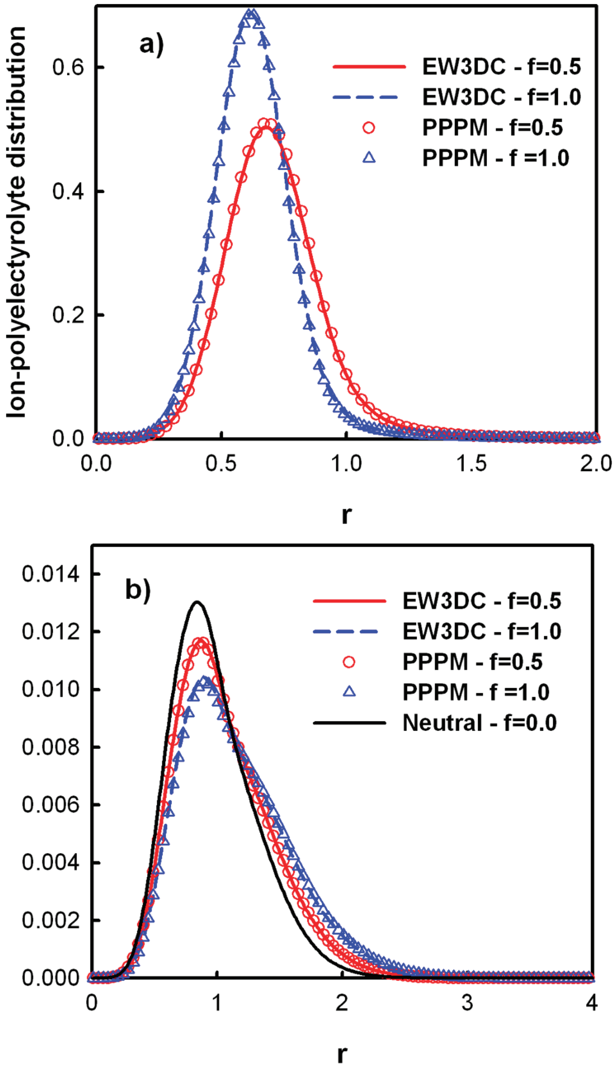


***Figure 9.*** (a) Ion−polyelectrolyte distribution function calculated according to the EW3DC and PPPM methods for a fully charged polymer chain and a half-charged polymer chain, where *r* is the separation distance between polyelectrolyte bond and counterion, and (b) bead−bead distribution function between connected beads within the polymer chain as a function of the separation distance.

counterions are further trapped inside the chains to screen the charged monomer interactions. The resulting bead−counterion separation distance is then reduced. This figure also shows that there is a strong correlation between polyelectrolyte beads and counterions. The result is an inhomogeneous distribution of the counterions and contrasts with Pincus' theory,[43] which assumes that counterions form an ideal gas. Part b of Figure 9 shows the distribution of the distance between neighboring beads in the polymer chains and confirms the fact that the average bond length increases with the strength of the Coulomb interactions. These two structural properties illustrate that the entropic pressure of the confined counterions stretches the chains against their elasticity.

*5.3.2. Scaling Properties.* The basic behavior of polyelectrolyte brushes can be understood on the basis of simple scaling theory. This rationalization through the use of simple scaling arguments is sometimes useful, but the approximations used (shape of the monomer density profiles, strength of the electrostatic interactions, etc.) have not been validated by molecular simulation or experiment. The Bjerrum length, given by $\lambda_B = e^2/(4\pi k_B T\varepsilon_0\varepsilon_r)$, characterizes the length scale at which the electrostatic interaction is equal to the thermal energy $k_BT$. According to the parameters used in this work, the Bjerrum length is equal to 1.11 ≡ 7.16 Å. The dimensionless Manning ratio[44−46] is defined as $\lambda_B/\langle b\rangle$, where $\langle b\rangle$ is the average bond length distance between beads in the polyelectrolyte chain. In Manning's theory, the condensation of counterions occurs at $\lambda_B/\langle b\rangle$ =1. The values calculated for the bond length lead to values of the Manning ratio very close to 1 and are in line with the counterion condensation observed in our simulations. The degree of condensed counterions can be estimated from Figure 9a by assuming them to be condensed if the polyion−counterion distance is smaller than $\lambda_B$. This counterion condensation is 97% with $f$ = 0.5 and $f$ = 1. The Debye screening length associated with the counterions is defined as $\lambda_D = 1/(4\pi\lambda_B c_{ci})^{1/2}$, where $c_{ci}$ can be estimated from the average density of counterions inside the brush. This parameter, defining the scale over which mobile charges are screened, is equal to 0.25 and 0.32 when the charge fraction is 0.5 and 1.0, respectively. This value, which is smaller than the bond length, explains in part why the counterions are trapped in the brush layer. Within the range of parameters used, the simulated brush is in the strong-charging and strong-streching limits. As a consequence, the model of polyelectrolyte brushes used follows the nonlinear osmotic brush regime.[47] This regime[47] combines the high-streching (nonlinear) version of the chain elasticity with the nonlinear entropic effects of the counterions inside the brush. This is also the case for previous molecular simulations of strongly charged polyelectrolyte brushes.[30,31,42,48]

Another typical length in the theory of polyelectrolyte brushes is the Gouy−Chapman length, $\lambda_{GC}$, defined as $1/(2\pi\lambda_B N_b f\rho_a)$, where $N_b$ and $\rho_a$ are the polymer chain length and the grafting density, respectively. This length defines the height at which counterions are effectively bound to a surface of charge density of $efN_b\rho_a$.[43] For a fully charged brush and a strong grafting density, the Gouy−Chapman length is on the order of 0.015 ≡ 0.1 Å and can increase up to 0.015 ≡ 0.1 Å for higher grafting densities. Within the nonlinear osmotic brush regime, the height of the counterion layer $H$ is equal to the brush height plus $3\lambda_{GC}/2$. The weak values of $\lambda_{GC}$ mean that the height of the brush is the same as that of a counterion layer with a high concentration of counterions inside the brush. It will be very interesting to investigate the dependence of the brush height on the grafting density within this regime, but we will consider this in a future work.

## 6. Conclusions

We have performed mesoscale modeling of different electrolyte systems: a bulk electrolyte, an electrolyte embedded between two surfaces, and a single polyelectrolyte brush. We have used the dissipative particle dynamics method to capture the physics of these complex systems at length and time scales that are outside the ranges of standard molecular simulations. Two methods were established recently for the calculation of the electrostatic interactions in the DPD formalism. The first method, initially introduced by Groot,[14] is an adaptation of the particle−particle particle-mesh (PPPM) method. The second method, initially developed by Alejandre and coworkers,[15] consists of using the standard Ewald summation method with charge distributions on particles to avoid the formation of artificial ion pairs (EW3DC).

We used the supercell approximation to allow the use of the three-dimensional EW3DC and PPPM methods in systems presenting a finite length along a given direction. We used two different definitions for the calculation of the local pressure. The Irving and Kirkwood definition is well-adapted for pairwise-additive forces and can be straitforwardly used with the EW3DC method. The PPPM method does not give a non-pairwise-additive electrostatic force, and the calculation of the local pressure in this method can be performed with the method of planes (MOP). We showed that the different techniques EW3DC and PPPM give similar profiles for the normal component of the pressure tensor for the configurational and electrostatic contributions. We also showed that the use of the supercell approximation with appropriate definitions of the pressure tensor allows for the calculation of the local pressure in agreement with that expected from a bulk electrolyte at the same density.

The calculation of the local pressure in a single polyelectrolyte brush demonstrated a positive configurational contribution to the pressure from the brush and a similar negative electrostatic part from the brush. The profile of the total pressure along the direction normal to the surface is flat, as expected for a system at mechanical equilibrium. We completed the study of the polyelectrolyte brush by calculating the heights of the brush and the counterion layer. We found that the counterions are mostly trapped in the brush and that the condensation of counterions increases with the fraction of charged monomers. The stretching of the polymer chain was found to about 85% of its contour length. These results were expected from the values of the dimensionless Manning ratio and the Debye screening length of the counterions. The weak value of the Gouy−Chapman length indicates that the height of the polymer brush must be equal to that of the counterion layer. The system model simulated here is in the strong-charging and strong-stretching limits and follows the nonlinear osmotic brush regime.

We also showed that the incorporation of the electrostatic interactions through the EW3DC and PPPM methods into the DPD methodology allows the main properties of a single strongly stretched polyelectrolyte brush made of strongly charged polymers to be recovered. This preliminary study calls for further investigation of the dependence of the brush height on the surface coverage, fraction of charge, and salt concentration. It is also interesting to see that an accurate calculation of the local pressure can be carried out in such systems at a mesoscopic scale. This represents an important step for the next calculation of the frictional forces in polyelectrolyte brushes under shear.

Additionally, the PPPM method is more efficient in CPU time than the EW3DC method as the number of charges increases. This makes the PPPM method a powerful and attractive method for the mesoscale modeling of polyelectrolytes.

**Acknowledgment.** The authors acknowledge A. Ghoufi, M. Gonzales-Melchor, F. Goujon, and R. D. Groot for helpful discussions.

## Appendix A. Expression of the Electrostatic Potential in the DPD Method with the PPPM and EW3DC Techniques

The empirical expressions of the electrostatic potential used in the PPPM method[14] are given by

$$\frac{4\pi r_e U(r)}{\Gamma r_c} = \begin{cases} \frac{52}{35} - \frac{4}{5}\left(\frac{rr_c}{r_e}\right)^2 + \frac{2}{5}\left(\frac{rr_c}{r_e}\right)^4 - 0.13587\left(\frac{rr_c}{r_e}\right)^{5.145} & (r < r_e/r_c) \\ \frac{r_e}{rr_c} - 3.2100\left(1 - \frac{rr_c}{2r_e}\right)^6 & (r_e/r_c < r < 2r_e/r_c) \\ \frac{rr_c}{r_e} & (r < r_e/r_c) \end{cases} \tag{A-1}$$

The potential and the corresponding force are represented in Figure 1.

The electrostatic potential used in the DPD method with the EW3DC technique[15] is given by

$$\frac{4\pi r_e U(r)}{\Gamma r_c} = \frac{1}{r}[1 - (1 + \beta r)\exp(-2\beta r)] \tag{A-2}$$

and its corresponding force is given by

$$\frac{4\pi r_e^2 f(r)}{\Gamma r_c^2} = \frac{1}{r^2}\{1 - \exp(-2\beta r)[1 + 2\beta r(1 + \beta r)]\} \tag{A-3}$$

The expressions for the energy and forces are represented in Figure 1 for comparison with those used in the PPPM method.

## Appendix B. Calculation of the Electrostatic Contributions to the Pressure of the EW3DC Method with the IK and Harasima Definitions

The contribution of the real space to the local pressure is given by

$$p_{\alpha\beta}^{\mathrm{E,R}}(z) = \frac{\Gamma}{4\pi}\left\langle \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} q_i q_j\left[\frac{2}{\sqrt{\pi}}\alpha r_{ij}\exp(-\alpha^2 r_{ij}^2) + \mathrm{erfc}(\alpha r_{ij})\right]\frac{(\mathbf{r}_{ij})_\alpha \cdot (\mathbf{r}_{ij})_\beta}{r_{ij}^3}\frac{1}{|z_{ij}|}\theta\left(\frac{z - z_i}{z_{ij}}\right)\theta\left(\frac{z_j - z}{z_{ij}}\right)\right\rangle \tag{B-1}$$

***Table C-1.*** DPD Parameters in Reduced Units, along with Some Typical Parameters in Real Units

| parameter | DPD value | real units | definition |
|---|---|---|---|
| DPD | | | |
| $a_{ij}$ | 25 | $1.59 \times 10^{-19}$ J m$^{-1}$ | repulsion parameter between particles *i* and *j* |
| $T$ | 1 | 298 K | temperature |
| $m$ | 1 | $8.97 \times 10^{-27}$ kg | particle mass |
| $N_m$ | | 3 | number of water molecules within one bead |
| $r_c$ | 1 | 6.4633 Å | range of repulsive interaction |
| $\rho$ | 3 | 1.0 g cm$^{-3}$ | density |
| $\delta t$ | 0.02−0.06 | 3.2−9.6 ps | time step |
| $\tau$ | 1 | 160 ps | time scale |
| $\sigma$ | 3 | $2.92 \times 10^{-11}$ J m$^{-1}$ s$^{1/2}$ | noise amplitude |
| $\gamma$ | 4.5 | $1.04 \times 10^{-1}$ J m$^{-2}$ s | friction coefficient |
| $v$ | 1.0 | 4.04 m s$^{-1}$ | velocity |
| $p$ | 1.0 | 15.22 MPa | pressure |
| $r_e$ | 1.6 | 10.34 Å | range over which charges are smeared |
| DPD + EW3DC | | | |
| $\alpha$ | 0.9695 | 0.15 Å$^{-1}$ | Ewald convergence factor |
| $r_c^R$ | 3.0 | 19.39 Å | electrostatic cutoff for the real part |
| $\mathbf{h}_x^{max} = 5$ | $\mathbf{h}_y^{max} = 5$ | $\mathbf{h}_z^{max} = 5$ | bulk electrolyte |
| $\mathbf{h}_x^{max} = 5$ | $\mathbf{h}_y^{max} = 5$ | $\mathbf{h}_z^{max} = 17$ | bulk electrolyte between two surfaces |
| $\mathbf{h}_x^{max} = 8$ | $\mathbf{h}_y^{max} = 3$ | $\mathbf{h}_z^{max} = 76$ | single polyelectrolyte brush |

and that of the reciprocal space is given by

$$p_{\alpha\beta}^{\mathrm{E,K}}(z) = \frac{\Gamma}{4\pi}\left\langle \frac{2\pi}{V^2}\sum_{\mathbf{h}\neq 0} H(z_i)\, Q(h)\, S(\mathbf{h})\, S(-\mathbf{h}) \times \left(\delta_{\alpha\beta} - \frac{2\mathbf{h}_\alpha \mathbf{h}_\beta}{h^2} - \frac{\mathbf{h}_\alpha \mathbf{h}_\beta}{2\alpha^2}\right)\right\rangle \tag{B-2}$$

where $H(z_i)$ is a top-hat function defined as

$$H(z_i) = \begin{cases} 1 & \text{if } z - \Delta z/2 < z_i < z + \Delta z/2 \\ 0 & \text{otherwise} \end{cases} \tag{B-3}$$

## Appendix C. DPD Parameters, Physical Length, and Time Scales

In the simulations, the particle mass, temperature, and interaction range were chosen as units of mass, energy, and length, respectively; hence $m = k_BT = r_c = 1$. The unit of time, $\tau$, then becomes $r_c(m/k_BT)^{1/2}$. The real length $r_c$ can be estimated from the volume of a DPD bead. If $N_m$ represents the number of water molecules within a DPD particle, then $r_c = (\rho^* N_m V_m / N_A)^{1/3}$, where $\rho^*$ is the reduced density of DPD particles, $V_m = 18$ cm $^3$ mol$^{-1}$, and $N_A$ is Avogadro's number. Groot and Rabone[8] and Groot[14] used $N_m = 3$ and a reduced density of 3. Using such values, $r_c = 6.46$ Å. Making the Poisson equation dimensionless[14] implies that the coupling constant $\Gamma$ is given by $e^2/(k_BT\varepsilon_0\varepsilon_r r_c)$, where $e$ is the electron charge, $\varepsilon_0 = 8.85418782 \times 10^{-12}$ C$^2$ J$^{-1}$ m$^{-1}$ is the dielectric constant of a vacuum, and $\varepsilon_r = 78.3$ is the relative permittivity of water at 298 K. Using $r_c = 6.46$ Å and $\Gamma = 13.87$, to match the interaction between two charge clouds at $r = 0$ within the PPPM and EW3DC methodologies, the $\beta$ parameter is 0.929.[15] As already dicussed by Groot and Rabone,[8] the time scale is fixed by matching the diffusion constant of water. For the repulsion parameter $a = 25$, we found that the natural unit of time $\tau$ is 160 ps.

The complete list of DPD parameters in reduced units is provided in Table C-1.

## References

(1) Hoogerbrugge, P. J.; Koelman, J. M. V. A. *Europhys. Lett.* **1992**, *19*, 155.

(2) Koelman, J. M. V. A.; Hoogerbrugge, P. J. *Europhys. Lett.* **1993**, *21*, 363.

(3) Espanol, P.; Warren, P. B. *Europhys. Lett.* **1995**, *30*, 191.

(4) Espanol, P. *Europhys. Lett.* **1997**, *40*, 631.

(5) Groot, R. D.; Madden, T. J. *J. Chem. Phys.* **1998**, *108*, 8713.

(6) Groot, R. D.; Madden, T. J.; Tildesley, D. J. *J. Chem. Phys.* **1999**, *110*, 9739.

(7) Groot, R. D. *Langmuir* **2000**, *16*, 7493.

(8) Groot, R. D.; Rabone, K. L. *Biophys. J.* **2001**, *81*, 725.

(9) Malfreyt, P.; Tildesley, D. J. *Langmuir* **2000**, *16*, 4732.

(10) Irfachsyad, D.; Tildesley, D. J.; Malfreyt, P. *Phys. Chem. Chem. Phys.* **2002**, *4*, 3008.

(11) Goujon, F.; Malfreyt, P.; Tildesley, D. J. *ChemPhysChem* **2004**, *5*, 100.

(12) Goujon, F.; Malfreyt, P.; Tildesley, D. J. *Mol. Phys.* **2005**, *103*, 2675.

(13) Raviv, U.; Giasson, S.; Kampf, N.; Gohy, J. F.; Jerome, R.; Klein, J. *Nature* **2003**, *425*, 163.

(14) Groot, R. D. *J. Chem. Phys.* **2003**, *118*, 11265.

(15) Gonzalez-Melchor, M.; Mayoral, E.; Velazquez, M. E.; Alejandre, J. *J. Chem. Phys.* **2006**, *125*, 224107/1.

(16) Ewald, P. P. *Ann. Phys.* **1921**, *64*, 253.

(17) Goujon, F.; Malfreyt, P.; J. Tildesley, D. *J. Chem. Phys.* **2008**, *129*, 034902.

(18) Irving, J. H.; Kirkwood, J. G. *J. Chem. Phys.* **1950**, *18*, 817.

(19) Harasima, A. *Adv. Chem. Phys.* **1958**, *1*, 203.

(20) Alejandre, J.; Tildesley, D. J.; Chapela, G. A. *J. Chem. Phys.* **1995**, *102*, 4574.

(21) Sonne, J.; Hansen, F. Y.; Peters, G. H. *J. Chem. Phys.* **2005**, *122*, 124903.

(22) Todd, B. D.; Evans, D. J.; Daivis, P. J. *Phys. Rev. E* **1995**, *52*, 1627.

(23) Yeh, I. C.; Berkowitz, M. L. *J. Chem. Phys.* **1999**, *111*, 3155.

(24) Crozier, P. S.; Rowley, R. L.; Spohr, E.; Henderson, D. *J. Chem. Phys.* **2000**, *112*, 9253.

(25) Goujon, F.; Bonal, C.; Limoges, B.; Malfreyt, P. *Mol. Phys.* **2008**, *106*, 1397.

(26) Hautman, J.; Klein, M. *Mol. Phys.* **1992**, *75*, 379.

(27) Lekner, J. *Physica A* **1991**, *176*, 485.

(28) Strebel, R.; Sperb, R. *Mol. Simul.* **2001**, *27*, 61.

(29) Arnold, A.; Holm, C. *Comput. Phys. Commun.* **2002**, *148*, 327.

(30) Kumar, N. A.; Seidel, C. *Macromolecules* **2005**, *38*, 9341.

(31) Kumar, N. A.; Seidel, C. *Phys. Rev. E* **2007**, *76*, 020801.

(32) Beckers, J. V. L.; Lowe, C. P.; de Leeuw, S. *Mol. Simul.* **1998**, *20*, 368.

(33) Deserno, M.; Holm, C. *J. Chem. Phys.* **1998**, *109*, 7678.

(34) Sagui, C.; Darden, T. *J. Chem. Phys.* **2001**, *114*, 6578.

(35) Eastwood, J. W.; Hockney, R. W.; Lauwrence, D. *Comput. Phys. Commun.* **1980**, *19*, 215.

(36) Smith, E. R. *Proc. R. Soc. London, Ser. A* **1981**, *375*, 475.

(37) Rowlinson, J. S.; Widom, B. *Molecular Theory of Capillarity*; Clarendon Press: Oxford, U.K., 1982.

(38) Walton, J. P. R. B.; Tildesley, D. J.; Rowlinson, J. S.; Henderson, J. R. *Mol. Phys.* **1983**, *48*, 1357.

(39) Walton, J. P. R. B.; Tildesley, D. J.; Rowlinson, J. S. *Mol. Phys.* **1986**, *58*, 1013.

(40) Toukmaji, A. Y.; Board, J. A., Jr. *Comput. Phys. Commun.* **1996**, *95*, 73.

(41) Perram, J.; Petersen, H.; Leeuw, S. D. *Mol. Phys.* **1988**, *65*, 875.

(42) Csajka, F. S.; Seidel, C. *Macromolecules* **2000**, *33*, 2728.

(43) Pincus, P. *Macromolecules* **1991**, *24*, 2912.

(44) Manning, G. S. *Biophys. Chem.* **1977**, *7*, 95.

(45) Manning, G. S. *Q. Rev. Biophys.* **1978**, *11*, 179.

(46) Heath, P. J.; Schurr, J. M. *Macromolecules* **1992**, *25,* 4149.

(47) Ahrens, H.; Förster, S.; Helm, C. A.; Kumar, N. A.; Naji, A.; Netz, R. R.; Seidel, C. *J. Phys. Chem. B* **2004**, *108*, 16870.

(48) Seidel, C. *Macromolecules* **2003**, *36*, 2536.

CT900296S

JCTC Journal of Chemical Theory and Computation

# "Reverse-Schur" Approach to Optimization with Linear PDE Constraints: Application to Biomolecule Analysis and Design

Jaydeep P. Bardhan,[†] Michael D. Altman,[‡] B. Tidor,*[,||,§] and Jacob K. White*[,||]

*Department of Molecular Biophysics and Physiology, Rush University Medical Center, Chicago, Illinois, Merck Research Laboratories, Boston, Massachusetts, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, and Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts*

Received March 12, 2009

**Abstract:** We present a partial-differential-equation (PDE)-constrained approach for optimizing a molecule's electrostatic interactions with a target molecule. The approach, which we call *reverse-Schur co-optimization*, can be more than 2 orders of magnitude faster than the traditional approach to electrostatic optimization. The efficiency of the co-optimization approach may enhance the value of electrostatic optimization for ligand-design efforts. In such projects, it is often desirable to screen many candidate ligands for their viability, and the optimization of electrostatic interactions can improve ligand binding affinity and specificity. The theoretical basis for electrostatic optimization derives from linear-response theory, most commonly continuum models, and simple assumptions about molecular binding processes. Although the theory has been used successfully to study a wide variety of molecular binding events, its implications have not yet been fully explored, in part due to the computational expense associated with the optimization. The co-optimization algorithm achieves improved performance by solving the optimization and electrostatic simulation problems simultaneously and is applicable to both unconstrained and constrained optimization problems. Reverse-Schur co-optimization resembles other well-known techniques for solving optimization problems with PDE constraints. Model problems as well as realistic examples validate the reverse-Schur method and demonstrate that our technique and alternative PDE-constrained methods scale very favorably compared to the standard approach. Regularization, which ordinarily requires an explicit representation of the objective function, can be included using an approximate Hessian calculated using the new BIBEE/P (boundary-integral-based electrostatics estimation by preconditioning) method.

## 1. Introduction

The problem of optimizing electrostatic interactions is a task of particular importance in molecular design. One asks whether a candidate designed molecule, or *ligand*, is optimal for binding the target molecule, which is called a *receptor*, and if not, what chemical modifications might be made to improve binding affinity or specificity. A variety of factors contribute to the binding free energy, including conformational entropy, although often the contributions are dominated by packing effects and electrostatics. Although the short-range packing interactions can be conceptualized relatively easily, analysis of the electrostatic component is more complex. The electrostatic component of the binding free energy can be particularly nonintuitive due to the interactions' long range and the trade-off between favorable ligand−receptor interactions in the bound state and the

* Corresponding authors: e-mail: tidor@mit.edu; white@mit.edu.

[†] Rush University Medical Center.

[‡] Merck Research Laboratories.

[||] Department of Electrical Engineering and Computer Science, MIT.

[§] Department of Biological Engineering, MIT.

10.1021/ct9001174 CCC: $40.75 © 2009 American Chemical Society
Published on Web 10/30/2009

unfavorable desolvation penalties paid on binding.[1] These nonintuitive features have led to the important but challenging goal of designing optimal electrostatic interactions as an approach to designing useful molecular binding partners.[1,2] Questions in molecular biology regarding the evolution of biomolecules, whether to serve specific functions or to bind targets with high affinity and specificity, may also be interpreted as questions regarding optimization of a particular objective function.[2–4]

Lee and Tidor presented the first work describing the possibility of optimizing electrostatic interactions between molecules,[1] showing that linear-response theory and simple assumptions about binding events—in particular, that the ligand binds rigidly and that no charge redistribution occurs on binding—give rise to a quadratic model for the electrostatic contribution to the binding free energy. Their primarily analytical study used a multipole-based representation of the ligand charge distribution and spherical geometries for the unbound ligand and the ligand−receptor complex. Chong et al. applied this theory to an idealized model of the protein barnase and found that small sets of biochemically reasonable charge distributions resembled the computed optimal charge distribution.[5] Kangas and Tidor later proved that the electrostatic component of the binding free energy is a convex function under reasonable assumptions and extended the theory to address nonspherical geometries, alternative basis sets, and measures of binding specificity.[6–8]

Following these developments, Lee and Tidor studied the interactions between two proteins, the extremely tight-binding partners barnase and barstar;[3,9] their analysis suggested that the inhibitor barstar is electrostatically optimized to bind to the enzyme barnase. In another application of the optimization theory, Kangas and Tidor studied the enzyme *B. subtilis* chorismate mutase.[2] This investigation indicated a particularly promising modification to improve the binding affinity of a transition-state analog inhibitor—the replacement of a carboxylate group by a nitro group. Mandal and Hilvert synthesized the proposed inhibitor; in agreement with the computational analysis, the resulting ligand bound the enzyme more tightly and was in fact the tightest-binding chorismate mutase inhibitor reported to date in the literature.[10]

Several groups have applied the optimization theory to study a number of other molecular systems. Sulea and Purisima have studied cation−protein binding, the optimization of protein−protein interfaces, and the use of the charge optimization framework as a means to identify "hot spots" for binding.[11,12] Sims et al. studied two protein kinases, protein kinase A (PKA) and cyclin-dependent kinase 2 (CDK2), and several inhibitors.[13] Green and Tidor have applied charge optimization theory to two systems.[4,14] In one study, they demonstrated that glutaminyl-tRNA synthetase is optimized for its substrates;[4] more recently, they proposed optimization-theory-based mutations to 5-Helix, which inhibits HIV-1 membrane fusion by gp41.[14] Armstrong et al. have studied several inhibitors of neuraminidase and the relation between charge optimization and lead progression.[15] Gilson has explored the theory allowing the optimization of flexible ligands.[16] Schreiber and collaborators have also focused on optimizing ligand−receptor electrostatics.[17,18] Brock et al. have used a theory similar to Lee and Tidor's in their analysis of protein−protein complexes.[19]

The computational expense associated with optimization has limited the broad application of the electrostatic-optimization theory. Traditional approaches to optimization begin with an explicit representation of the second-derivative, or *Hessian*, matrix or a means by which to multiply a vector by the Hessian. For electrostatic optimization problems, a large cost associated with calculating the Hessian matrix explicitly is the successive simulation of the bound and unbound systems with each of the point charges (more generally, the basis functions) used to describe the ligand charge distribution.[1] The cost to form the Hessian must be paid before optimization can be performed, and it scales essentially linearly with the number of basis functions for reasonably sized problems. Lee and Tidor emphasized, in their original optimization paper, the importance of using sufficiently complete basis function sets to achieve convergence to the optimum affinity.[1] However, the fixed-location point-charge basis sets used in most electrostatic optimization work offer little insight into geometric sensitivity or to basis-set completeness. The availability of more efficient computational methods may therefore enable not only greater numbers of ligands to be optimized in design efforts but also a more thorough exploration of the optimization theory itself and the extent to which biology may have employed electrostatic optimization to achieve desired binding affinities and specificities.

This paper presents a new, highly efficient approach, which we call *reverse-Schur co-optimization*, to solving the electrostatic optimization problem. The theory and implementation of efficient methods for optimization problems constrained by partial differential equations (PDEs) have become a progressively more important research topic over the past several years,[20–23] and we show that electrostatic optimization is actually a special case of a PDE-constrained optimization problem. Most PDE-constrained optimization techniques follow one of two approaches. *All-at-once* approaches incorporate the PDE state variables (for electrostatic optimization, the electrostatic potentials in the bound and unbound states) directly into the optimization problem. The state variables together with the decision variables (the point charge values) satisfy the PDE, which is included as an equality constraint.[22–25] Such techniques are often termed *simultaneous analysis and design* (SAND) approaches.[20] The second general strategy, sometimes called a *black-box* approach, hides the PDE from the optimization algorithm.[23] The *nested analysis and design* (NAND) paradigm is a black-box method,[26,27] as are techniques that directly invert the PDE constraint before initiating optimization. The calculation of an explicit Hessian is effectively a black-box approach, because the mathematical details of the PDE simulation are entirely hidden from the optimization procedure. In electrostatic optimization problems, the decision variables and the state variables are related by a linear matrix equation. As we show in this paper, this linearity allows optimal charge distributions to be found without calculating the Hessian explicitly and without using the discretized PDE as an equality constraint.

To demonstrate the co-optimization method's performance on problems of therapeutic relevance, we have applied the optimization methodology to two protein–small-molecule-ligand complexes. The first is a complex between HIV-1 protease and the small-molecule inhibitor darunavir (TMC-114).[28] HIV-1 protease is an essential enzyme in the life cycle of HIV, and small-molecule inhibitors of the protease have been successful components of combinatorial strategies for treating HIV infection.[29,30] The second protein−ligand system studied is a complex between the protein cyclin-dependent kinase 2 (CDK2) and a small-molecule inhibitor.[31] The CDK family of proteins are involved in regulating cell growth, and inhibitors of these enzymes are potential cancer therapies.[32,33] It may be possible to use charge optimization to identify regions of these small-molecule ligands that are suboptimal for binding their protein target, and chemical modification at these locations may lead to improved inhibitors.

The following section describes a linear-response continuum model for biomolecule electrostatics, two boundary-integral formulations of the PDE problem, boundary-element methods (BEM) for solving the integral equations numerically, the optimization problem based on the linear-response model, and methods for convex quadratic optimization. Section 3 presents the new *reverse-Schur co-optimization* method. In addition, we describe two more widely used approaches to PDE-constrained optimization problems, partially to highlight differences between these methods and the reverse-Schur approach and partially to illustrate that the performance gains are not necessarily specific to the reverse-Schur method. Techniques for constrained co-optimization are also presented. Important details of the implementation—regularization methods and preconditioning—are described in Section 4. In Section 5 we present computational results that validate the method, demonstrate its computational efficiency, and show that realistic problems in biomolecule design can be studied using PDE-constrained optimization methods. Section 6 summarizes the paper and suggests future research directions.

## 2. Theory

**2.1. A Linear-Response Model for Estimating the Electrostatic Contribution to the Free Energy of Binding Between Biomolecules.** Free energies of binding are commonly estimated using a thermodynamic cycle such as that shown in Figure 1.[34] The lower set of images represents the ligand, receptor, and ligand−receptor complex in aqueous solvent, and the lower horizontal arrow represents the binding free energy $\Delta G^0_{\text{bind}}$ to be estimated. The unbound ligand and receptor and the bound complex are assumed to be at infinite dilution. The upper cartoons represent the three species in a homogeneous low-dielectric environment with zero ionic strength throughout, and the horizontal arrow denoted by $\Delta G^{0,\text{ref}}_{\text{bind}}$ represents the free energy change on binding in the low-dielectric environment, the electrostatic component of which is simply the ligand−receptor Coulomb-interaction energy. The three steps illustrated by vertical arrows involve the transfer of a molecule or complex between the low-dielectric environment and the solvent. The difference in a molecule's free energy as it is transferred into solvent from the reference low-dielectric medium is called its solvation free energy,[34] and this free energy is frequently decomposed into nonpolar and electrostatic terms so that

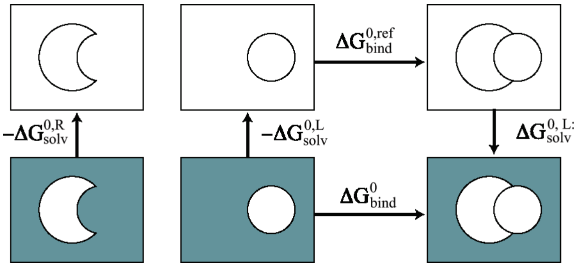

$\Delta G^{0,\text{ref}}_{\text{bind}}$
$-\Delta G^{0,\text{R}}_{\text{solv}}$
$-\Delta G^{0,\text{L}}_{\text{solv}}$
$\Delta G^{0,\text{L:R}}_{\text{solv}}$
$\Delta G^{0}_{\text{bind}}$

*Figure 1.* A thermodynamic cycle for estimating binding free energies. The shaded regions on the lower set of cartoons represent aqueous solvent. The upper cartoons represent a uniform low dielectric (the same as that of the ligand and receptor) with zero ionic strength throughout.

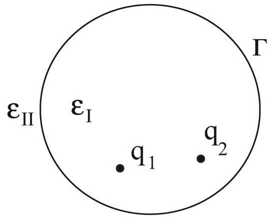


*Figure 2.* A mixed discrete-continuum model for estimating the electrostatic component of a solute's solvation free energy; $\varepsilon_I$ and $\varepsilon_{II}$ represent the dielectric constants of the solute and solvent regions, Γ is the boundary between the dielectric regions and is typically a molecular (solvent-excluded) surface, and $q_1$ and $q_2$ are representative discrete point charges in the solute.

$$\Delta G^0_{\text{solv}} = \Delta G^{0,\text{np}}_{\text{solv}} + \Delta G^{0,\text{es}}_{\text{solv}} \tag{1}$$

In many models, the nonpolar free energy $\Delta G^{0,\text{np}}_{\text{solv}}$ is proportional to the molecular surface area, although recently more sophisticated models have been developed and parametrized (see, for example, refs 35 and 36). The electrostatic component $\Delta G^{0,\text{es}}_{\text{solv}}$ is often estimated using a macroscopic continuum electrostatic model,[34–37] shown in Figure 2.

The molecule-solvent boundary, denoted as Γ in Figure 2, is taken to be the Richards molecular surface[38] and separates the molecular interior, region *I*, from the solvent exterior, region *II*. The interior is modeled as a homogeneous dielectric with low dielectric constant $\varepsilon_I$ and a charge distribution $\rho(r)$; in this work, we assume that the charge distribution consists of $n_c$ discrete point charges, the *i*th of which is located at $r_i$ and has charge $q_i$. For many biomolecules, $n_c$ ranges from a few dozen to several thousand. The electrostatic potential in region *I*, $\varphi_I(r)$, satisfies a Poisson equation

$$\nabla^2 \varphi_I(r) = -\sum_{i=1}^{n_c} \frac{q_i}{\varepsilon_I} \delta(r - r_i) \tag{2}$$

The solvent region is modeled as a homogeneous dielectric with high dielectric constant $\varepsilon_{II}$; in this region, the electrostatic potential $\varphi_{II}(r)$ satisfies the Laplace equation

$$\nabla^2\varphi_{II}(r) = 0 \tag{3}$$

for nonionic solutions, or for dilute ionic solutions, the linearized Poisson–Boltzmann equation (LPBE)

$$\nabla^2\varphi_{II}(r) = \kappa^2\varphi_{II}(r) \tag{4}$$

where $\kappa$ is the inverse Debye length. The continuity of the potential and normal displacement furnishes boundary conditions for both regions.[39] In the remainder of this paper, we assume that $\kappa = 0$, noting that the PDE-constrained optimization techniques apply equally well when the LPBE is used to model the potential in the solvent.

This set of coupled partial differential equations (PDEs) cannot be solved analytically except for relatively simple geometries. For general problems and realistic treatments of molecular geometries, numerical methods such as the finite-difference, finite-element, or boundary-element methods must be employed.[39–61] In contrast to the finite-difference and finite-element methods, which discretize the differential form of the PDE, boundary-element methods discretize boundary-integral-equation formulations of the PDE problem.[39,40,42,43,47,53–55,59,62–66] Boundary-integral formulations possess attractive theoretical and numerical properties such as reduced dimensionality, the possibility of exact treatment of the dielectric boundary, and exact treatment of point charge effects and boundary conditions at infinity.[39] Many integral-equation approaches have been described in the literature;[39,40,42,47,55,62,67–69] in this paper, we present only the polarizable continuum model (PCM) formulation introduced by Miertus et al.[40,51,70,71] (which was independently derived by Shaw and Zauhar, who also called it the apparent-surface-charge (ASC) formulation[42,43,45]) and the formulation introduced by Yoon and Lenhoff.[47]

Using the electrostatic model in Figure 2, the computational challenge associated with calculating $\Delta G_{\text{solv}}^{0,\text{es}}$, the electrostatic contribution to a solute's solvation free energy, is evaluating the *reaction potentials* at each of the $n_c$ charge locations that is induced by solvent polarization in response to the charges themselves. Because we have assumed linear response, the vector of reaction potentials at the charge locations, $\varphi_R$, can be written as a weighted combination of the responses due to each of the individual point charges

$$\varphi_R = Sq \tag{5}$$

where we have defined $S$ to be the *reaction-potential* or *solvation* matrix. The electrostatic free energy is then a quadratic function of $q$

$$\Delta G_{\text{solv}}^{0,\text{es}} = \frac{1}{2}q^TSq \tag{6}$$

We now derive expressions for the reaction-potential matrix $S$ such that it may be written as

$$S = M_3M_2^{-1}M_1 \tag{7}$$

where $M_3$, $M_2^{-1}$, and $M_1$ are linear operators.

*2.1.1. The Apparent-Surface-Charge Formulation.* Numerous groups have derived the boundary-integral equation for the surface charge that develops at a dielectric boundary in response to a distribution of charge;[40,42,72,73] it is known variously as the polarizable-continuum model (PCM) and apparent-surface-charge (ASC) formulation[40,42,51,53] and has been widely used in biomolecular simulations.[43,45,53,63,68,74] Rather than solving for the potential throughout space in the original mixed-dielectric PDE problem, one solves an equivalent problem with uniform dielectric constant $\varepsilon_I$ everywhere, finding a distribution of charge $\sigma_p(r)$ on $\Gamma$ such that $\sigma_p(r)$ reproduces the continuity conditions of the original mixed-dielectric problem. This surface charge satisfies the second-kind integral equation[42]

$$\frac{\varepsilon_I + \varepsilon_{II}}{2\varepsilon_I(\varepsilon_I - \varepsilon_{II})}\sigma_p(r) + ⨍_{\Gamma}\frac{\partial}{\partial n(r)}\frac{\sigma_p(r')dA'}{4\pi\varepsilon_I\|r - r'\|} = -\frac{\partial}{\partial n(r)}\sum_{i=1}^{n_c}\frac{q_i}{4\pi\varepsilon_I\|r - r'\|'} \tag{8}$$

where ⨍ denotes a principal-value integral,[75,76] and $n(r)$ denotes the outward normal direction into solvent. The surface charge distribution $\sigma_p(r)$ produces in the molecular interior (region I) a potential equal to that induced by the polarization of the solute. The *reaction potential* at a solute charge location $r_i$ is the result of convolving the free-space Green's function with the surface charge distribution

$$\varphi_R(r_i) = \int_{\Gamma}\frac{\sigma_p(r')}{4\pi\varepsilon_I\|r_i - r'\|}dA' \tag{9}$$

The set of reaction potentials at all the charge locations is therefore the image of the charge distribution under three linear operators

$$\varphi_R = M_3M_2^{-1}M_1q \tag{10}$$

The operator $M_1$ maps the solute charge distribution to the induced normal-displacement field at the dielectric boundary; that is, the application of $M_1$ to $q$ generates the right-hand side (RHS) in eq 8.

The operator $M_2$ generates the left-hand side in eq 8 when applied to $\sigma_p$, and $M_2^{-1}$ is used to denote the operator's inverse. That is, $M_2^{-1}$ applied to the RHS in eq 8 generates $\sigma_p(r)$. Finally, the integral operator $M_3$ maps the induced surface charge to the reaction potentials at the charge locations via eq 9. Note that $M_3M_2^{-1}M_1$ is an $n_c$-by-$n_c$ matrix, even though $M_1$, $M_2$, and $M_3$ are operators.

Because the charge distribution is a set of discrete point charges, the difference in electrostatic free energy between the uniform $\varepsilon_I$ domain and the mixed-dielectric problem is a finite-dimensional inner product

$$\Delta G_{\text{solv}}^{0,\text{es}} = \frac{1}{2}\varphi_R^Tq \tag{11}$$

where $\varphi_R$ denotes the vector of reaction potentials computed at the $n_c$ charge locations. For problems in which $\varepsilon_{II} > \varepsilon_I$, $S = M_3M_2^{-1}M_1$ is symmetric negative definite.

*2.1.2. The Nonderivative Green's-Theorem Formulation.* Using Green's theorem, Yoon and Lenhoff derived a pair of coupled integral equations capable of modeling solutes in dilute ionic solutions (that is, when the LPBE holds in the solvent).[47] The integral equations are

$$\frac{1}{2}\varphi(r) + \fint_\Gamma \varphi(r')\frac{\partial G_I}{\partial n}(r,r')dA' - \fint_\Gamma \frac{\partial \varphi}{\partial n}(r')G_I(r,r')dA' = \sum_{i=1}^{n_c}\frac{q_i}{\varepsilon_I}G_I(r,r') \quad (12)$$

$$\frac{1}{2}\varphi(r) + \fint_\Gamma \varphi(r')\frac{\partial G_{II}}{\partial n}(r,r')dA' + \frac{\varepsilon_I}{\varepsilon_{II}}\fint_\Gamma \frac{\partial \varphi}{\partial n}(r')G_{II}(r,r')dA' = 0 \quad (13)$$

where $G_I(r; r')$ and $G_{II}(r; r')$ are the free-space Green's functions in the solute and solvent regions, and the unknown surface variables $\varphi(r)$ and $(\partial\varphi)/(\partial n)(r)$ are the surface potential and its normal derivative. These equations are derived by applying Green's theorem in regions *I* and *II*, finding the potential at arbitrary points in these regions by substituting the relevant Green's functions, and then letting the points approach the surface by taking appropriate limits.[47,64,66] After solving eqs 12 and 13 for $\varphi$ and $(\partial\varphi)/(\partial n)$, the reaction potential at the *i*th charge location induced by solvent polarization can be written as

$$\varphi_R(r_i) = \int_\Gamma \left[G_I(r_i;r')\frac{\partial\varphi}{\partial n}(r') - \varphi(r')\frac{\partial G_I}{\partial n}(r_i;r')\right]dA' \quad (14)$$

and again the electrostatic solvation free energy can be written as a product of three linear operators as in eq 7.

**2.2. Numerical Solution of the Integral Equations Using Boundary-Element Methods and Fast Algorithms.** The boundary-element method (BEM) is a popular technique for solving boundary-integral equations numerically. To solve an integral equation such as eq 8 using the BEM, one first introduces a set of basis functions defined on the surface. Representing the unknown surface variable as a weighted combination of the basis functions reduces the exact infinite-dimensional problem to an approximation problem with a finite number of unknowns, the weights used to scale the basis functions. A set of constraints on the weights is then written to force the approximate representation of the surface variable to satisfy the discretized integral equation as closely as possible in some metric (see, for example, ref 75). The resulting problem—that of finding the basis function weights that minimize some function of the residual—is a finite-dimensional matrix equation.

Usually, it is convenient to discretize the surface into a set of surface patches, or *boundary elements*, before defining the basis functions. In biomolecule electrostatic simulations, these elements are commonly planar triangles,[47,63] although curved-element discretizations of molecule-solvent interfaces have been described by several groups.[39,56,59,61,77,78] We present a boundary-element method for solving the ASC formulation. The Green's-theorem formulation (eqs 12 and 13) can be solved analogously but requires two weights for each basis function: one for the potential and one for its normal derivative. Full details for solving the Green's-theorem formulation numerically can be found in refs 47, 64, and 66.

First, the molecule-solvent interface is discretized using $n_p$ boundary elements, and then a set of $n_p$ piecewise-constant basis functions is defined such that

$$\chi_i(r) = \begin{cases} 1 & \text{if } r \text{ is on panel } i \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

The unknown surface charge density $\sigma_p(r)$ is then represented approximately as

$$\sigma_p(r) \approx \sum_{i=1}^{n_p} x_i\chi_i(r) \quad (16)$$

where the weights $x_i$ are unknown. Using a Galerkin discretization[75] of the PCM/ASC formulation in which the inner integral is evaluated via one-point quadrature,[68,79,80] one obtains the dense linear system $M_2x = M_1q$, with the entries of $M_2$ and $M_1$ given by

$$M_{2,ii} = \frac{\hat{\varepsilon}}{2\varepsilon_I}\alpha_i + \fint_{\text{panel } i}\frac{\partial}{\partial n(r)}\frac{\alpha_i dA'}{4\pi\varepsilon_I||r - r_{c_i}||} \quad (17)$$

$$M_{2,ij} = \int_{\text{panel } i}\frac{\partial}{\partial n(r)}\frac{\alpha_j dA'}{4\pi\varepsilon_I||r - r_{c_j}||}(i \neq j) \quad (18)$$

$$M_{1,ij} = -\int_{\text{panel } i}\frac{\partial}{\partial n(r)}\frac{q_j dA'}{4\pi\varepsilon_I||r - r_j||'} \quad (19)$$

where $\alpha_i$ denotes the area of panel *i*, $\hat{\varepsilon} = (\varepsilon_I + \varepsilon_{II})/(\varepsilon_I - \varepsilon_{II})$, $n(r)$ denotes the outward normal at *r*, and $r_{c_i}$ denotes the centroid of panel *i*. The approach presented here differs slightly from the commonly used centroid-collocation method, which essentially approximates the outer Galerkin integral using one-point quadrature; the method described here offers superior accuracy.[68,79] We note that the matrix entries of eqs 17, 18, and 19 are specific to the PCM/ASC formulation; if the Green's theorem formulation[47,66] or other boundary-integral formulations are employed to define the solvation matrix, similar matrices are defined that play analogous roles.[80]

Protein-sized systems often require more than $10^5$ unknowns and boundary elements to accurately represent the molecule–solvent interfaces and surface variables. Because solving the *n*-dimensional dense matrix equation $M_2x = M_1q$ using LU factorization requires $O(n^3)$ time, and even storing $M_2$ requires prohibitively large $O(n^2)$ memory, more efficient methods have been developed whose time and memory requirements scale linearly or near-linearly in the number of unknowns.[54,63,65,81,82] These fast-solver approaches combine Krylov-subspace iterative methods[83] such as GMRES[84] with fast, approximate algorithms to apply the discretized integral operator matrix to a vector. At the *k*th iteration of a Krylov-subspace algorithm, one finds an approximate solution $x^{(k)}$ that lies in the *k*th Krylov subspace, which is formed by repeatedly applying *A* to *b*

$$x^{(k)} \in \{b, Ab, A^2b, ..., A^{k-1}b\} \quad (20)$$

The fast multipole method[81,85] and the precorrected-FFT algorithm[86] represent two algorithms that compute BEM matrix-vector products in linear or near-linear time. The results in this paper were computed using the FFTSVD algorithm, which offers several advantages for biomolecule electrostatic problems.[65,66]

Krylov-subspace iterative methods are commonly preconditioned so that instead of solving $Ax = b$ for $x$, one solves $PAx = Pb$, where $P$ is a matrix that approximates $A^{-1}$ such that the iterates $x^{(k)}$ converge more rapidly toward the exact solution; that is, the use of a preconditioner reduces the number of matrix-vector products required to find a suitably accurate approximation to the actual solution $x$. Preconditioning the ASC formulation is easily accomplished using a diagonal matrix in which $P_{ii} = M_{2,ii}^{-1}$. Methods for preconditioning the Green's-theorem formulation are presented in refs 64 and 66.

**2.3. Biomolecule Electrostatic Optimization.** Using the thermodynamic cycle in Figure 1, the total electrostatic contribution to the binding free energy can be written as

$$\Delta G_{\text{bind}}^{0,\text{es}} = (-\Delta G_{\text{solv}}^{0,R,\text{es}} - \Delta G_{\text{solv}}^{0,L,\text{es}}) + \Delta G_{\text{bind}}^{0,\text{ref,es}} + \Delta G_{\text{solv}}^{0,L:R,\text{es}} \tag{21}$$

where the solvation free energies for the ligand, receptor, and complex are denoted by the superscripts $L$, $R$, and $L : R$, and the ligand−receptor Coulomb-interaction energy is written $\Delta G_{\text{bind}}^{0,\text{ ref, es}}$.[1] Substituting the appropriate ligand, receptor, and ligand−receptor reaction potential matrices, one obtains

$$\Delta G_{\text{bind}}^{0,\text{es}} = -\frac{1}{2}q_R^T R_{\text{unbound}} q_R - \frac{1}{2}q_L^T L_{\text{unbound}} q_L + (Gq_R)^T q_L + \frac{1}{2}q_C^T C_{\text{bound}} q_C \tag{22}$$

where $q_L$ and $q_R$ denote the $n_{c_L}$- and $n_{c_R}$-length vectors of ligand and receptor charge values, $q_C = (q_L, q_R)^T$, and $L_{\text{unbound}}$, $R_{\text{unbound}}$, and $C_{\text{bound}}$ denote the appropriate solvation matrices; the electrostatic component of the low-dielectric binding free energy has been written $(Gq_R)^T q_L$, where the $n_{c_L}$-by-$n_{c_R}$ Coulomb matrix $G$ maps receptor-charge values to Coulomb potentials at the ligand-charge locations given the bound-state geometry.

The optimizable component of $\Delta G_{\text{bind}}^{0,\text{ es}}$, which is the portion of $\Delta G_{\text{bind}}^{0,\text{ es}}$ that is dependent on the ligand charges, is called the variational electrostatic binding free energy $\Delta G_{\text{bind}}^{0,\text{ var.}}$.[6] The first term in eq 22 does not contribute to $\Delta G_{\text{bind}}^{0,\text{ var.}}$, nor does the component of the final term that depends only on the receptor charges. Writing $^1/_2(q_C^T C_{\text{bound}} q_C)$ as

$$\Delta G_{\text{solv},L-R}^{0,\text{es}} = \frac{1}{2}[q_L^T \quad q_R^T]\begin{bmatrix} L_{\text{bound}} & C_{\text{bound}}^{L,R} \\ C_{\text{bound}}^{R,L} & R_{\text{bound}} \end{bmatrix}\begin{bmatrix} q_L \\ q_R \end{bmatrix} \tag{23}$$

and exploiting the symmetry of $C_{\text{bound}}$ allows the variational electrostatic binding free energy to be written as

$$\Delta G_{\text{bind}}^{0,\text{var}} = -\frac{1}{2}q_L^T L_{\text{unbound}} q_L + \frac{1}{2}q_L^T L_{\text{bound}} q_L + q_R^T G^T q_L + q_R^T C_{\text{bound}}^{L,R} q_L \tag{24}$$

The final two terms in eq 24 are linear in the ligand charge values, and the vector

$$c = Gq_R + C_{\text{bound}}^{L,R} q_R \tag{25}$$

which represents the total field induced by the receptor charges at the ligand-charge locations in the bound state, may be used to further simplify eq 24

$$\Delta G_{\text{bind}}^{0,\text{var}} = \frac{1}{2}q_L^T(L_{\text{bound}} - L_{\text{unbound}})q_L + c^T q_L \tag{26}$$

Eq 26 is the objective function for optimizing the electrostatic component of the free energy of binding. Kangas and Tidor showed that the difference $L_{\text{bound}} - L_{\text{unbound}}$, which is the Hessian of the objective function, is positive definite if one assumes that the ligand binds rigidly, that the ligand charge distribution is unchanged on binding, and that the molecules have finite size.[6] The variational electrostatic binding free energy $\Delta G_{\text{bind}}^{0,\text{ var}}$ is therefore a convex function with respect to the ligand charge distribution, and there exists a unique minimal free energy.

Often, it is of interest to enforce sum-of-charge constraints over subsets of the charges and possibly over the entire set.[1,3,5] Defining the matrix $H = L_{\text{bound}} - L_{\text{unbound}}$ and including the linear constraint $Aq = b$ gives rise to the constrained optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}q^T H q + c^T q \\ \text{subject to} & Aq = b \end{array} \tag{27}$$

In eq 27 and for the remainder of the paper the vector $q$ is used instead of $q_L$ to represent the ligand charges. In addition to linear equality constraints, linear inequality constraints are sometimes imposed on the variables to ensure that the computed charges are physically reasonable. The resulting inequality-constrained problem

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}q^T H q + c^T q \\ \text{subject to} & Aq = b \\ \text{and} & m_i \le q_i \le M_i, \forall i \in \{1, ..., n_c\} \end{array} \tag{28}$$

where $m_i$ and $M_i$ represent the lower and upper bounds for $q_i$. Assuming without loss of generality that $A$ has full rank, this problem can be transformed into the standard form for a convex quadratic problem

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}x^T \hat{H} x + \hat{c}^T x \\ \text{subject to} & \hat{A}x = \hat{b} \\ \text{and} & x \ge 0 \end{array} \tag{29}$$

using the substitutions

$$
\begin{aligned}
x &= \begin{bmatrix} t \\ r \end{bmatrix} \\
\hat{H} &= \begin{bmatrix} H & 0 \\ 0 & 0 \end{bmatrix} \\
\hat{c} &= \begin{bmatrix} c + Hm \\ 0 \end{bmatrix} \\
\hat{A} &= \begin{bmatrix} A & 0 \\ I & I \end{bmatrix} \\
\hat{b} &= \begin{bmatrix} b - Am \\ M - m \end{bmatrix}
\end{aligned}
\tag{30}
$$

where the slack variables $t$ and $r$ satisfy $m + t = q$ and $q + r = M$. This notation for inequality-constrained biomolecule optimization problems will be used throughout the rest of the paper.

Titratable chemical groups in the ligand warrant a brief discussion. One of the assumptions inherent in the electrostatic optimization theory is that the ligand geometry and charge distribution do not change on binding.[1] Thus, for the charge optimization scheme presented here, the number of charges $n_c$ remains unchanged during optimization, as do their locations, and therefore charge optimization is performed for a particular titration state. In reality, of course, the ligand−receptor binding event may perturb the ligand geometry, its charge distribution, or state of protonation, or any combination of these. Exploring the dependence of the optimal charges (and the optimized binding free energy) on the titration state is the most obvious way to assess the impact of the assumptions underlying the theory, although in general such an undertaking is likely to be computationally expensive.

### 2.4. Solving Convex Quadratic Optimization Problems.

This section presents methods for minimizing the quadratic function

$$
f(x) = \frac{1}{2}x^T H x + c^T x \tag{31}
$$

where $x$ is a vector of length $n_{\text{primal}}$, and the matrix $H$, also known as the *Hessian* matrix, is symmetric and positive definite (SPD). The global minimizer $x^*$ can be found by setting $\nabla f(x) = 0$ and solving the resulting linear system

$$
Hx^* = -c \tag{32}
$$

Optimization problems with constraints require more sophisticated approaches (see, for example, refs 87 and 88). The quadratic program

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2}x^T H x + c^T x \\
\text{subject to} \quad & Ax = b
\end{aligned}
\tag{33}
$$

can be solved using Lagrange multipliers.[87] The optimal solution is a point $x^*$ and a corresponding vector of multipliers $\lambda^*$ that together satisfy the matrix equation

$$
\begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} -c \\ b \end{bmatrix} \tag{34}
$$

Inequality-constrained problems require the introduction of a vector of slack variables $s$ in addition to the Lagrange multipliers $\lambda$. Because the optimization problem in eq 29 satisfies a constraint qualification,[87] an optimal solution ($x^*$, $\lambda^*$, $s^*$) can be calculated by finding a point that satisfies the Karush-Kuhn-Tucker (KKT) optimality conditions

$$
\begin{aligned}
s^* &= Hx^* + c - A^T \lambda^* \\
Ay^* &= b \\
x_i^* s_i^* &= 0, \ \forall i \in \{1, 2, ..., n_{\text{primal}}\} \\
(x^*, s^*) &\geq 0
\end{aligned}
\tag{35}
$$

These conditions can be interpreted as the zeros of the nonlinear function

$$
F(x, \lambda, s) = \begin{bmatrix} Hx + c - A^T\lambda - s \\ b - Ax \\ Xs \end{bmatrix} \tag{36}
$$

where $X$ represents the diagonal matrix with $X_{ii} = x_i$. Primal-dual interior point methods find the roots of this equation using a modified Newton−Raphson iteration, with the Newton−Raphson updates biased to ensure convergence and scaled to ensure positivity of the elements of $x$ and $s$.[88] The $k$th update of a primal-dual iterative method satisfies the linear system of equations

$$
\begin{bmatrix} H & -A^T & -I \\ A & 0 & 0 \\ S^k & 0 & X^k \end{bmatrix} \begin{bmatrix} \Delta x^{k+1} \\ \Delta \lambda^{k+1} \\ \Delta S^{k+1} \end{bmatrix} = \begin{bmatrix} -c + s^k - Hx^k + A^T \lambda^k \\ b - Ax^k \\ X^k S^k e \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \sigma \dfrac{x^{k,T} s^k}{n_{primal}} e \end{bmatrix} \tag{37}
$$

where $e$ is the vector of ones, $S^k$ is diagonal with $S_{ii}^k = s_i^k$, and the second term on the right-hand side (RHS) biases the update toward a point with equal pairwise products $x_i s_i$.[88] The parameter $\sigma$, which is between 0 and 1, determines the strength of the bias. It can be a fixed value over all iterations or set dynamically based on the progress of the previous iterations.[89] Smaller values for $\sigma$ allow faster convergence in most cases, but larger values offer superior robustness.[88]

The reverse-Schur optimization method is specialized to PDE-constrained problems in which the relationships between the decision variables $x$, the PDE state variables $y_I$, and the external state variables $y_E$ are affine. That is, the three vectors satisfy a matrix equality

$$
\begin{bmatrix} B & A & 0 \\ D & C & -I \end{bmatrix} \begin{bmatrix} x \\ y_I \\ y_E \end{bmatrix} = \begin{bmatrix} z_I \\ z_E \end{bmatrix} \tag{38}
$$

for some vectors $z_I$ and $z_E$; in the electrostatic optimization, $z_I = 0$ and $z_E = 0$.

## 3. The Reverse-Schur Method for Electrostatic Optimization

Some problems in computational science can be solved more efficiently using a Schur complement, such that one solves not a block linear system such as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} \quad (39)$$

but rather a smaller (or better-conditioned) system like

$$(D - CA^{-1}B)y = b - CA^{-1}a \quad (40)$$

The reverse-Schur co-optimization method uses the exactly opposite approach. An unconstrained quadratic program

$$\text{minimize } \frac{1}{2}x^T Hx + c^T x \quad (41)$$

in which the Hessian is of the form $H = M_3 M_2^{-1} M_1$, can be solved by setting the gradient equal to zero, leading to the linear system

$$M_3 M_2^{-1} M_1 x^* = -c \quad (42)$$

Eq 42 resembles eq 40 with $D = 0$ and, therefore, the Hessian may be said to have the structure of a Schur complement. In reverse-Schur co-optimization, one solves the larger "reverse-Schur-complement" system

$$\begin{bmatrix} 0 & M_3 \\ -M_1 & M_2 \end{bmatrix}\begin{bmatrix} x^* \\ y \end{bmatrix} = \begin{bmatrix} -c \\ 0 \end{bmatrix} \quad (43)$$

As discussed in Section 2, the Hessian matrix $H = L_{\text{bound}} - L_{\text{unbound}}$ is a difference of two matrices, each of which has Schur structure. The reverse-Schur form of the unconstrained electrostatic optimization problem therefore has two reverse-Schur complements, and the optimizing ligand distribution $q^*$ can be found by solving

$$\begin{bmatrix} 0 & M_{3,b} & -M_{3,u} \\ -M_{1,b} & M_{2,b} & 0 \\ -M_{1,u} & 0 & M_{2,u} \end{bmatrix}\begin{bmatrix} q^* \\ y_b \\ y_u \end{bmatrix} = \begin{bmatrix} -c \\ 0 \\ 0 \end{bmatrix} \quad (44)$$

where the subscripts $b$ and $u$ denote the bound and unbound systems and the variables $y_b$ and $y_u$ are the surface variables for the corresponding BEM problems when the ligand charge distribution is the optimizing distribution $q^*$. In the ASC formulation, for instance, $y_b$ represents the weights for the bound-state basis functions. The bound- and unbound-geometry state variables are therefore found simultaneously with the optimal decision variables $q^*$. The co-optimization approach in this respect resembles "all-at-once" methods; however, the corresponding state variables are not explicitly included in the optimization problem. Section 3.1.2 details that approach to Hessian-implicit optimization. It is important to note that the matrices associated with reverse-Schur co-optimization are not necessarily symmetric. Also, we note that the bound-state and unbound-state surface are different, with the bound state representing the larger ligand−receptor complex; consequently, the vector $y_b$ is typically a longer vector than $y_u$.

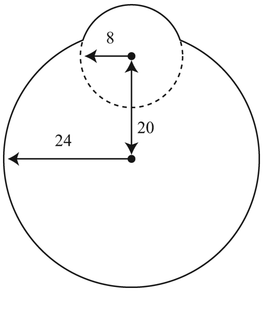


***Figure 3.*** Schematic of a model ligand−receptor complex for studying implementation details for the co-optimization method. The surface of the ligand−receptor complex is defined by rolling a probe sphere of radius 1.4 Å over the union of the two spheres. All distances are in Angstroms.

As discussed in Section 2, electrostatic charge optimization problems usually have many fewer decision variables than there are degrees of freedom associated with the BEM problems. Reverse-Schur co-optimization systems such as eq 44 are therefore only slightly larger than the corresponding BEM systems. Efficient preconditioning strategies, which will be described in Section 4, allow systems such as eq 44 to be solved with approximately the same computational cost as would be required to solve a single bound-state and a single unbound-state electrostatic problem. We emphasize that the dense boundary-element matrices $M_{2,b}$ and $M_{2,u}$ are almost always too large to be calculated or stored, and so their LU factorizations cannot be computed. Forming $H$ in eq 41 explicitly requires $n_c$ solves of the bound and unbound geometries. The reverse-Schur method is therefore more computationally efficient than the explicit-Hessian approach.

The next section presents alternative Hessian-free PDE-constrained methods, such as a nested-Krylov method, or NAND-like approach, and a traditional PDE-constrained formulation, following the SAND paradigm. Section 3.2 describes biomolecule co-optimization techniques for constrained problems.

**3.1. Alternative PDE-Constrained Approaches.** *3.1.1. Nested-Krylov Approach.* One alternative to the co-optimization approach would be to use nested Krylov methods to solve the linear systems associated with the explicit-Hessian approach. The unconstrained problem

$$(M_{3,b}M_{2,b}^{-1}M_{1,b} - M_{3,u}M_{2,u}^{-1}M_{1,u})q^* = -c \quad (45)$$

would then require two inner Krylov solves for every matrix-vector multiplication required for the outer Krylov method: one for the bound-state problem and one for the unbound-state problem. This approach represents an implementation of a *nested analysis and design* (NAND) approach to PDE-constrained optimization.

*3.1.2. Incorporating the PDE as Constraints.* The electrostatic optimization problem can also be formulated in a traditional PDE-constrained approach in which the surface variables of the bound- and unbound-state boundary-element problems are included as optimization variables, with the boundary-element method equations added as equality constraints. The resulting problem

$$\text{minimize } \frac{1}{2}\begin{bmatrix} q \\ y_b \\ y_u \end{bmatrix}^T \begin{bmatrix} 0 & \frac{1}{2}M_{3,b} & -\frac{1}{2}M_{3,u} \\ \frac{1}{2}M_{3,b}^T & 0 & 0 \\ -\frac{1}{2}M_{3,u}^T & 0 & 0 \end{bmatrix} \begin{bmatrix} q \\ y_b \\ y_u \end{bmatrix} + \begin{bmatrix} c \\ 0 \\ 0 \end{bmatrix}^T \begin{bmatrix} q \\ y_b \\ y_u \end{bmatrix}$$
$$\text{subject to } \begin{bmatrix} -M_{1,b} & M_{2,b} & 0 \\ -M_{1,u} & 0 & M_{2,u} \end{bmatrix} \begin{bmatrix} q \\ y_b \\ y_u \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{46}$$

can be solved by setting the gradient to zero and solving the resulting linear system

$$\begin{bmatrix} 0 & -\frac{1}{4}M_{3.b} & -\frac{1}{4}M_{3.u} & -M_{1,b}^T & -M_{1,u}^T \\ \frac{1}{4}M_{3,b}^T & 0 & 0 & M_{2,b}^T & 0 \\ -\frac{1}{4}M_{3,u}^T & 0 & 0 & 0 & M_{2,u}^T \\ -M_{1.b} & M_{2.b} & 0 & 0 & 0 \\ -M_{1.u} & 0 & M_{2.u} & 0 & 0 \end{bmatrix} \begin{bmatrix} q^* \\ y_b^* \\ y_u^* \\ \lambda_b^* \\ \lambda_u^* \end{bmatrix} = \begin{bmatrix} -c \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{47}$$

The matrix in eq 47 is symmetric, which allows the system to be solved using specialized Krylov-subspace iterative methods. However, every matrix-vector product requires twice as much calculation as the matrix-vector products required for Krylov methods to solve eq 44. It can also be difficult or impractical to apply the transposed matrices $M_{2,b}^T$ and $M_{2,u}^T$. Thus, both the reverse-Schur method and the SAND-type approach have strengths and weaknesses.

**3.2. Constrained Co-Optimization.** It is straightforward to use the reverse-Schur method to solve the constrained optimization problems presented in Section 2.4. After transforming eq 34, the co-optimization system for solving the problem with linear-equality constraints is

$$\begin{bmatrix} 0 & A^T & M_{3,b} & -M_{3,u} \\ A & 0 & 0 & 0 \\ -M_{1,b} & 0 & M_{2,b} & 0 \\ -M_{1,u} & 0 & 0 & M_{2,u} \end{bmatrix} \begin{bmatrix} q^* \\ \lambda^* \\ y_b \\ y_u \end{bmatrix} = \begin{bmatrix} -c \\ b \\ 0 \\ 0 \end{bmatrix} \tag{48}$$

Similarly, the co-optimization system associated with the *k*th iteration of a primal-dual interior-point method is transformed from eq 37 to

$$\begin{bmatrix} 0 & -\hat{A}^T & -I & \hat{M}_{3,b} & -\hat{M}_{3,u} \\ \hat{A} & 0 & 0 & 0 & 0 \\ S^k & 0 & X^k & 0 & 0 \\ -\hat{M}_{1,b} & 0 & 0 & M_{2,b} & 0 \\ -\hat{M}_{1,u} & 0 & 0 & 0 & M_{2,u} \end{bmatrix} \begin{bmatrix} \Delta x^{k+1} \\ \Delta \lambda^{k+1} \\ \Delta s^{k+1} \\ \Delta y_b^{k+1} \\ \Delta y_u^{k+1} \end{bmatrix} = \begin{bmatrix} -\hat{c} - s^k - \hat{H}x^k + \hat{A}^T\lambda^k \\ \hat{b} - \hat{A}x^k \\ X^kS^ke + \sigma\frac{x^{k,T}s^k}{n_c}e \\ \hat{M}_{1,b}x^k - M_{2,b}y_b^k \\ \hat{M}_{1,u}x^k - M_{2,u}y_u^k \end{bmatrix} \tag{49}$$

where

$$\hat{M}_{1,b} = [M_{1,b} \quad 0] \tag{50}$$

$$\hat{M}_{3,b} = \begin{bmatrix} M_{3,b} \\ 0 \end{bmatrix} \tag{51}$$

with the zero submatrices of the appropriate size given the transformation of the inequality-constrained problem into standard form, and $\hat{M}_{1,u}$ and $\hat{M}_{3,u}$ are similarly defined. The term $\hat{H}x^k$ on the right-hand side is computed as

$$\hat{H}x^k = \hat{H}x^{k-1} + \hat{M}_{3,b}\Delta y_b^k - \hat{M}_{3,u}\Delta y_u^k \tag{52}$$

and $\hat{H}x^0$ must be found before the first iteration.

## 4. Implementation

In this section we present two implementation details that are important for the co-optimization method to achieve high efficiency and accuracy. First, regularization of the optimization presents a critical challenge. The need for regularization arises due to numerical error in simulation; although the exact Hessian is positive definite, typically a numerically computed $H$ is not. Given $H$ explicitly, the eigendecomposition or singular value decomposition can be used to penalize or eliminate the nonphysical part of the matrix. However, for the reverse-Schur (or any other implicit-Hessian) method to produce results comparable to those obtained by current methods, accurate regularization techniques must be found so that the appropriate directions can be penalized. Second, effective preconditioning schemes need to be developed because the co-optimization linear systems are solved using Krylov iterative methods.

A simple model geometry, shown in Figure 3, is used to demonstrate the performance of the presented regularization and preconditioning methods. The unbound ligand is a sphere of radius 8 Å; the ligand−receptor complex is modeled as the solvent-excluded surface produced by rolling a 1.4-Å probe sphere over the union of the ligand sphere and a 24-Å radius sphere representing the receptor, where the sphere centers are separated by 20 Å.[38] The internal dielectric constant is taken to be 4 and the solvent external dielectric constant to be 80. For simplicity in this paper, we assume that the Laplace equation holds in the solvent region (that is, that there are no mobile ions in solution). However, the implicit-Hessian methods can be used also for the case when the LPBE holds in the solvent region.

The receptor charge distribution consists of 2000 randomly placed charges located such that no charge is within 1.5 Å of another charge, the dielectric boundary, or the ligand volume; the charge values have been chosen from a uniform distribution on [−1*e*, 1*e*]. The ligand charge distribution is built from a set of charge locations randomly placed in the ligand sphere subject to the constraints that no charges are within 1.5 Å of one another or of the ligand boundary. Several discretizations of each geometry have been generated. One surface discretization uses planar triangle boundary elements generated using MSMS,[90] with 71922 and 7924 elements used to approximate the bound and unbound surfaces; the other uses 2132 and 1810 curved boundary elements that exactly represent the two geometries.[78] Coarser discretizations have been used for some examples and are described where appropriate. Full computational details are deferred to Section 5.

**4.1. Regularization.** The explicit-Hessian approach to biomolecule electrostatic optimization allows straightforward regularization. The numerically calculated Hessian matrix is first symmetrized to remove numerical-error-based asymmetries. Because the Hessian is available explicitly, it is possible to calculate its eigendecomposition. The eigenspace corresponding to the smallest eigenvectors is then heavily penalized but not removed explicitly, so that the existence of a feasible solution is assured even in the presence of other constraints. The co-optimization approach, in contrast, does not permit the use of direct eigendecomposition, because the Hessian matrix is not available. Co-optimization achieves its performance advantage by leaving the Hessian implicit, and therefore little information is available regarding its spectrum or corresponding eigenvectors. In order to solve the same optimization problems using the current protocols without sacrificing performance, the minimal eigenspace must be approximated inexpensively.

In ref 89, linear-equality constrained co-optimization problems were preconditioned using an approximate Hessian $\hat{H}$ of the form

$$\hat{H} = M_{3,b}P_{2,b}M_{1,b} - M_{3,u}P_{2,u}M_{1,u} \tag{53}$$

where $P_{2,b}$ and $P_{2,u}$ represent the bound- and unbound-state BEM preconditioners. Using a diagonal approximation of the ASC integral equation as the preconditioner $P_2$ corresponds to the recent BIBEE/P electrostatic model.[91] When performing co-optimization using the Green's theorem formulation, the operators $M_1$, $M_2$, $M_3$, and $P_2$ differ from those employed in the ASC formulation (i.e., the entries of $M_2$ are no longer defined according to eq 17). However, the Green's theorem co-optimization linear systems are written the same way, and an approximate Hessian can still be defined according to eq 53.

To illustrate that an approximation to the ASC formulation can be used to regularize the optimization method but not the Green's-theorem method approximation, explicit Hessians and their approximations were calculated using both integral formulations and both planar and curved boundary elements. The four Hessian matrices and their approximations were decomposed using the singular value decomposition (SVD), and the right singular vectors of the approximate Hessians were projected onto those from the corresponding explicitly calculated Hessians. The Green's-theorem and ASC formulations produced very similar explicit actual Hessians ($\|H_A - H_G\|/\|H_A\| < 0.01$), regardless of whether planar or curved boundary elements were used.

To ensure comparable regularization between explicit-Hessian and implicit-Hessian optimization procedures, implicit-Hessian methods must penalize not only the same number of search directions as explicit-Hessian methods but also the search directions themselves. To illustrate how the singular vectors of an approximate Hessian $\hat{H}$ are aligned with the singular vectors of the actual Hessian $H$, we calculate the matrix

$$V_H^T V_{\hat{H}} \tag{54}$$

which represents the projection of the singular vectors of $\hat{H}$ onto those of $H$. Perfect alignment between the sets of vectors would produce a diagonal matrix $X$ whose diagonal entries all have unit magnitude. Similarly, the degree to which the singular vectors of $\hat{H}$ are imperfectly aligned with those of $H$ is reflected in the presence of nonzero entries off the diagonal. Figure 4(a) is a pseudocolor plot of the magnitudes of the entries of $X$, using approximate and actual Hessians computed from the Green's-theorem formulation and the planar-element discretization. The analogous plot, computed using the ASC formulation and planar elements, is shown in Figure 4(b). The ASC-based approximate singular vectors are clearly much better aligned with the corresponding singular vectors of the explicitly computed ASC Hessian. Similarly, plotted in Figure 5(a),(b) are the results of curved-element simulations of the Green's-theorem and ASC formulations, respectively, with the approximate Hessian right singular vectors projected onto the explicit-Hessian right singular vectors. Because the ASC formulation generates superior approximate Hessians to the mixed formulation for both kinds of discretizations, we attribute the fidelity to the superior conditioning of purely second-kind integral operators (see, for instance, ref 56). In principle, it is possible to regularize the Green's-theorem co-optimization system using the penalty matrix derived from the ASC-based approximate Hessian. However, the results in Section 4.2 illustrate that the superior conditioning of the second-kind integral equation produces convergence of the co-optimization GMRES in many fewer iterations than are required for the Green's-theorem based co-optimization.

Figure 6(a),(b) contains plots of the singular values for the planar- and curved-element discretizations. Predictability of the relation between the approximate singular values $\hat{\sigma}_i$ and the actual values $\sigma_i$ is important so that the appropriate number of search directions can be penalized. The singular values of $\hat{H}_A$ were much closer to those of $H_A$ than the singular values of $\hat{H}_G$ were to $H_G$.

Based on the results in Figures 4, 5, and 6 we adopted the following scheme to regularize the co-optimization solutions. The ASC-based Hessian approximation $\hat{H}_A$ was computed first, and the eigendecomposition of the symmetrized matrix $^1/_2(\hat{H}_A + \hat{H}_A^T)$ was taken. The first right singular vector $\hat{v}_1$ was multiplied by the Hessian $H_A$ using BEM simulation of the bound- and unbound-states, and the Rayleigh quotient

$$\hat{\lambda}_1 = \hat{v}_1^T H_A \hat{v}_1 \tag{55}$$

was then used as an estimate for the maximum eigenvalue of $H_A$. The penalty matrix

$$W = \alpha V_{\{:,I\}} V_{\{:,I\}}^T \tag{56}$$

was then created, where the penalty paramater $\alpha = 100$ kcal/mol/$e^2$, the eigenvalue tolerance $\gamma = 10^{-4}$, and the set of penalized directions $I = \{i|\hat{\sigma}_i < \gamma\hat{\lambda}_1\}$. The quadratic penalty term $^1/_2 q^T W q$ was then added to the objective function, and optimization could begin. The unconstrained co-optimization system with a penalty matrix is

$$\begin{bmatrix} W & A_{3,b} & -A_{3,u} \\ -A_{1,b} & A_{2,b} & 0 \\ -A_{1,u} & 0 & A_{2,u} \end{bmatrix} \begin{bmatrix} q^* \\ \sigma_{p,b} \\ \sigma_{p,u} \end{bmatrix} = \begin{bmatrix} -c \\ 0 \\ 0 \end{bmatrix} \tag{57}$$

and the systems for constrained problems are similarly modified.

Even though $\hat{\lambda}_1$ usually approximated $\lambda_1$ to within a few percent, sometimes the number of directions penalized was slightly different from the number that would be penalized in an explicit-Hessian method using the same tolerance $\gamma$. As a result, it was desirable to have an inexpensive means of obtaining approximations to the optimal distributions for problems where different numbers of directions were penalized. For the unconstrained and linear-equality constrained problems, such estimates could be obtained using an approximation to the Sherman-Woodbury-Morrison formula

$$(H + UV^T)^{-1} = H^{-1} - H^{-1}U(I + V^T H^{-1} U)^{-1} V^T H^{-1} \tag{58}$$

which specifies how the inverse of a matrix $H$ changes when $H$ is perturbed by the low-rank update $UV^T$. Update-approximation methods for inequality-constrained problems represent an area of current research.

**4.2. Preconditioning.** The Hessian-implicit linear systems in eqs 44, 48, and 49 share a similar structure and therefore can be preconditioned by similar methods. For a problem with $n_{\text{primal}}$ decision variables and $\Sigma n_{SV}$ total unknowns associated with the BEM simulations, we define the desolvation operators

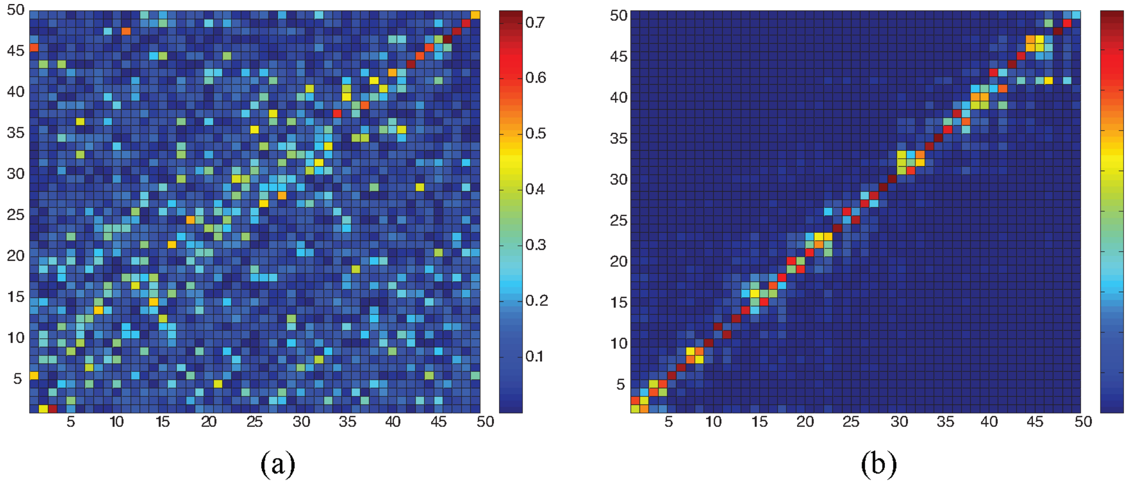


*Figure 4.* Comparison of the explicit and approximate Hessians of the sample problem in Figure 3 when discretized using planar boundary elements. The alignment between the singular vectors of exact and approximate Hessians is obtained by projecting the right singular vectors of an approximate Hessian $\hat{H}$ onto the right singular vectors of the explicit Hessian $H$ and taking the magnitude of the resulting entries. Each row and column therefore has 2-norm of one. (a) Explicit and approximate Hessians obtained using the Green's theorem formulation. (b) Explicit and approximate Hessians obtained using the polarizable continuum model/apparent surface charge formulation.

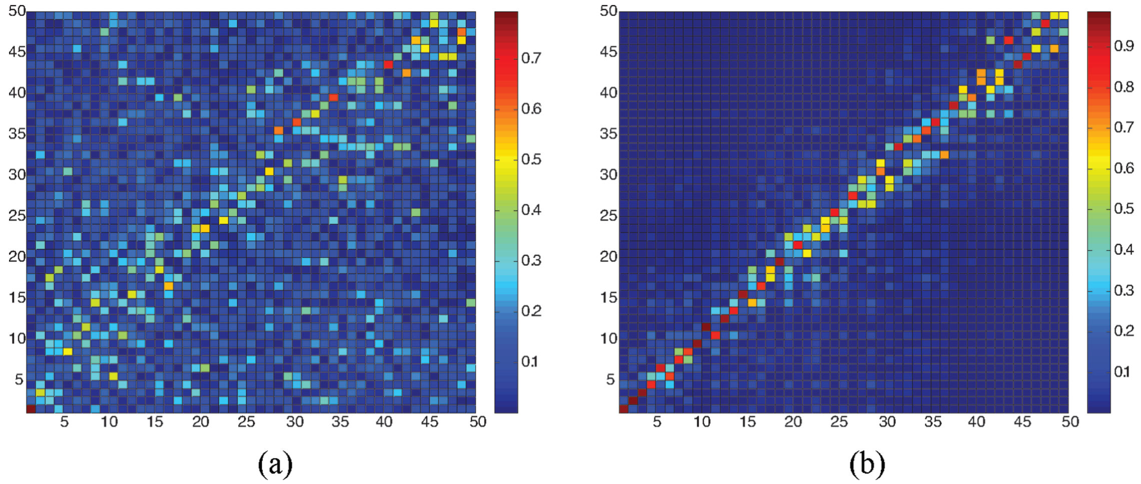


*Figure 5.* Comparison of the explicit and approximate Hessians of the sample problem in Figure 3 when discretized using curved boundary elements. The alignment between the singular vectors of exact and approximate Hessians is obtained by projecting the right singular vectors of an approximate Hessian $\hat{H}$ onto the right singular vectors of the explicit Hessian $H$ and taking the magnitude of the resulting entries. Each row and column therefore has 2-norm of one. (a) Explicit and approximate Hessians obtained using the Green's theorem formulation. (b) Explicit and approximate Hessians obtained using the polarizable continuum model/apparent surface charge formulation.

$$\hat{J}_1 = \begin{bmatrix} -M_{1,b} & 0 \\ -M_{1,u} & 0 \end{bmatrix} \tag{59}$$

$$\hat{J}_2 = \begin{bmatrix} M_{2,b} & 0 \\ 0 & M_{2,u} \end{bmatrix} \tag{60}$$

$$\hat{P}_2 = \begin{bmatrix} p_{2,b} & 0 \\ 0 & p_{2,u} \end{bmatrix} \tag{61}$$

$$\hat{J}_3 = \begin{bmatrix} M_{3,b} & -M_{3,u} \\ 0 & 0 \end{bmatrix} \tag{62}$$

where the zero blocks are sized such that $\hat{M}_1 \in \mathscr{R}^{\Sigma n_{SV} \times n_{\text{primal}}}$ and $\hat{M}_3 \in \mathscr{R}^{n_{\text{primal}} \times \Sigma n_{SV}}$.

We define the preconditioners by block-factorizing the corresponding linear systems using the BEM preconditioners rather than the inverses of the BEM matrices. The Hessian-implicit preconditioners would therefore be exact if the BEM preconditioners were actually the BEM matrix inverses. The resulting preconditioners can be written as the product $\tilde{P} = \tilde{P}_4\tilde{P}_3\tilde{P}_2\tilde{P}_1$. Other preconditioners, which for example block-triangularize the Hessian-implicit linear systems, can also be used but are generally less effective than the block factorization. Figure 7(a),(b) contains plots of the preconditioned relative GMRES residuals as a function of iteration count for the unconstrained problem using the Green's theorem formulation and the apparent surface-charge formulation. Each solves the same 50-charge unconstrained problem using curved-element discretizations of the model problem in Figure 3, using different preconditioners. It is clear that the ASC co-optimization converges in many fewer iterations than the corresponding Green's-theorem co-optimization, regardless of which approach to preconditioning is employed.

## 5. Computational Results

**5.1. Efficiency of Co-Optimization and PDE-Constrained Approaches.** The standard "all-at-once" approach and nested-Krylov approach to the electrostatic optimization problem were implemented in MATLAB using the geometry in Figure 3 and relatively coarse discretizations of 142 and 124 curved elements for the bound and unbound geometries. The methods' implementations were verified by direct inspection of the optimal charges computed by dense factorization of the systems in eqs 45 and 47. These techniques and the co-optimization solver were then used with preconditioned GMRES to solve a set of unconstrained problems of varying dimension. Computational expense was measured by counting the total number of applications of the BEM operator $\hat{M}_2$, because the computational cost of optimization is dominated by the application of the BEM operators. The all-at-once solver required two $\hat{M}_2$ matrix-vector products for every Krylov iteration.

The number of required applications for the nested-Krylov method was estimated using the fact that the nested-Krylov method, which relies on an implicit Hessian, and an explicit-Hessian Krylov method require the same number of GMRES iterations to achieve convergence. Therefore we estimated the nested-Krylov computational expense using the explicitly calculated Hessian, rather than a true nested-Krylov code. The GMRES solve was preconditioned using the ASC-based Hessian approximation $\hat{H}_A$. The all-at-once system was preconditioned using a block-factorization method similar to the co-optimization preconditioning schemes, so that the all-at-once preconditioner would be exact if the BEM preconditioner were exact. Every GMRES iteration for the all-at-once system requires two applications of the BEM operator $\hat{M}_2$, as shown in eq 47. Figure 8 is a plot of the computational cost of these methods for solving unconstrained problems as the number of optimization variables varies from 5 to 130. The PDE-constrained approaches scale very favorably compared to the explicit-Hessian approach and exhibit essentially comparable reductions in cost. It should be noted that the performance differences between the PDE-constrained approaches may reflect the relatively unoptimized implementations of the methods, and no significant conclusions should be drawn regarding the merits

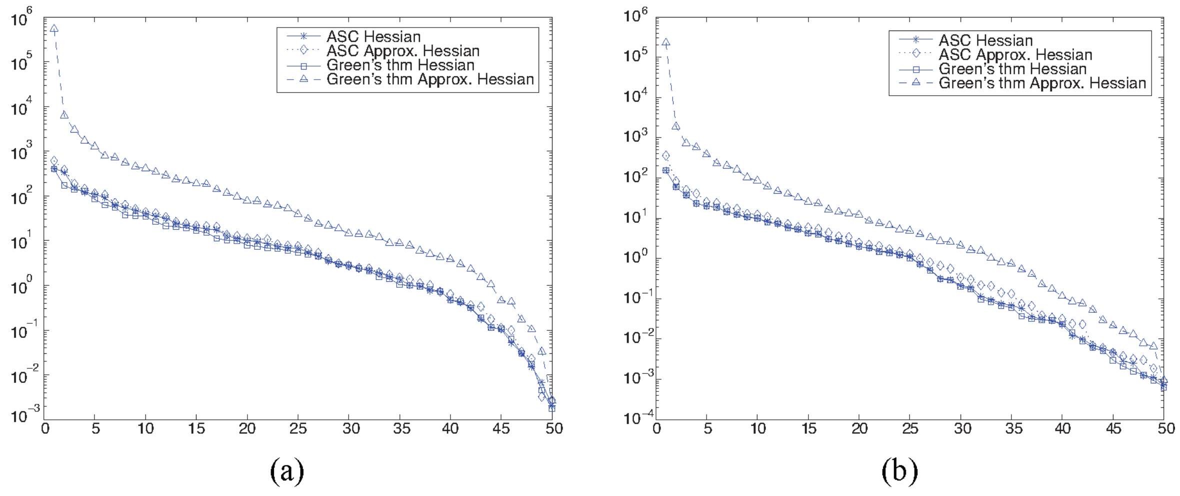


***Figure 6.*** The magnitudes of the singular values of the explicit and approximate Hessians computed using (a) planar boundary elements and (b) curved boundary elements. The singular values of the approximate Hessians calculated using the Green's-theorem formulation are less accurate than the singular values of the approximate Hessians computed using the apparent-surface-charge (ASC) formulation, regardless of whether planar or curved boundary elements are employed.

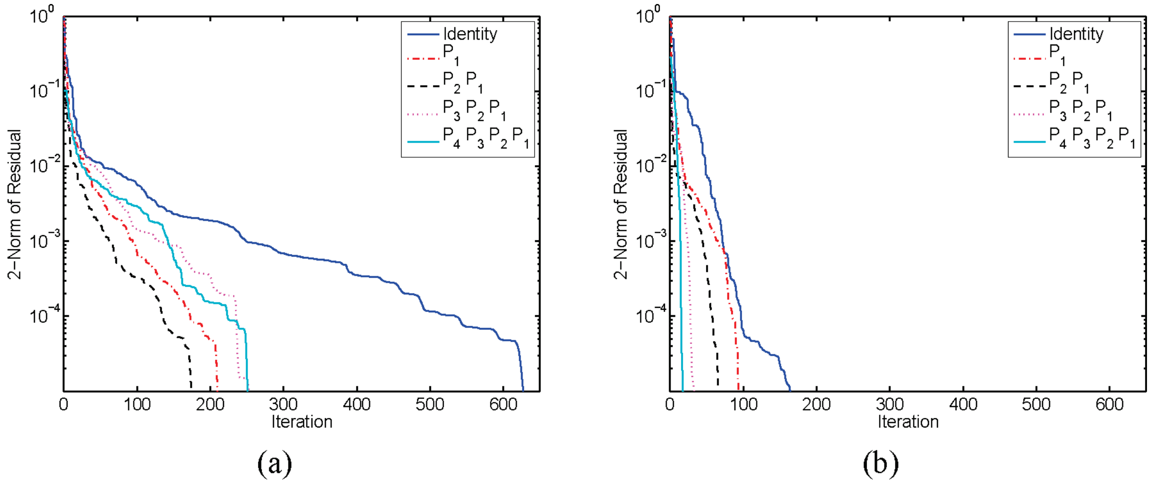


***Figure 7.*** Preconditioning effects on GMRES convergence for a 50-charge unconstrained optimization problem using (a) the nonderivative Green's-theorem formulation of Yoon and Lenhoff[47] and (b) the Shaw apparent-surface-charge formulation.[42]

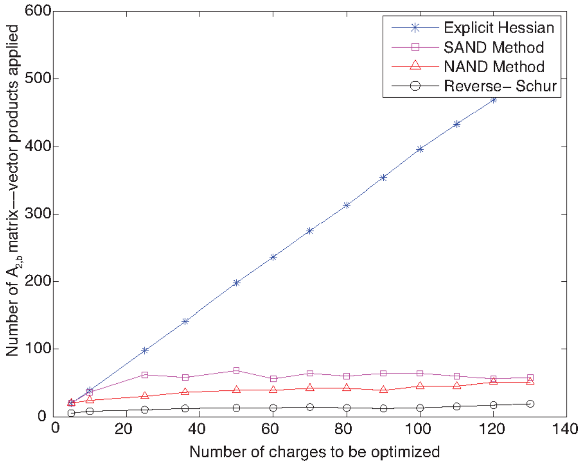


***Figure 8.*** The cost to solve unconstrained optimization problems of varying dimension using the co-optimization method, the two alternative implicit-Hessian approaches presented in Section 3.1, and by calculating the Hessian explicitly.

of one PDE-constrained technique over another. We note also that constrained problems exhibit similar performance trends.[89,92]

**5.2. Realistic Protein−Ligand Systems.** The three-dimensional structures of a complex between HIV-1 protease and darunavir (accession code 1T3R) and a complex between CDK2 and a small-molecule inhibitor (accession code 1OIT) were obtained from the Protein Data Bank (PDB). For both structures, protein side chains with missing density were rebuilt in their default geometry using the CHARMM computer program,[93] and in cases of multiple occupancy, the first entry listed was used. The final chi angles for asparagine, glutamine, and histidine side chains were flipped by 180 degrees as necessary to improve the hydrogen bonding network. Hydrogen atoms were added to both structures using the HBUILD module[94] of CHARMM and the PARAM22 parameter set,[95] using a distance-dependent dielectric constant of 4. Ionizable residues were left in their standard states at pH 7. In the case of HIV-1 protease, the catalytic dyad was left doubly deprotonated. The receptor protonation states are assumed to be the same in both the bound and unbound states. For electrostatic simulations, atomic radii were taken from the PARSE parameter set.[96] Partial atomic charges for protein atoms were also taken from the PARSE parameter set; quantum-mechanically derived partial atomic charges for the small-molecule inhibitors were calculated as follows. The geometry of each small-molecule inhibitor was optimized using quantum mechanical calculations at the RHF/6-31G* level of theory as implemented in the program Gaussian 98.[97] After geometry optimization, partial atomic charges were fit to the quantum mechanical electrostatic potential using the RESP methodology.[98,99] Both the CDK2 and HIV-protease systems were optimized using the co-optimization method and an explicit Hessian; the same curved-element discretizations were used for both methods. As in Section 4, solute−solvent interfaces were defined as solvent-excluded (molecular) surfaces using a probe of radius 1.4 Å, and the solute and solvent dielectric constants were 4 and 80, respectively.

*5.2.1. CDK2 and Inhibitor.* The CDK2 inhibitor described by Anderson et al. has 40 atoms.[31] Optimization of the partial atomic charges for a small-molecule inhibitor of CDK2 did not lead to significantly improved predicted electrostatic binding free energy; this inhibitor appears to already be very well optimized for its protein target. From finite-difference simulations, the total electrostatic contribution to binding for the quantum-mechanically derived (nominal) charge distribution, which has net zero charge, is 8.75 kcal/mol, and the optimal charge distribution leads to an electrostatic binding free energy of 5.54 kcal/mol.

Figure 9 is a plot of the inhibitor with RESP-derived nominal charge values (in red) and the co-optimization unconstrained optimal charge values (in blue) labeling each atom, computed using curved boundary elements. It can be seen that the inhibitor atoms that directly hydrogen bond to the protein, especially those in the aminopyrimidine core, have optimal partial atomic charges that closely match those determined through quantum mechanics.

The wild-type molecule has zero net charge, and the co-optimization optimal solution from Figure 10 has a net charge of −0.26*e*. For comparison, the explicit-Hessian boundary-element approach leads to an unconstrained optimum with

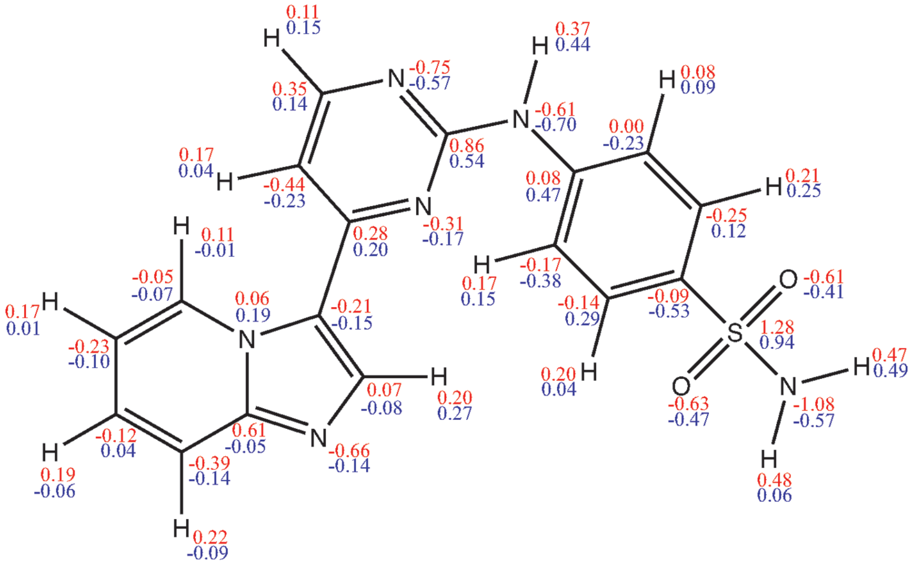


***Figure 9.*** The Anderson et al. inhibitor of CDK2.[31] The atoms are labeled with (red) partial atomic charge values derived from quantum mechanics via RESP[98,99] fitting and (blue) optimized partial atomic charges computed with unconstrained minimization using curved-element BEM and co-optimization.

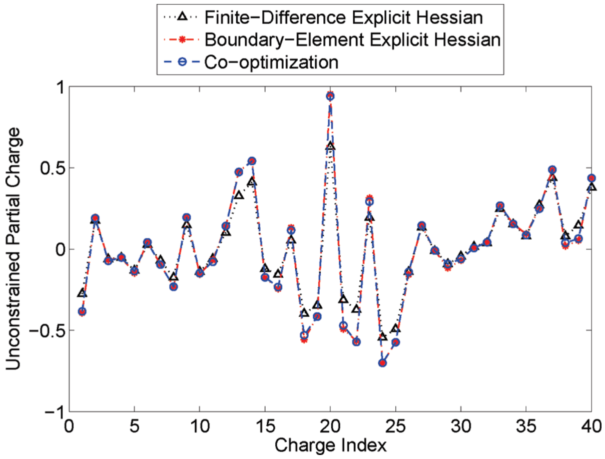


***Figure 10.*** The unconstrained optimal partial atomic charges computed using finite-difference and boundary-element explicit Hessians as well as using the reverse-Schur co-optimization method. The boundary-element simulations employed curved boundary element discretizations.

$-0.33e$ net charge, and the finite-difference optimal charges sum to $0.06e$; the preponderance of the difference is localized in a small number of atoms whose optimal charges are of large magnitude (Figure 10). As expected, optimal charges computed using the explicit Hessian and the co-optimization methods were nearly identical. Explicit-Hessian calculations were performed using both finite-difference methods[100] and boundary-element methods,[66] as was the boundary-element based co-optimization approach. The first instructive comparison, of the explicit-Hessian approaches, demonstrates that even the vastly different approaches to numerical simulation produce optimal charges that correspond very closely. The methods must give exactly the same results in the limit of infinitely fine discretizations, of course, but it is valuable to know that such good agreement can be obtained even for discretizations that can be easily solved on a personal workstation. The excellent agreement between the boundary-element methods demonstrates the correct implementation of the co-optimization approach and that numerical errors associated with the implicit representation of the dense boundary-integral operators do not materially affect the computed optimal solution.

The total charge on the inhibitor was also constrained to different net charge values. For these problems with different constraints, very little change was observed in the partial atomic charges for atoms directly interacting with the receptor, and the calculated optimal binding free energy (the objective function at the optimum) changed minimally. Figure 11 shows the box-constrained optimal charges when the total inhibitor charge was constrained to be −1, 0, or +1, computed using co-optimization. The co-optimization charges again correspond closely with calculations performed using either boundary-element or finite-difference methods with explicit Hessians (data not shown). The finite-difference calculations gave rise to optimized binding free energies of 6.88 kcal/mol for the $-1e$-constrained problem, 5.54 for the zero-charge problem, and 6.55 for the $+1e$-constrained problem.

The partial atomic charge values of the solvent-exposed sulfonamide group changed the most to accommodate these different net charges, largely because their solvent-exposed nature results in small desolvation penalties on binding. It should be noted that the $0.85e$ bound on the sulfur partial charge may be too stringent. These box constraints were introduced following earlier work[3] and the observation that few biomolecular systems are modeled as having partial charges of larger magnitude, regardless of whether the partial charges are taken from molecular mechanics force fields such as CHARMM[95] or derived from electronic structure calculations and RESP fitting. Our primary purpose here, however, is to demonstrate that the co-optimization method is fully capable of treating inequality constraints using a primal-dual interior point method. Also, the constraining of the total

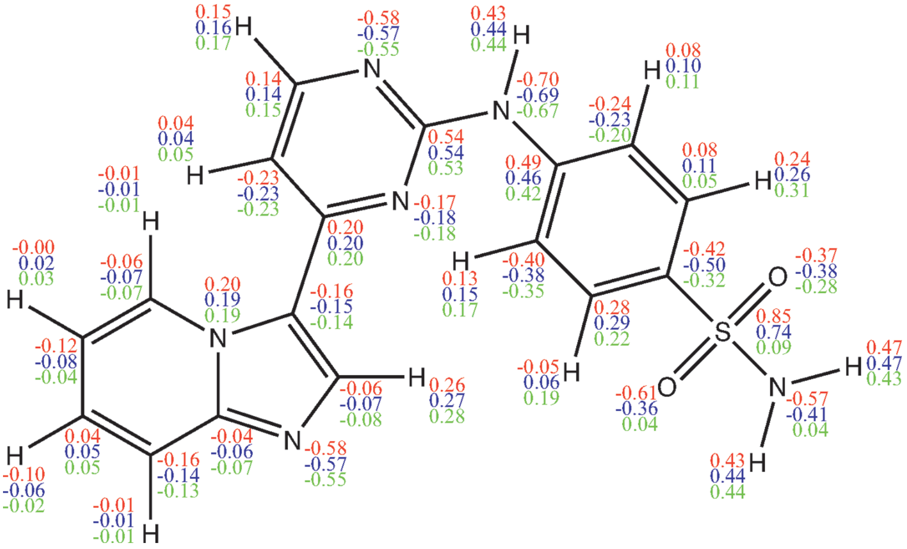


***Figure 11.*** The Anderson et al. inhibitor of CDK2.[31] Partial atomic charges have been optimized using box inequality constraints to enforce that charge values are less than 0.85*e* in magnitude, and sum-of-charge constraints have been imposed such that the total inhibitor charge is −1 (red label), 0 (blue label), or +1 (green label).

charge to either −1, 0, or +1 is not meant to test protonation states, given that the same number of charges and the same set of charge locations are used in each test. Optimization under these varying constraints suggests whether the optimized binding affinity or the optimal charge distribution is sensitive to the overall inhibitor charge, and for this problem neither appears to be the case.

The unbound- and bound-state geometries consisted of 3821 and 138,770 curved boundary elements, respectively. Calculating the explicit Hessian required 461 applications of the bound-state integral operator $M_{2,b}$. In contrast, using reverse-Schur co-optimization, the unconstrained and linear-equality constrained problems each required at most 2 $\hat{M}_2$ matrix-vector products, which is essentially the same cost as 2 $M_{2b}$ matrix-vector products owing to the small size of the unbound system. The more than 200-fold reduction in the number of matrix-vector products for such a small problem suggests that the model problems based in Figure 3 may actually be more computationally challenging than realistic problems, because a model problem of comparable size showed an acceleration of only a factor of 10 using co-optimization over an explicit-Hessian method (see Figure 8).

This example illustrates one weakness of our chosen metric for performance improvement—the reduction in the number of $\hat{M}_2$ matrix-vector products. This metric neglects the startup cost that must be paid regardless of whether one intends to use explicit-Hessian or co-optimization methods and thus the overall compute time required to obtain an unconstrained optimium is not being reduced by a factor of 200. Our decision to use the number of $\hat{M}_2$ matrix-vector products as an improvement metric is based on the consideration that it is impossible to perform the optimization without paying the initialization cost. Thus, from a theoretical perpective it is appropriate to neglect this cost in comparing optimization methods.

Nevertheless, as a practical matter, it is important to understand the impact of co-optimization on overall computational cost. On a 2-GHz Intel MacBook pro, the planar-boundary-element simulations require approximately 100 s of setup time, and the bound-state simulations require an average 5 s (6 GMRES iterations are generally required for bound-state simulations). The overall unconstrained co-optimization cost is thus about 102 s (100 s for initialization and 2 GMRES iterations to solve the unconstrained co-optimization problem), whereas the overall optimization cost for the BEM explicit-Hessian approach is about 300 s (100 s for initialization and 5 s for calculating each of the 40 columns of the Hessian). Thus the total reduction in computational cost is almost a factor of 3. For large problems such as optimizing protein−protein interactions, it can be expected that the total reduction will be even larger.

**5.2.2. HIV-1 Protease and Inhibitor.** The 75-atom inhibitor darunavir binds tightly to HIV-1 protease.[28] In Figure 12, the atoms are labeled with indices corresponding to the entries in Table 1, which lists the RESP-derived charge values, the unconstrained co-optimized partial atomic charge values, optimal charges under a zero total charge equality constraint, and the optimal charges computed with box constraints such that no charge exceeded 0.85*e* in magnitude, with sum-of-charge constraints set to −1, 0, and 1.

We emphasize that the equality constraints have not been introduced to evaluate protonation effects but only to estimate the influence of total charge on the optimal solution (and the associated binding free energy). Our results illustrate that the faster co-optimization method generates results consistent with the traditional approach. The unbound- and bound-state geometry discretizations consisted of 5892 and 133,067 curved boundary elements. Computation of the explicit Hessian required 576 applications of the operator $M_{2,b}$, and unconstrained co-optimization required 15 $M_{2\,b}$ matrix-vector products. In Figure 13 are plotted the unconstrained optimal charges computed using the explicit-Hessian and the co-optimization methods and again, the answers agree extremely well.

Electrostatic optimization of darunavir in the HIV-1 protease active site led to a significant improvement in the

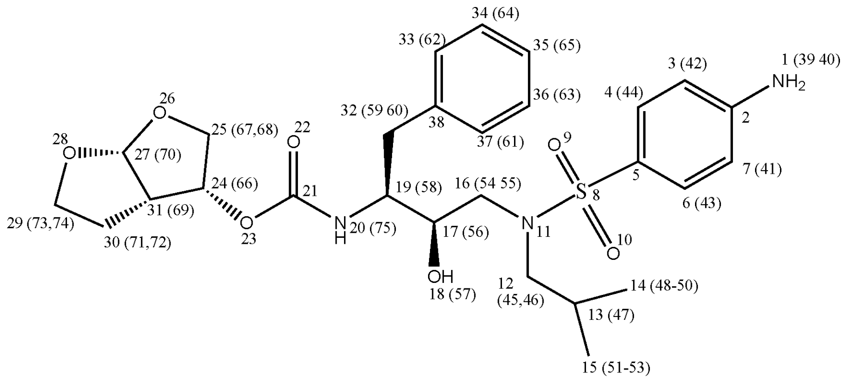


***Figure 12.*** The HIV-1 protease inhibitor darunavir[28] with atom indices for reference to Table 1. Hydrogen atom indices are indicated in parentheses adjacent to the atoms to which they are bonded.

predicted electrostatic binding free energy. The ligand is net neutral, and the wild-type charge distribution gives a finite-difference-calculated electrostatic binding free energy of 27.54 kcal/mol, and the unconstrained optimal solution computed using co-optimization gives an electrostatic binding free energy of 5.48 kcal/mol. The resulting optimized binding free energies for the bound-constrained problems were for the $-1e$ problem, 10.40 kcal/mol; for the neutral problem, 6.89 kcal/mol; for the $+1e$ problem, 5.73 kcal/mol.

The improvement on optimization can be attributed mainly to an accumulation of positive charge in the center of the ligand, near the negatively charged aspartyl dyad. These atoms are buried in both the bound and unbound states due to the molecular shape and consequently can take larger charge values without incurring significant desolvation penalties. For the unconstrained problems, the net ligand charge for the co-optimization method was $0.64e$, $0.64e$ for the curved-boundary-element explicit-Hessian method, and $1.00e$ using the finite-difference method.

Inhibitor atoms that make direct hydrogen-bonding interactions with the protease, such as the aniline nitrogen and hydrogen atoms (atom indices 1, 39, and 40), hydroxyl group (indices 18 and 57), and bis-tetrahydrofuran oxygen atoms (indices 26 and 28) have optimal charges very similar to their quantum-mechanically derived values.

## 6. Discussion

In this paper we have described an efficient technique, which we call reverse-Schur co-optimization, for calculating the molecular charge distribution that optimizes the electrostatic component of the free energy of binding to another molecule. The approach exhibits substantially better performance than traditional optimization approaches, which explicitly calculate the Hessian matrix before optimization. The co-optimization approach is a PDE-constrained optimization technique and exhibits comparable performance to alternative PDE-constrained optimization techniques that are well-known in other areas of computational science and engineering. Although this paper has presented an approach based on a boundary-element method (BEM) for solving the electrostatics problem, no fundamental issues seem to preclude the use of other numerical methods in a co-optimization approach. The critical elements for efficient co-optimization appear to be the availability of good preconditioners and sufficiently accurate but computationally inexpensive, Hessian approximations.

The reverse-Schur approach to PDE-constrained optimization is only possible because the PDE state variables and the decision variables are linearly related. This structure enables the PDE to be incorporated as a final algebraic manipulation before numerically solving the linear systems associated with quadratic programming. The all-at-once and nested approaches PDE-constrained optimization techniques, in contrast, are much more flexible with respect to the relationships between the state and decision variables.

Regularization—the penalization of certain search directions associated with the smallest eigenvalues—is more complicated for PDE-constrained approaches than for methods that rely on explicit Hessians. However, the BIBEE/P approach to estimating electrostatic interactions[91] has been demonstrated to generate a sufficiently accurate Hessian approximation whose eigendecomposition can be used as the basis for deriving a penalty function. The superior conditioning of purely second-kind integral-equation formulations[56,75] relative to first-kind or mixed first-second-kind equations has an unexpected consequence in that the Yoon and Lenhoff formulation cannot be used to generate a preconditioner-based Hessian approximation.

A number of extensions to the co-optimization technique may make it still more efficient. For instance, primal-dual interior-point methods are less efficient than active-set methods for "warm start" problems, in which one begins optimizing from a neighborhood of the optimal solution. Coupling co-optimization to an active-set solver might therefore significantly reduce the cost required to solve problems that differ only by the inclusion of varying constraints. Furthermore, a co-optimization warm-start method may be faster than the present implementation because there exist ways (such as Gasteiger-Marsili charges[101]) to rapidly estimate a wild-type charge distribution that could be used as an initial guess for the optimal distibution. One might also save the Hessian-vector products as they are formed, in essence allowing the Hessian to be "built" such that after a sufficient number of optimizations have been performed, the solver is using a completely explicit Hessian.

The efficiency gains afforded by PDE-constrained methods in general, not just the reverse-Schur approach presented here, should allow significantly larger and more complex

***Table 1.*** Partial Atomic Charges for Darunavir[a]

| atom index | RESP-fit | unconstrained optimal | equality-constrained $\Sigma q_i = 0$ | box-constrained $\Sigma q_i = -1$ | box-constrained $\Sigma q_i = 0$ | box-constrained $\Sigma q_i = +1$ |
|---|---|---|---|---|---|---|
| 1 | −0.87 | −0.52 | −0.45 | −0.33 | −0.44 | −0.55 |
| 2 | 0.40 | 0.16 | 0.09 | −0.05 | 0.06 | 0.16 |
| 3 | −0.29 | 0.14 | 0.22 | 0.35 | 0.22 | 0.10 |
| 4 | −0.18 | −0.38 | −0.46 | −0.61 | −0.48 | −0.37 |
| 5 | 0.02 | 0.37 | 0.43 | 0.62 | 0.55 | 0.47 |
| 6 | −0.18 | −0.09 | −0.12 | −0.24 | −0.20 | −0.16 |
| 7 | −0.29 | −0.14 | −0.12 | −0.03 | −0.08 | −0.11 |
| 8 | 0.86 | 0.09 | 0.07 | −0.33 | −0.33 | −0.35 |
| 9 | −0.52 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 |
| 10 | −0.50 | 0.07 | 0.06 | 0.08 | 0.09 | 0.10 |
| 11 | −0.36 | −2.35 | −2.25 | −0.85 | −0.85 | −0.85 |
| 12 | −0.06 | 0.45 | 0.19 | −0.75 | −0.68 | −0.48 |
| 13 | 0.15 | 0.55 | 0.80 | 0.85 | 0.81 | 0.49 |
| 14 | −0.14 | −0.19 | −0.22 | −0.02 | −0.11 | −0.09 |
| 15 | −0.31 | −0.22 | −0.31 | −0.16 | −0.20 | −0.09 |
| 16 | −0.08 | 1.62 | 1.66 | 0.85 | 0.85 | 0.85 |
| 17 | 0.09 | −0.17 | −0.22 | −0.21 | −0.23 | −0.19 |
| 18 | −0.67 | 0.39 | 0.37 | 0.31 | 0.33 | 0.35 |
| 19 | 0.05 | −1.08 | −0.87 | 0.15 | 0.10 | −0.11 |
| 20 | −0.38 | −0.41 | −0.46 | −0.85 | −0.85 | −0.79 |
| 21 | 0.68 | 0.76 | 0.76 | 0.82 | 0.85 | 0.85 |
| 22 | −0.56 | −0.38 | −0.39 | −0.42 | −0.41 | −0.39 |
| 23 | −0.39 | −0.13 | −0.13 | −0.03 | −0.05 | −0.07 |
| 24 | 0.12 | −0.71 | −0.67 | −0.82 | −0.85 | −0.85 |
| 25 | 0.05 | 0.88 | 0.86 | 0.79 | 0.83 | 0.85 |
| 26 | −0.49 | −0.61 | −0.58 | −0.52 | −0.56 | −0.59 |
| 27 | 0.44 | 0.21 | 0.17 | −0.03 | 0.06 | 0.16 |
| 28 | −0.51 | −0.34 | −0.32 | −0.26 | −0.30 | −0.33 |
| 29 | 0.07 | −0.35 | −0.48 | −0.54 | −0.37 | −0.19 |
| 30 | −0.12 | 1.70 | 1.89 | 0.85 | 0.85 | 0.85 |
| 31 | −0.01 | −0.14 | −0.18 | 0.43 | 0.35 | 0.25 |
| 32 | −0.08 | 1.36 | 1.26 | 0.84 | 0.85 | 0.85 |
| 33 | −0.16 | 0.24 | 0.26 | 0.23 | 0.17 | 0.12 |
| 34 | −0.19 | −0.09 | −0.13 | −0.15 | −0.07 | −0.01 |
| 35 | −0.18 | 0.11 | 0.16 | 0.20 | 0.10 | 0.01 |
| 36 | −0.19 | −0.11 | −0.10 | −0.04 | −0.04 | −0.05 |
| 37 | −0.05 | 0.12 | 0.10 | −0.03 | −0.02 | −0.00 |
| 38 | −0.06 | −0.63 | −0.61 | −0.43 | −0.39 | −0.35 |
| 39 | 0.39 | 0.22 | 0.19 | 0.16 | 0.19 | 0.22 |
| 40 | 0.39 | 0.10 | 0.03 | −0.08 | 0.03 | 0.14 |
| 41 | 0.18 | 0.14 | 0.13 | 0.11 | 0.13 | 0.14 |
| 42 | 0.19 | −0.13 | −0.22 | −0.37 | −0.22 | −0.07 |
| 43 | 0.21 | 0.03 | 0.03 | 0.04 | 0.04 | 0.04 |
| 44 | 0.21 | 0.20 | 0.20 | 0.21 | 0.21 | 0.21 |
| 45 | 0.07 | 0.14 | 0.17 | 0.25 | 0.26 | 0.25 |
| 46 | 0.11 | 0.06 | 0.09 | 0.17 | 0.18 | 0.18 |
| 47 | 0.02 | 0.07 | 0.03 | 0.05 | 0.05 | 0.11 |
| 48 | 0.03 | 0.04 | 0.03 | −0.01 | 0.02 | 0.04 |
| 49 | 0.00 | −0.07 | −0.06 | −0.09 | −0.07 | −0.07 |
| 50 | 0.06 | 0.04 | 0.03 | −0.15 | −0.11 | −0.10 |
| 51 | 0.09 | 0.03 | 0.01 | −0.04 | 0.02 | 0.06 |
| 52 | 0.06 | −0.12 | −0.19 | −0.33 | −0.18 | −0.06 |
| 53 | 0.08 | 0.05 | 0.06 | 0.03 | 0.03 | 0.01 |
| 54 | 0.14 | −0.04 | −0.06 | 0.04 | 0.04 | 0.05 |
| 55 | 0.11 | −0.28 | −0.28 | −0.15 | −0.15 | −0.16 |
| 56 | 0.14 | 0.34 | 0.34 | 0.28 | 0.29 | 0.30 |
| 57 | 0.41 | 0.25 | 0.25 | 0.28 | 0.29 | 0.29 |
| 58 | 0.18 | 0.13 | 0.08 | −0.22 | −0.22 | −0.18 |
| 59 | 0.07 | −0.21 | −0.20 | −0.16 | −0.16 | −0.15 |
| 60 | 0.07 | −0.12 | −0.11 | −0.09 | −0.08 | −0.06 |
| 61 | 0.09 | −0.06 | −0.05 | −0.03 | −0.03 | −0.04 |
| 62 | 0.16 | 0.01 | 0.00 | 0.01 | 0.02 | 0.03 |
| 63 | 0.17 | −0.00 | −0.01 | −0.04 | −0.02 | −0.00 |
| 64 | 0.17 | −0.08 | −0.08 | −0.09 | −0.09 | −0.08 |
| 65 | 0.17 | −0.10 | −0.22 | −0.40 | −0.20 | −0.01 |
| 66 | 0.10 | 0.14 | 0.12 | 0.14 | 0.15 | 0.16 |
| 67 | 0.06 | −0.05 | −0.06 | −0.05 | −0.04 | −0.03 |
| 68 | 0.11 | −0.14 | −0.14 | −0.14 | −0.14 | −0.13 |
| 69 | 0.04 | 0.03 | 0.04 | −0.04 | −0.03 | −0.02 |
| 70 | 0.06 | 0.22 | 0.18 | 0.12 | 0.19 | 0.25 |
| 71 | 0.07 | −0.43 | −0.43 | −0.09 | −0.17 | −0.24 |
| 72 | 0.05 | −0.44 | −0.70 | −0.69 | −0.37 | −0.05 |
| 73 | 0.11 | 0.08 | 0.11 | 0.19 | 0.13 | 0.07 |
| 74 | 0.09 | 0.13 | 0.16 | 0.22 | 0.18 | 0.13 |
| 75 | 0.23 | 0.24 | 0.24 | 0.27 | 0.28 | 0.29 |

[a] All charge values are given as multiples of the electron charge magnitude *e* and have been rounded to the nearest 0.01*e*. Box-constrained optimization enforced that each charge had maximum magnitude 0.85*e*.

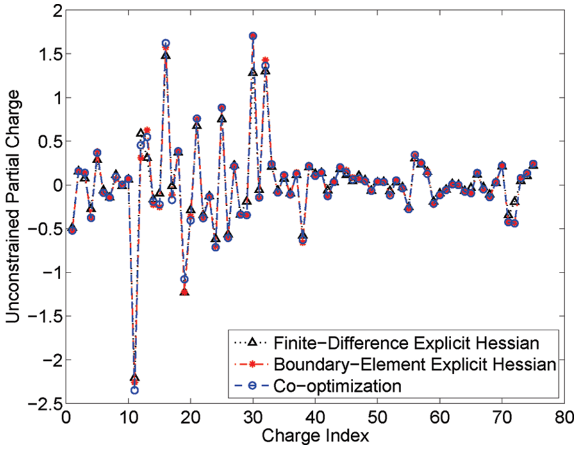


***Figure 13.*** Unconstrained optimal partial atomic charges for the HIV-1 protease inhibitor darunavir computed using finite-difference and boundary-element explicit Hessians as well as using the reverse-Schur co-optimization method. The boundary-element simulations employed curved boundary elements.

biological systems to be studied. Computational redesign of proteins, for instance, may produce optimization problems with dimension greater than one thousand. For these problems, co-optimization can offer a cost reduction of over 2 orders of magnitude;[92] such an acceleration may allow the evaluation of many more candidate ligands or ligand poses. Also, geometry can now be varied to assess the best-case energetic cost of adding of a functional group as it changes a binding partner's desolvation penalty or the effects of molecular flexibility.[16] For example, it may be computationally feasible to use co-optimization to study the influence of different protonation states on binding free energies and on the optimal charge distributions associated with each state. Finally, the original electrostatic optimization paper by Lee and Tidor noted an unexpected asymmetry between the receptor charge distribution and the calculated optimal ligand distribution.[1] Because co-optimization allows the use of substantially larger basis sets, it may be able to develop techniques that can identify optimal charge placement as well as value.

**Acknowledgment.** The authors thank J. Nocedal, J.-H. Lee, D. J. Willis, D. F. Green, M. Knepley, and S. Leyffer for useful discussions. J. P. Bardhan gratefully acknowledges support from a Department of Energy Comptutational Science Graduate Fellowship as well as from a Wilkinson Fellowship in Scientific Computing funded by the Mathematical, Information, and Computational Sciences Division Subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Dept. of Energy, under Contract DE-AC02-06CH11357. J. K. White acknowledges support from grants from the National Science Foundation and the Singapore-MIT Alliance. This work was supported in part by the National Institutes of Health (GM065418).

## References

(1) Lee, L.-P.; Tidor, B. *J. Chem. Phys.* **1997**, *106*, 8681–8690.

(2) Kangas, E.; Tidor, B. *J. Phys. Chem. B* **2001**, *105*, 880–888.

(3) Lee, L.-P.; Tidor, B. *Nat. Struct. Biol.* **2001**, *8*, 73–76.

(4) Green, D. F.; Tidor, B. *J. Mol. Biol.* **2004**, *342*, 435–452.

(5) Chong, L. T.; Dempster, S. E.; Hendsch, Z. S.; Lee, L.-P.; Tidor, B. *Protein Sci.* **1998**, *7*, 206–210.

(6) Kangas, E.; Tidor, B. *J. Chem. Phys.* **1998**, *109*, 7522–7545.

(7) Kangas, E.; Tidor, B. *Phys. Rev. E* **1999**, *59*, 5958–5961.

(8) Kangas, E.; Tidor, B. *J. Chem. Phys.* **2000**, *112*, 9120–9131.

(9) Lee, L.-P.; Tidor, B. *Protein Sci.* **2001**, *10*, 362–377.

(10) Mandal, A.; Hilvert, D. *J. Am. Chem. Soc.* **2003**, *125*, 5598–5599.

(11) Sulea, T.; Purisima, E. O. *J. Phys. Chem. B* **2001**, *105*, 889–899.

(12) Sulea, T.; Purisima, E. O. *Biophys. J.* **2003**, *84*, 2883–2896.

(13) Sims, P. A.; Wong, C. F.; McCammon, J. A. *J. Comput. Chem.* **2004**, *25*, 1416–1429.

(14) Green, D. F.; Tidor, B. *Proteins* **2005**, *60*, 644–657.

(15) Armstrong, K. A.; Tidor, B.; Cheng, A. C. *J. Med. Chem.* **2006**, *49*, 2470–2477.

(16) Gilson, M. K. *J. Chem. Theory Comput.* **2006**, *2*, 259–270.

(17) Selzer, T.; Albeck, S.; Schreiber, G. *Nat. Struct. Biol.* **2000**, *7*, 537–541.

(18) Shaul, Y.; Schreiber, G. *Proteins* **2005**, *60*, 341–352.

(19) Brock, K.; Talley, K.; Coley, K.; Kundrotas, P.; Alexov, E. *Biophys. J.* **2007**, *93*, 3340–3352.

(20) Haftka, R. *AIAA J.* **1985**, *23*, 1099–1103.

(21) Orozco, C.; Ghattas, O. *AIAA J.* **1992**, *30*, 1877–1885.

(22) Biegler, L.; Nocedal, J.; Schmid, C. *SIAM J. Optim.* **1995**, *5*, 314–347.

(23) Dennis, J.; Heinkenschloss, M.; Vicente, L. *SIAM J. Control Optim.* **1998**, *36*, 1750–1794.

(24) Biros, G.; Ghattas, O. *SIAM J. Sci. Comput.* **2005**, *27*, 687–713.

(25) Biros, G.; Ghattas, O. *SIAM J. Sci. Comput.* **2005**, *27*, 714–739.

(26) Brayton, R. K.; Hachtel, G. D.; Sangiovanni-Vincentelli, A. L. *Proc. IEEE* **1981**, *69*, 1334–1362.

(27) Fisher, M.; Nocedal, J.; Tremolet, Y.; Wright, S. J. Data assimilation in weather forecasting: a case study in PDE-constrained optimization. Optimization and Engineering, 2008. DOI: 10.1007/s11081-008-9051-5. http://www.springerlink.com/content/e47441q543236t31/ (accessed August 1, 2009).

(28) Surleraux, D. L.; Tahri, A.; Verschueren, W. G.; Pille, G. M.; de Kock, H. A.; Jonckers, T. H.; A. Peeters, A.; Meyer, S. D.; Azijn, H.; Pauwels, R.; de Bethune, M. P.; King, N. M.; Prabu-Jeyabalan, M.; Schiffer, C. A.; Wigerinck, P. B. *J. Med. Chem.* **2005**, *48*, 1813–1822.

(29) Huff, J. R. *J. Med. Chem.* **1991**, *34*, 2305–2314.

(30) Flexner, C. *N. Engl. J. Med.* **1998**, *338*, 1281–1292.

(31) Anderson, M.; Beattie, J.; Breault, G.; Breed, J.; Byth, K.; Culshaw, J.; Ellston, R.; Green, S.; Minshull, C.; Norman, R.; Pauptit, R.; Stanway, J.; Thomas, A.; Jewsbury, P. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3021.

(32) Vandenheuvel, S.; Harlow, E. *Science* **1993**, *262*, 2050–2054.

(33) Hartwell, L. H.; Kastan, M. B. *Science* **1994**, *266*, 1821–1828.

(34) Honig, B.; Sharp, K.; Yang, A. S. *J. Phys. Chem.* **1993**, *97*, 1101–1109.

(35) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. *J. Am. Chem. Soc.* **2003**, *125*, 9523–9530.

(36) Wagoner, J. A.; Baker, N. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8331–8336.

(37) Sharp, K. A.; Honig, B. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301–332.

(38) Richards, F. M. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151–176.

(39) Juffer, A. H.; Botta, E. F. F.; van Keulen, B. A. M.; van der Ploeg, A.; Berendsen, H. J. C. *J. Comput. Phys.* **1991**, *97*, 144–171.

(40) Miertus, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117–129.

(41) Warwicker, J.; Watson, H. C. *J. Mol. Biol.* **1982**, *157*, 671–679.

(42) Shaw, P. B. *Phys. Rev. A* **1985**, *32*, 2476–2487.

(43) Zauhar, R. J.; Morgan, R. S. *J. Mol. Biol.* **1985**, *186*, 815–820.

(44) Klapper, I.; Hagstrom, R.; Fine, R.; Sharp, K.; Honig, B. *Proteins* **1986**, *1*, 47–59.

(45) Zauhar, R. J.; Morgan, R. S. *J. Comput. Chem.* **1988**, *9*, 171–187.

(46) Gilson, M. K.; Honig, B. *Proteins* **1988**, *4*, 7–18.

(47) Yoon, B. J.; Lenhoff, A. M. *J. Comput. Chem.* **1990**, *11*, 1080–1086.

(48) Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435–445.

(49) Holst, M. J. Ph.D. Thesis, Univ. of Ill. at Urbana-Champaign, 1993.

(50) You, T. J.; Harvey, S. C. *J. Comput. Chem.* **1993**, *14*, 484–501.

(51) Cammi, R.; Tomasi, J. *J. Comput. Chem.* **1995**, *16*, 1449–1458.

(52) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Ridgway-Scott, L.; McCammon, J. A. *Comput. Phys. Commun.* **1995**, *91*, 57–95.

(53) Purisima, E. O.; Nilar, S. H. *J. Comput. Chem.* **1995**, *16*, 681–689.

(54) Bharadwaj, R.; Windemuth, A.; Sridharan, S.; Honig, B.; Nicholls, A. *J. Comput. Chem.* **1995**, *16*, 898–913.

(55) Cances, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032–3041.

(56) Liang, J.; Subramaniam, S. *Biophys. J.* **1997**, *73*, 1830–1841.

(57) Holst, M.; Baker, N.; Wang, F. *J. Comput. Chem.* **2000**, *21*, 1319–1342.

(58) Rocchia, W.; Alexov, E.; Honig, B. *J. Phys. Chem. B* **2001**, *105*, 6507–6514.

(59) Bordner, A. J.; Huber, G. A. *J. Comput. Chem.* **2003**, *24*, 353–367.

(60) Boschitsch, A. H.; Fenley, M. O.; Zhou, H.-X. *J. Phys. Chem. B* **2002**, *106*, 2741–54.

(61) Boschitsch, A. H.; Fenley, M. O. *J. Comput. Chem.* **2004**, *25*, 935–955.

(62) Zhou, H. X. *Biophys. J.* **1993**, *65*, 955–963.

(63) Purisima, E. O. *J. Comput. Chem.* **1998**, *19*, 1494–1504.

(64) Kuo, S. S.; Altman, M. D.; Bardhan, J. P.; Tidor, B.; White, J. K. Fast Methods for Simulation of Biomolecule Electrostatics. *Proc. Int. Conf. Comput.-Aided Des. (ICCAD)*, San Jose, CA, ACM: New York, NY, 2002; pp 466−473.

(65) Altman, M. D.; Bardhan, J. P.; Tidor, B.; White, J. K. *IEEE Trans. Comput.-Aided Des.* **2006**, *25*, 274–284.

(66) Altman, M. D.; Bardhan, J. P.; White, J. K.; Tidor, B. *J. Comput. Chem.* **2009**, *30*, 132–153.

(67) Chipman, D. M. *J. Chem. Phys.* **2004**, *120*, 5566–5575.

(68) Altman, M. D.; Bardhan, J. P.; White, J. K.; Tidor, B. An Efficient and Accurate Surface Formulation for Biomolecule Electrostatics in Non-ionic Solution. *Proc. 2005 IEEE Eng. Med. Biol. Conf. (IEEE-EMBS) 2005*, Shanghai; IEEE, Piscataway, NJ, 2005; pp 7591−7595.

(69) Grandison, S.; Penfold, R.; Vanden-Broeck, J.-M. *J. Comput. Phys.* **2007**, *224*, 663–680.

(70) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210–3221.

(71) Mennucci, B.; Cancès, E.; Tomasi, J. *J. Phys. Chem. B* **1997**, *101*, 10506–10517.

(72) Levitt, D. G. *Biophys. J.* **1978**, *22*, 209–219.

(73) Attard, P. *J. Chem. Phys.* **2003**, *119*, 1365–1372.

(74) Boda, D.; Voliskó, M.; Eisenberg, B.; Nonner, W.; Henderson, D.; Gillespie, D. *J. Chem. Phys.* **2006**, *125*, 034901.

(75) Atkinson, K. E. *The Numerical Solution of Integral Equations of the Second Kind*; Cambridge University Press: 1997.

(76) Hsiao, G. C.; Wendland, W. L. *Encyclopedia of Computational Mechanics*; 2004.

(77) Zauhar, R. J. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 149–159.

(78) Bardhan, J. P.; Altman, M. D.; White, J. K.; Tidor, B. *J. Chem. Phys.* **2007**, *127*, 014701.

(79) Tausch, J.; Wang, J.; White, J. *IEEE Trans. Comput.-Aided Des.* **2001**, *20*, 1398–1405.

(80) Bardhan, J. P. *J. Chem. Phys.* **2009**, *130*, 094102.

(81) Nabors, K.; Korsmeyer, F. T.; Leighton, F. T.; White, J. *SIAM J. Sci. Comput.* **1994**, *15*, 713–735.

(82) Lu, B. Z.; Cheng, X. L.; Huang, J.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 19314–19319.

(83) Golub, G. H.; Loan, C. F. V. *Matrix Computations*, 3rd ed.; The Johns Hopkins University Press: Baltimore, MD, 1996.

(84) Saad, Y.; Schultz, M. *SIAM J. Sci. Stat. Comput.* **1986**, *7*, 856–869.

(85) Greengard, L.; Rokhlin, V. *J. Comput. Phys.* **1987**, *73*, 325–348.

(86) Phillips, J. R.; White, J. K. *IEEE Trans. Comput.-Aided Des.* **1997**, *16*, 1059–1072.

(87) Bertsekas, D. P. *Nonlinear Programming*, 2nd ed.; Athena Scientific: Nashua, NH, 1999.

(88) Wright, S. J. *Primal-Dual Interior Point Methods*; SIAM: Philadelphia, PA, 1997.

(89) Bardhan, J. P.; Lee, J. H.; Altman, M. D.; Benson, S.; Leyffer, S.; Tidor, B.; White, J. K. Biomolecule Electrostatic Optimization with an Implicit Hessian. In *Tech. Proc. 2004 Nanotech. Conf. Trade Show, v. 1, Boston MA, 2004*; Laudon, M., Romanowicz, B., Eds.; NSTI: Cambridge MA, 2004; pp 164−167.

(90) Sanner, M.; Olson, A. J.; Spehner, J. C. *Biopolymers* **1996**, *38*, 305–320.

(91) Bardhan, J. P. *J. Chem. Phys.* **2008**, *129*, 144105.

(92) Bardhan, J. P.; Lee, J. H.; Kuo, S. S.; Altman, M. D.; Tidor, B.; White, J. K. *Fast Methods for Biomolecule Charge Optimization. In Tech. Proc. 2003 Nanotech. Conf. Trade Show, v. 2; San Francisco, CA, 2003*; Laudon, M., Romanowicz, B., Eds.; NSTI: Cambridge, MA, 2003; pp 508−511.

(93) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(94) Brünger, A. T.; Karplus, M. *Proteins* **1988**, *4*, 148–156.

(95) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher III, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(96) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem. B* **1994**, *98*, 1978–1988.

(97) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komáromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*; Gaussian, Inc.: Pittsburgh, PA, 1998.

(98) Green, D. F.; Tidor, B. *J. Phys. Chem. B* **2003**, *107*, 10261–10273.

(99) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(100) Altman, M. D. Unpublished results.

(101) Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. *J. Chem. Theory Comput.* **2006**, *2*, 128–139.

CT9001174

JCTC Journal of Chemical Theory and Computation

# Coarse Point Charge Models For Proteins From Smoothed Molecular Electrostatic Potentials

Laurence Leherte* and Daniel P. Vercauteren

*Laboratoire de Physico-Chimie Informatique, Groupe de Chimie Physique Théorique et Structurale, University of Namur (FUNDP), Rue de Bruxelles 61, B-5000 Namur, Belgium*

Received April 21, 2009

**Abstract:** To generate coarse electrostatic models of proteins, we developed an original approach to hierarchically locate maxima and minima in smoothed molecular electrostatic potentials. A charge-fitting program was used to assign charges to the so-obtained reduced representations. Templates are defined to easily generate coarse point charge models for protein structures, in the particular cases of the Amber99 and Gromos43A1 force fields. Applications to four small peptides and to the ion channel KcsA are presented. Electrostatic potential values generated by the reduced models are compared with the corresponding values obtained using the original sets of atomic charges.

## I. Introduction

The design of protein coarse-grain (CG) models and their corresponding interaction potential functions are nowadays an active field of research, especially for solving problems such as protein folding and docking through, e.g., molecular mechanics (MM) and molecular dynamics (MD) methods.[1] Indeed, all-atom simulations may be out of practical computational resources for macromolecules, and a strategy to consider large size systems and long time scales in a simulation consists in limiting the number of interacting particles. Among the essential parameters involved in all-atom and CG potentials, electrostatic interactions are of crucial importance since they govern local and global properties, e.g., their stability, flexibility, etc. Various approaches to evaluate electrostatic interactions are, e.g., reviewed by Dong et al.[2] Nevertheless, evaluating the adequacy of a particular method is not straightforward; a presentation of this problem can, e.g., be found in the work of Schutz and Warshel[3] who discussed the choice of dielectric constants.

Common approaches used to design a CG description of a protein consist in reducing groups of atoms into single interaction sites. For example, in the work by Skepö et al.,[4] each amino acid (AA) is represented by a single spherical site, with unit or nul electric charge. The authors studied a proline-rich protein PRP-1 interacting with a mica surface using Monte Carlo simulations. Curcó et al.[5] developed a CG model of $\beta$-helical protein fragments, where the AAs are represented by two, three, or four blobs depending upon the AA type, in accordance with a best fitting between Monte Carlo (MC) all-atom and CG energies. In their work, the AAs are depicted by the amide hydrogen atom, the oxygen atom, the geometric center of the side chain (except for Gly), and a fourth blob whose position depends on the AA type (except for Gly, Ala, and Val). In the Basdevant et al. paper,[6] each AA residue is modeled using one sphere located on the geometric center of the backbone and one or two spheres located on the geometric centers of the side chain fragments (except for Gly). Differently, Pizzitutti et al.[7] represented each AA of a protein sequence by a charged dipolar sphere. For each AA, one CG sphere is located on the center-of-mass (com) of uncharged residues, while for charged residues, one CG is assigned to the com of the neutral part of the AA, and one CG is assigned to the com of the charged part. Charged residues are Arg, Asp, Glu, Lys, and terminal AAs. The authors show that, in protein association, their model provides a good approximation of the all-atom potential, if the distance between the protein surfaces is larger than the diameter of a solvent molecule. In a recent work, Zhang et al.[8] proposed a method to define CGs that reflect the collective motions computed by a principal component analysis of an atomistic MD trajectory. Each CG site is the com of a domain, i.e., a group of contiguous Cα atoms that

* Corresponding author. Telephone: +32-81-724560. Fax: +32-81-725466. E-mail: laurence.leherte@fundp.ac.be.

10.1021/ct900193m CCC: $40.75 © 2009 American Chemical Society
Published on Web 10/29/2009

move in a highly correlated fashion. Very recently, Bereau and Deserno[9] presented a generic CG model for proteins with one grain located on each of the N, Cα, and C atoms and a fourth grain on Cβ. Such a simple side chain description was aimed at facilitating the parametrization of the corresponding CG potential function.

The development of CG interaction potential functions is generally made either from atomistic interaction potential[10] or MD results[11−13] via experimental data, such as B-factors,[14] or from the fitting of a potential function achieved by matching CG and atomistic distributions.[13,15] For example, Lyman et al.[16] presented a new method for fitting spring constants to mean square CG−CG distance fluctuations computed from atomistic MD. One can also cite the inverse MC approach[17] used for iteratively adjusting an effective CG potential function until it matches a target radial distribution function. Noid et al.[18,19] proposed a statistical mechanics theory to check the consistency between CG and all-atom models. More specifically, the parametrization of the well-known MARTINI force field (FF), dedicated to MD simulations of biomolecular systems, is based on the reproduction of partitioning free energies between polar and apolar phases of a large number of chemical systems.[20,21] In that model, groups of four heavy atoms are represented by a single interaction center, except for small ring-like fragments. AAs, thus, consist of one to four side chain beads and one backbone bead.[21] Only four main types of interaction sites are defined: polar (P), nonpolar (N), apolar (C), and charged (Q). Each particle type has a number of subtypes, which allow for a more accurate representation of the chemical nature of the underlying atomistic structure. In MARTINI, only AA residues Arg, Asp, Glu, and Lys are charged. Such a description was, e.g., applied to protein channels embedded in a lipid membrane environment.[22] In the UNRES model,[23] a peptidic chain is represented by a sequence of backbone beads located at peptide bonds, while side chains are modeled as single beads attached to the Cα atoms, which are considered only to define the molecular geometry. In the so-called SimFold CG description and energy function, a mixed representation is used.[24,25] Residues of aqueous proteins are represented by backbone atoms N, Cα, C, O, and H and by one side chain centroid. In UNRES and SimFold, electrostatic interactions are not explicitely calculated using the Coulomb term like they are in the MARTINI FF for charged AAs.

Basdevant et al.[6] proposed an approach to determine charges for their reduced models built on the com of backbone and side chain groups of AA residues. Prior to these studies, Gabdoulline et al.[26] used a model that consisted of a small number of point charges (monopoles) suitable for the description of the intermolecular electrostatic interactions. As later applied by Basdevant et al.,[6] these charges were derived from a fitting procedure applied to reproduce the molecular electrostatic potential (MEP) obtained by solving the Poisson−Boltzmann equation. In their example, the charges are located at the geometric centers of the head groups of the charged residues. The mimick of all-atom electrostatic interactions using a limited set of point charges was also proposed by Berardi et al.,[27] who applied a genetic algorithm to determine the location and the values of a given number of charges for molecules involved in liquid-crystalline materials. Extended multipolar models are also reported, such as the one described by Golubkov et al.[28] In that approach, illustrated for small molecules, such as water, methanol, and benzene, the charge distribution is represented by a point multipole expression with charges (nul for the hereabove examples), by a dipole, and by quadrupole moments placed at the molecular com. These dipole and quadrupole moments were set equal to the corresponding average moments obtained from analyses of all-atom MD simulations. In the framework of proteins, Cascella et al.[29] presented a method to parametrize an AA reduced model that allows reproduction of all-atom electrostatic properties evaluated as averages during MD simulations for the side chains and statistically for the backbones. Reviews on the progresses of CG dynamical models can also be found in additional references.[30,31]

Multiscale methods, that combine several levels of description, are also appealing since they allow to model limited regions of space with details while representing the outer regions by coarser models.[32,33] The consideration of outer influences, such as external stresses[34] or solvent effects,[35] can also be treated with CG approaches.

In the present paper, we propose a method to elaborate coarse point charge models for protein structures from smoothed MEPs. The quality of such models is approached by comparing CG-based MEP and dipole values with the corresponding all-atom properties. In a previous work,[36] a protein structure was decomposed into separate molecular fragments that were determined through a merging/clustering procedure of atom trajectories generated in progressively smoothed electron density (ED) distribution functions. This was followed by a second study[37] where atoms were clustered according to their trajectories defined in a smoothed MEP function. That procedure allowed to locate the corresponding MEP local maxima (peaks) and minima (pits). A fitting algorithm was applied to evaluate the peak and pit charges. Results, presented for the twenty AAs, were derived from the all-atom Amber charges reported in Duan et al.[38]

With respect to that second paper, we have extended, refined, and automated our approach, which now consists of the three following steps. First, extrema corresponding to the AA backbone are located in the smoothed MEP of a β-pentadecapeptide $Gly_{15}$ model. For the AA side chains, the CGs are identified as extrema in the smoothed MEPs of each of the 20 natural AA residues in their isolated state from both the Amber99,[39] as available in PDB2PQR,[40,41] and Gromos43A1[42] sets of charges. Gromos charges were taken from the files provided with the software SwissPDB-Viewer.[43,44] Second, charges are assigned to each of the CGs through a charge-fitting procedure applied to reproduce unsmoothed MEP grid values and dipole moments. Third, a library of the resulting AA point charge templates, including CG locations and their charge values, is built for further modeling of proteins.

In Section II, we present a brief overlook of the theoretical background. In Section III, we describe the methodology to design the CG templates of the AAs from the pentadecapeptide $\beta$-Gly$_{15}$ and the isolated AA models as well as from their associated atom charges. Finally, in Section IV, we detail applications to four small peptides and to the ion channel KcsA. Let us finally note that further in the text: (i) we will use the expressions "reduced" and "CG" indifferently, and (ii) that all three-dimensional (3D) illustrations were generated with OpenDX[45] unless otherwise stated.

## II. Theoretical Background

In this section, we present the mathematical formalism that was used to design a protein-reduced representation and its corresponding point charges. First, the smoothing algorithm is described. This description is followed by the mathematical expressions that are specific to the Coulomb electrostatic interaction function. Finally, we detail the approach applied to calculate the CG point charges.

**A. Smoothing Algorithm.** To follow the trajectories of the local maxima and minima in a MEP function, as a function of the degree of smoothing, we implemented an algorithm initially described by Leung et al.[46] The authors initially proposed a method to model the blurring effect in human vision. This was achieved by filtering a digital image $p(x)$ through a convolution product with a Gaussian function and by assigning each data point of the resulting $p(x, t)$ image to a cluster via a dynamical equation built on the gradient of the convoluted image:

$$x(n + 1) = x(n) + h\nabla_x p(x, t) \quad (1)$$

where $h$ is defined as the step length. We adapted this idea to 3D images, such as ED and MEP functions, $f$, such as:

$$\mathbf{r}_{f(t)} = \mathbf{r}_{f(t-\Delta t)} + \frac{\Delta}{f(t)}\nabla f(t) \quad (2)$$

where $\mathbf{r}$ stands for the location vector of a point in a 3D function, such as a MEP field.

The various steps of the resulting merging/clustering algorithm are as follows: First, at scale $t = 0$, each atom of a molecular structure is considered as either a local maximum (peak) or minimum (pit) of the MEP function. All atoms are consequently taken as the starting points of the merging procedure. Second, as $t$ increases from 0 to a given maximal value $t_{\max}$, each point moves continuously along a gradient path to reach a location in the 3D space where $\nabla f(t) = 0$. From a practical point of view, this consists of following the trajectory of the peaks and pits on the MEP distribution surface calculated at $t$ according to eq 2. The trajectory search is stopped when $|\nabla f(t)|$ is lower or equal to a limit value, grad$_{\text{lim}}$. Once all peak/pit locations are found, close points are merged if their interdistance is lower than the initial value of $\Delta^{1/2}$. The procedure is repeated for each selected value of $t$. If the initial $\Delta$ value is too small to allow convergence toward a local maximum or minimum within the given number of iterations, then its value is doubled (a scaling factor that is arbitrarily selected), and the procedure is repeated until final convergence.

**B. Molecular Electrostatic Potentials.** The electrostatic potential function generated by a molecule $A$ is simply calculated as a summation over its atomic contributions:

$$V_A(\mathbf{r}) = \sum_{a\in A} \frac{q_a}{|\mathbf{r} - \mathbf{R}_a|} \quad (3)$$

where $\mathbf{R}_a$ is the position vector of atom $a$, and $q_a$ is the electric charge. A smoothed version can be expressed as:

$$V_{A,t}(\mathbf{r}) = \sum_{a\in A} \frac{q_a}{|\mathbf{r} - \mathbf{R}_a|} \operatorname{erf}\left(\frac{|\mathbf{r} - \mathbf{R}_a|}{2\sqrt{t}}\right) \quad (4)$$

where the error function erf can be calculated using the analytically derivable expression:[47]

$$\operatorname{erf}(x) = 1 - (a_1T + a_2T^2 + a_3T^3 + a_4T^4 + a_5T^5)e^{-x^2} \quad \text{with } T = \frac{1}{1 + px} \quad (5)$$

The values of the parameters $p$ and $a$ are: $p = 0.3275911$, $a_1 = 0.254829595$, $a_2 = -0.284496736$, $a_3 = 1.421413741$, $a_4 = -1.453152027$, and $a_5 = 1.061405429$.[47] Equation 4 is identical to the expression found in the potential smoothing approach, a well-known technique used in MM applications.[48]

**C. Calculation of Point Charges.** Charge values were obtained using the charge-fitting program QFIT.[49] Among the approaches that are reported in the literature, e.g., either excluding the MEP grid points that are located too close or too far from the molecular structure under consideration or including grid points located at large distances up to 30−45 Å from the molecular center,[26] we selected the first approach to modulate the influence of the neighborhood of the AA under interest. Indeed, we wished to establish AA CG charges that are as independent as possible on the selected models. All MEP grids were built using either the Amber99[39] or Gromos43A1[42,43] point charges, assigned using the software PDB2PQR,[40,41] with a grid step of 0.5 Å. Fittings were achieved by considering points located at distances between 1.4 and 2.0 times the van der Waals (vdW) radius of the atoms. These two limiting distance values were selected after the so-called Merz−Singh−Kollman scheme.[50]

In all fittings presented, the total electric charge and the magnitude of the molecular dipole moment were constrained to be equal to the corresponding all-atom Amber99 or united-atom Gromos43A1 MEP values. The quality of the fittings was evaluated by two root-mean-square deviation (rmsd) values, i.e., the rmsdV determined between the MEP grid values obtained using the fitted charges and the reference unsmoothed MEP grid values and the rmsd$\mu$ evaluated between the dipolar value calculated from the fitted CG charges and the reference dipole moment of the molecular structure. All dipole moment components were calculated with the origin of the atom coordinates set to (0. 0. 0.).

## III. Results and Discussion

This section is dedicated to the elaboration of coarse point charge models of proteins based on the local maxima and minima observed in their smoothed MEP functions. After selection of the best smoothing degree to work at, the first two steps of our strategy rely on the CG description of the protein backbone and the development of side chain CG

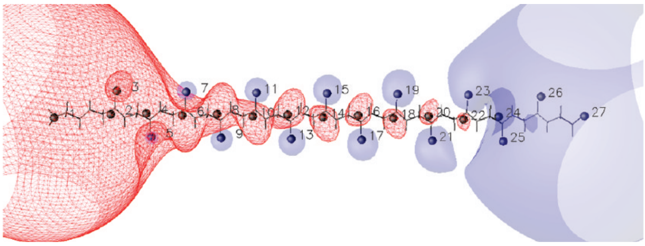


***Figure 1.*** Amber99 MEP isocontours (blue plain surface: −0.03; red mesh: 0.03 $e^-$/bohr) of $\beta$-Gly$_{15}$ with charged $NH_3^+$ and $COO^-$ ends, as obtained by smoothing the original MEP at $t = 1.25$ bohr$^2$. Local maxima and minima (black spheres) were obtained using the hierarchical merging/clustering algorithm applied to the all-atom Amber99 MEP function.

models. Each stage involves the determination of CG locations and corresponding electrostatic point charges. The final part of the section focuses on the application of our CG models to four small peptides (PDB access codes 2EVQ, 1BXX, 1BC5, and 2RD4) and to the tetrameric ion channel KcsA (PDB access code 1BL8).

As mentioned earlier, to determine the backbone-reduced representation, we limited our study to a fully extended peptide model made of 15 amino acids, i.e., $\beta$-Gly$_{15}$. That particular peptide sequence was chosen to minimize the interference between the central Gly residue Gly8 and the whole peptide structure as well as to get a nul charge on Gly8. The concept of "interference" is solely based on the CG description that can be obtained for various secondary structures. We indeed showed that the MEP-based clustering results are highly dependent on the peptide conformation.[37] For the studied pentadecapeptide $\beta$-Gly$_{15}$, end residues were not charged. At first, this may sound artificial, but the presence of a large negative or positive charge in the structure strongly affects the homogeneity of the CG distribution along the peptide chain.[37] Figure 1 illustrates the local extrema observed in the MEP of $\beta$-Gly$_{15}$ characterized by charged ends built from the Amber99 charges and smoothed at $t =$ 1.25 bohr$^2$ using eq 4. In this figure, one notices the presence of point charges in the close neighborhood of the C and O atoms of all residues but the two end ones. The two terminal AAs involve point charges on $NH_3^+$ and $COO^-$ only. The volume embedded by the negative and positive isocontours is also varying along the chain, increasing or decreasing toward $COO^-$, respectively. This reflects variations in the corresponding CG point charge values.

The structures of the isolated AAs involved the $(C\alpha{-}C{=}O)_{AA}(N{-}H)_{AA+1}$ backbone atoms so as to allow the merging of the $(C{=}O)_{AA}$ and $(N{-}H)_{AA+1}$ atoms, as observed in $\beta$-Gly$_{15}$. The consideration of isolated AAs is part of a strategy to favor CG models to approximate all-atom representations. That strategy was selected to reduce the mutual influence of the backbone atoms on the side chain descriptions. It was indeed observed, in a previous study on Gly$_7$−AA−Gly$_7$ structures,[37] that for AAs like Asp and Phe, the side chain CG representation is dependent on its conformation and on the presence of the backbone, respectively. Let us also mention that treating separately backbone and side chain CG descriptions was, e.g., applied by Cascella et al.[29] in their method to evaluate protein electrostatic potentials as summations over backbone dipolar and side chain multipolar contributions.

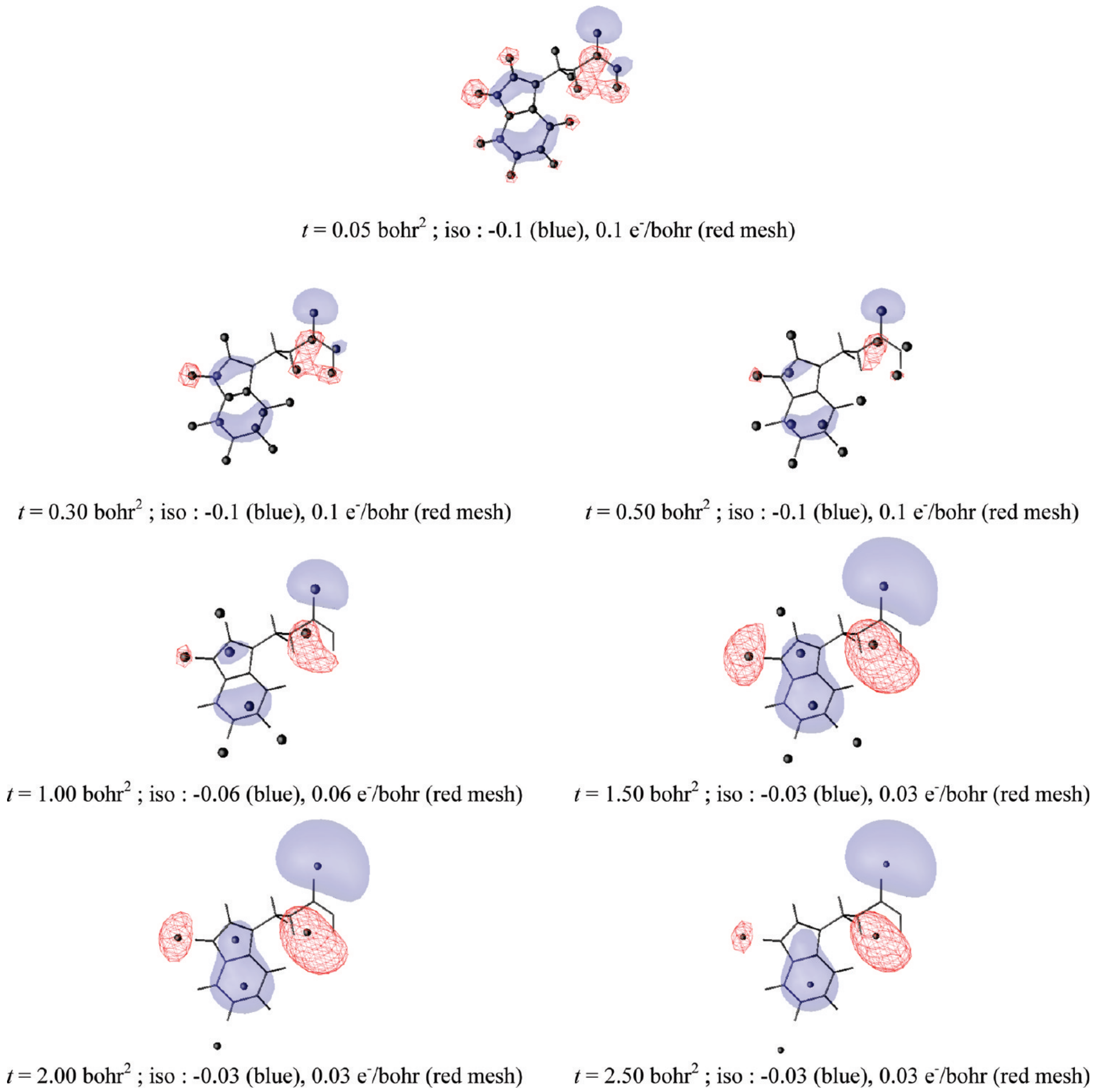
To generate the 3D structure of all AAs studied in this work, the simulated annealing (SA) procedure implemented in the program SMMP05[51,52] was applied to pentadecapeptide models, i.e., Gly$_7$−AA−Gly$_7$ structures, with $\Omega$, $\Phi$, $\Psi$, and $\chi$ dihedrals constrained to predefined values. The ECEPP/3 FF[53] and SA default running parameters were selected. Each SA run consisted in a first 100-step equilibration MC Metropolis stage carried out at 1 000 K. Then the procedure was continued for 50 000 MC Metropolis iterations until the final temperature, 100 K, was reached. The lowest potential energy structure generated during each run was kept. Isolated AA structures were then obtained by pruning the optimized pentadecapeptides.

The hierarchical decompositions of the molecular structures from MEP functions were carried out with the following parameters: $t = 0.05{-}3.0$ bohr$^2$, $\Delta_{init} = 10^{-4}$ bohr$^2$, grad$_{lim}$ $= 10^{-6}$ $e^-$/bohr$^2$.

**A. Selection of the Smoothing Degree.** As illustrated in Figure 2 for residue Trp, the CG description of an AA is dependent on the smoothing value $t$. At $t = 0.05$ bohr,$^2$ peaks and pits observed in the MEP are closely located on the atoms of the molecular structure. Starting at $t = 0.3$ bohr,$^2$ the extrema begin to move away from the atomic centers and their number decreases. At $t = 2.5$ bohr$^2$, there are only three extrema left on the side chain of the AA.

To select the optimal smoothing degree for the building of the reduced models, we used the charge-fitting algorithm QFIT[49] and applied it, with the same conditions as reported in Section II, to each set of peaks and pits obtained for the $\beta$-Gly$_{15}$ structure at various smoothing levels. The resulting minimal objective function (MOF) values are reported in Figure 3. The MOF function is built on the rmsdV and rmsd$\mu$ values defined in Section II. The best fittings, corresponding to a dipolar description of each AA backbone (Figure 4), were obtained at $t = 1.25$ and 1.3 bohr$^2$ for Amber99 and Gromos43A1, respectively, i.e., MOF = 1.76 and 0.48. For Amber99, the loss of one CG between $t = 1.25$ and 1.3 bohr$^2$ involves a steep rise in the MOF value, followed by a slower decrease observed up to $t = 1.9$ bohr.$^2$ Between $t = 1.5$ to 1.9 bohr,$^2$ the better fit is due only to a more adequate arrangement of peaks and pits, their number being constant, i.e., equal to 30 (Figure 3). For Gromos43A1, the MOF values are well below the corresponding values obtained with the Amber99 FF. This is due to the fact that Gromos43A1 is already a united-atom FF. Indeed, most of the atoms in alkyl groups, for instance, have a nul electric charge. Beyond $t = 1.3$ bohr,$^2$ the fitting is less and less efficient due to a progressive change in the location of the CGs with respect to the original structure. Models obtained for $\beta$-Gly$_{15}$ at $t =$ 1.25 and 1.3 bohr$^2$ for Amber99 and Gromos43A1, respectively, contain 32 and 31 CGs (Figure 4 and Table 1). In this sense, the application of the smoothing algorithm to the MEP function levels out the differences between the all-atom and united-atom FFs, but the CG charge values differ (Table 1), as explained in the next paragraph.

**B. Protein Backbone Modeling.** As announced hereabove, to generate a regular point charge distribution for the



$t$ = 0.05 bohr$^2$ ; iso : -0.1 (blue), 0.1 e$^-$/bohr (red mesh)

$t$ = 0.30 bohr$^2$ ; iso : -0.1 (blue), 0.1 e$^-$/bohr (red mesh)

$t$ = 0.50 bohr$^2$ ; iso : -0.1 (blue), 0.1 e$^-$/bohr (red mesh)

$t$ = 1.00 bohr$^2$ ; iso : -0.06 (blue), 0.06 e$^-$/bohr (red mesh)

$t$ = 1.50 bohr$^2$ ; iso : -0.03 (blue), 0.03 e$^-$/bohr (red mesh)

$t$ = 2.00 bohr$^2$ ; iso : -0.03 (blue), 0.03 e$^-$/bohr (red mesh)

$t$ = 2.50 bohr$^2$ ; iso : -0.03 (blue), 0.03 e$^-$/bohr (red mesh)

***Figure 2.*** Amber99 MEP isocontours of Trp in a conformational state corresponding to the *g*−,*g*−rotamer, smoothed at various values of *t*. Local maxima and minima (black spheres) were obtained using the hierarchical merging/clustering algorithm applied to the all-atom Amber99 MEP function.

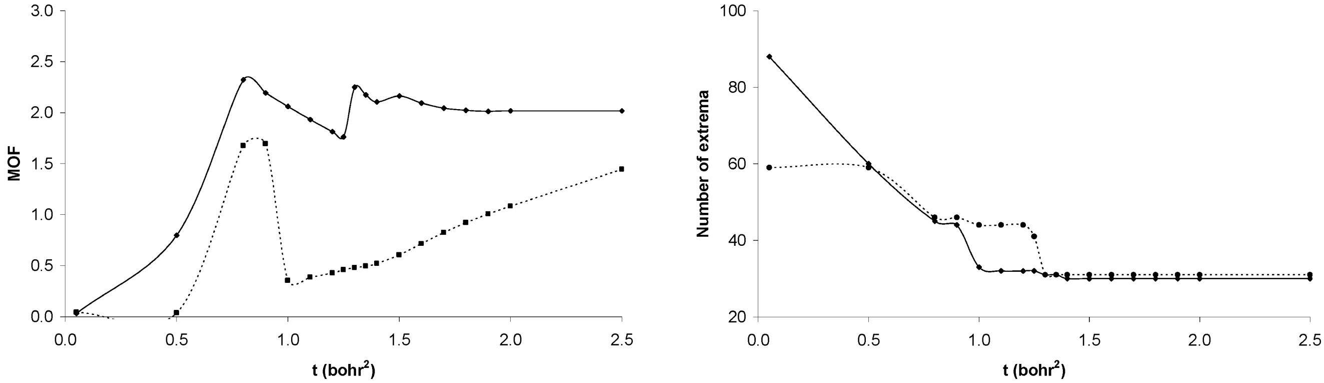


***Figure 3.*** (Left) The MOF of the charge fittings of $\beta$-Gly$_{15}$ CG points vs the unsmoothed Amber99 (plain line) and the Gromos43A1 (dashed line) MEP values. (Right) Number of local minima and maxima observed in the smoothed MEPs, as a function of the smoothing degree *t*.

backbone, an extended geometry characterized by $\Omega = 180°$, $\Phi = -139°$, and $\Psi = 135°$ was considered. Indeed, for MEP analyses, the conformation of the peptide appeared to be extremely important on the results of the merging/clustering algorithm applied to MEP functions.[37] Fitted CG charges of structure $\beta$-Gly$_{15}$, depicted in Figure 4, are reported in Table

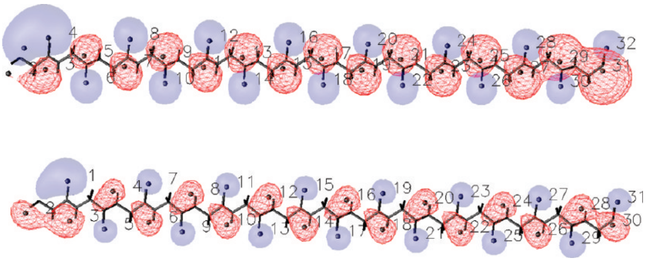


***Figure 4.*** MEP isocontours (blue plain surface: −0.03; red mesh: 0.03 $e^-$/bohr) of $\beta$-Gly$_{15}$ smoothed (top) at $t$ = 1.25 bohr$^2$ using Amber99 and (bottom) at $t$ = 1.3 bohr$^2$ using Gromos43A1. Local maxima and minima (black spheres) were obtained using the hierarchical merging/clustering algorithm applied to the original MEP functions. CG points are numbered as in Table 1.

***Table 1.*** CG Charges $q$ (in $e^-$) of $\beta$-Gly$_{15}$ Fitted vs the Unsmoothed Amber99 and Gromos43A1 MEP Grids Using the Program QFIT[a]

| | Amber99 | | | | Gromos43A1 | | | |
|---|---|---|---|---|---|---|---|---|
| no. | closest atom | | $d$ | $q$ | closest atom | | $d$ | $q$ |
| 1 | N | Gly1 | 0.978 | −0.1401 | O | Gly1 | 0.522 | −0.2470 |
| 2 | H | Gly1 | 1.038 | 0.0762 | H | Gly1 | 0.698 | 0.0885 |
| 3 | Cα | Gly1 | 1.104 | 0.2616 | H | Gly2 | 0.555 | 0.1695 |
| 4 | O | Gly1 | 0.538 | −0.2753 | H | Gly3 | 0.441 | 0.1959 |
| 5 | C | Gly2 | 0.769 | 0.2543 | O | Gly2 | 0.555 | −0.1772 |
| 6 | O | Gly2 | 0.545 | −0.2612 | H | Gly4 | 0.461 | 0.1750 |
| 7 | C | Gly3 | 0.803 | 0.2479 | O | Gly3 | 0.568 | −0.1828 |
| 8 | O | Gly3 | 0.560 | −0.2452 | H | Gly5 | 0.463 | 0.1794 |
| 9 | C | Gly4 | 0.794 | 0.2421 | O | Gly4 | 0.565 | −0.1778 |
| 10 | O | Gly4 | 0.556 | −0.2427 | H | Gly6 | 0.464 | 0.1769 |
| 11 | C | Gly5 | 0.798 | 0.2408 | O | Gly5 | 0.567 | −0.1781 |
| 12 | O | Gly5 | 0.559 | −0.2413 | H | Gly7 | 0.465 | 0.1786 |
| 13 | C | Gly6 | 0.796 | 0.2428 | O | Gly6 | 0.566 | −0.1773 |
| 14 | O | Gly6 | 0.558 | −0.2425 | H | Gly8 | 0.465 | 0.1766 |
| 15 | C | Gly7 | 0.797 | 0.2430 | O | Gly7 | 0.567 | −0.1783 |
| 16 | O | Gly7 | 0.558 | −0.2435 | H | Gly9 | 0.465 | 0.1785 |
| 17 | C | Gly8 | 0.796 | 0.2440 | O | Gly8 | 0.567 | −0.1771 |
| 18 | O | Gly8 | 0.558 | −0.2439 | H | Gly10 | 0.466 | 0.1771 |
| 19 | C | Gly9 | 0.797 | 0.2450 | O | Gly9 | 0.567 | −0.1782 |
| 20 | O | Gly9 | 0.559 | −0.2445 | H | Gly11 | 0.465 | 0.1777 |
| 21 | C | Gly10 | 0.796 | 0.2417 | O | Gly10 | 0.567 | −0.1773 |
| 22 | O | Gly10 | 0.558 | −0.2433 | H | Gly12 | 0.468 | 0.1776 |
| 23 | C | Gly11 | 0.796 | 0.2479 | O | Gly11 | 0.567 | −0.1777 |
| 24 | O | Gly11 | 0.558 | −0.2426 | H | Gly13 | 0.466 | 0.1754 |
| 25 | C | Gly12 | 0.796 | 0.2343 | O | Gly12 | 0.567 | −0.1764 |
| 26 | O | Gly12 | 0.558 | −0.2448 | H | Gly14 | 0.468 | 0.1766 |
| 27 | C | Gly13 | 0.796 | 0.2588 | O | Gly13 | 0.567 | −0.1729 |
| 28 | O | Gly13 | 0.560 | −0.2454 | H | Gly15 | 0.436 | 0.1586 |
| 29 | C | Gly14 | 0.786 | 0.2580 | O | Gly14 | 0.576 | −0.1802 |
| 30 | O | Gly14 | 0.565 | −0.2574 | C | Gly15 | 0.542 | 0.1637 |
| 31 | C | Gly15 | 0.585 | 0.3158 | O | Gly15 | 0.665 | −0.1673 |
| 32 | O | Gly15 | 0.666 | −0.2406 | | | | |
| rmsdV | | | | 1.33 | | | | 0.69 |
| rmsd$\mu$ | | | | 0.16 | | | | 0.15 |

[a] Local maxima and minima at $t$ = 1.25 and 1.3 bohr$^2$, respectively, were obtained using the hierarchical merging/clustering algorithm applied to the original Amber99 and Gromos43A1 MEP functions. For each point, the distance vs the closest atom, $d$, is given in Å. RmsdV and rmsd$\mu$ are given in kcal/mol and D, respectively. Point numbers (no.) refer to Figure 4.

1. For Amber99, positive and negative charges located near the C and O atoms of the central residue Gly8 are equal to ± 0.244 $e^-$ and are separated by a distance of 2.52 Å. rmsdV and rmsd$\mu$ values are equal to 1.33 kcal/mol and 0.16 D. For Gromos43A1, the two-site CG description of each AA backbone differs. Rather than being located along the C=O axis of a residue, as in the case of Amber99, it is displaced such as the positive charge is closer to the H atom of the neighboring residue (Table 1), and the two opposite charges, equal to ± 0.177 $e^-$, are separated by a distance of 3.66 Å.

Our CG models are, thus, of an intermediate description level between representations that involve only one grain per AA backbone, like in the MARTINI[20,21] and the Basdevant's[6] models, and finer descriptions that allow to more precisely account for the various secondary structure elements of a protein.[9] A dipolar representation of the backbone of AAs will appear to be useful in applications where the dipolar character of AAs is important, as further illustrated in the KcsA case.

**C. Protein Side Chain Modeling.** CG representations of each of the 20 AA side chains were obtained by considering the AAs in specific conformational states. Except for AA = Gly and Ala, most recurrent rotamers were generated by taking into account the angular constraints given in Table 2. These rotamers were selected according to their occurrence degree in protein structures as reported in the Structural Library of Intrinsic Residue Propensities (SLIRP).[54,55] As already mentioned, from the pentadecapeptide chains $\beta$-Gly$_7$−AA−Gly$_7$ generated using SMMP05,[51,52] only the central AA residue was kept with backbone atoms $(C\alpha{-}C{=}O)_{AA}(N{-}H)_{AA+1}$. This was achieved to avoid the generation of side chain CGs that might depend on a particular secondary structure motif. As already specified above, we considered the following protonation states: Lys(+1), Arg(+1), Glu(−1), and Asp(−1). For Gln, it appeared that both specific conformations first selected to represent classes $g-$, $t$, N$g+$ and $g-$, $t$, O$g+$ led, through the program SMMP05, to an identical 3D structure. We, thus, kept only one structure, $g-$, $t$, O$g+$ and summed over the two initial weights reported in SLIRP to get a value of 28.6. Similarly, Gln conformations representing classes $g-$, $t$, O$g-$ and $g-$, $t$, N$g-$ led to only one rotamer, with a total weight of 33.2. This occurred for another AA, His, for which two conformers, depicting classes $g-$, N$g-$ and $g-$, C$g-$, are characterized by a total weight of 35.8.

In a further step, we determined the charge values for the CG descriptions of each AA through a fitting procedure carried out using QFIT[49] vs unsmoothed all-atom MEP grids. In this procedure, and for each of the AAs, all rotamer descriptions in terms of peaks and pits observed in the Amber99 and Gromos43A1 MEPs, smoothed at $t$ = 1.25 and 1.3 bohr$^2$, respectively, were considered according to their occurrence probability (Table 2). This step was carried out in four stages. First, isolated AA structures were assigned atom charges using PDB2PQR.[40,41] Side chain extrema were located using our merging/clustering algorithm. Second, the corresponding charge values were fitted vs the all-atom MEP generated from the side chain atoms only. Third, the backbone CGs were added in accordance with the motif found for Gly8 in $\beta$-Gly$_{15}$, and fourth, a second charge-fitting procedure, now carried out vs the MEP calculated using all the AA atoms, was achieved to determine the charge values

***Table 2.*** Geometrical Parameters and Occurrence Probability of the Selected AA Side Chain Rotamers[54,55] with the Exception of Ala and Gly[a]

| | conformation | $\chi$1 (°) | $\chi$2 (°) | $\chi$3 (°) | $\chi$4 (°) | occurrence (%) |
|---|---|---|---|---|---|---|
| Arg | *g*−, *t*, *g*−, *g*− | 300 | 180 | 300 | 300 | 9.5 |
| | *g*−, *t*, *g*−, *t* | 300 | 180 | 300 | 180 | 11.9 |
| | *g*−, *t*, *g*+, *t* | 300 | 180 | 60 | 180 | 12.2 |
| | *g*−, *t*, *t*, *t* | 300 | 180 | 180 | 180 | 12.2 |
| Asn | *t*, *Nt* | 180 | 0 | | | 11.1 |
| | *t*, *Og*− | 180 | 300 | | | 21.3 |
| | *t*, *Og*+ | 180 | 60 | | | 23.6 |
| Asp | *t*, *g*+ | 180 | 60 | | | 62.8 |
| Cys | *g*− | 300 | | | | 56.3 |
| | *g*+ | 60 | | | | 15.1 |
| | *t* | 180 | | | | 28.7 |
| Gln | *g*−, *t*, *Nt* | 300 | 180 | 0 | | 11.2 |
| | *g*−, *t*, *Og*− | 300 | 180 | 300 | | 33.2 |
| | *g*−, *t*, *Og*+ | 300 | 180 | 60 | | 28.6 |
| Glu | *g*−, *t*, *g*− | 300 | 180 | 120 | | 29.9 |
| | *g*−, *t*, *g*+ | 300 | 180 | 60 | | 25.3 |
| His | *g*−, *Ng*− | 300 | 300 | | | 35.8 |
| | *t*, *Ng*+ | 180 | 60 | | | 15.0 |
| Ile | *g*−, *g*− | 300 | 300 | | | 22.7 |
| | *g*−, *t* | 300 | 180 | | | 28.3 |
| | *g*+, *t* | 60 | 180 | | | 42.5 |
| Leu | *g*−, *t* | 300 | 180 | | | 65.2 |
| | *t*, *g*+ | 180 | 60 | | | 24.1 |
| Lys | *g*−, *g*−, *t*, *g*− | 300 | 300 | 180 | 300 | 8.5 |
| | *g*−, *g*−, *t*, *g*+ | 300 | 300 | 180 | 60 | 6.5 |
| | *g*−, *t*, *t*, *g*− | 300 | 180 | 180 | 300 | 21.7 |
| | *g*−, *t*, *t*, *g*+ | 300 | 180 | 180 | 60 | 14.3 |
| Met | *g*−, *g*−, *g*− | 300 | 300 | 300 | | 15.5 |
| | *g*−, *g*−, *t* | 300 | 300 | 180 | | 11.6 |
| | *g*−, *t*, *g*− | 300 | 180 | 300 | | 19.4 |
| | *g*−, *t*, *g*+ | 300 | 180 | 60 | | 16.4 |
| | *g*−, *t*, *t* | 300 | 180 | 180 | | 15.4 |
| Phe | *g*−, *g*− | 300 | 300 | | | 37.8 |
| | *t*, *g*+ | 180 | 60 | | | 31.5 |
| Pro | *g*+ | 0 | | | | 66.8 |
| Ser | *g*− | 300 | | | | 73.1 |
| | *g*+ | 30 | | | | 24.8 |
| Thr | *g*− | 300 | | | | 51.6 |
| | *g*+ | 30 | | | | 46.3 |
| Trp | *g*−, *g*− | 300 | 90 | | | 28.2 |
| | *g*−, *t* | 300 | 0 | | | 16.5 |
| | *t*, *g*− | 180 | 60 | | | 11.6 |
| | *t*, *g*+ | 180 | 300 | | | 13.8 |
| | *t*, *t* | 180 | 0 | | | 11.2 |
| Tyr | *g*−, *g*− | 300 | 120 | | | 38.3 |
| | *t*, *g*+ | 180 | 60 | | | 31.7 |
| Val | *g*− | 300 | | | | 46.4 |
| | *t* | 180 | | | | 51.9 |

[a] *g* and *t* stand for *gauche* and *trans*, respectively.

of the two backbone CGs, while preserving the side chain CG charge values first obtained.

It is to be specified that, for some AA residues, the initial peak/pit-based CG representation obtained for their side chain was replaced by a simpler model consisting of one point centered on a selected atom, as detailed below. This was achieved as a first stage in the easy design of a CG protein model from its atom coordinates retrieved from the PDB,[56,57] as in ref 37.

In Figures 5 and 6, we report the so-obtained original or simplified CG representations for the 20 AA residues as derived from the results of our hierarchical merging/clustering algorithm applied to the Amber99 and Gromos43A1 MEP functions, smoothed at $t$ = 1.25 and 1.3 bohr$^2$, respectively. Corresponding CG charges and deviations of the electrostatic properties vs the all-atom ones are reported in the tables provided in the Supporting Information. With Amber99 (Figure 5), all noncyclic C−H based residues, i.e., Ala, Ile, Leu, and Val, have no side chain points. This was chosen because of the low charge values obtained initially for their CGs and was an easy way to model those specific residues in possible MM applications. For Lys, we also simplified the model by setting the positive charge exactly on the Nς atom (point three). For all other AAs, the original point locations observed in the smoothed MEP functions were kept for the charge-fitting procedures. As illustrated in Figure 5, we note that for hydroxyl-containing residues, i.e., Ser, Thr, and Tyr, there are two charges located near but not exactly on the O and H atoms (points three and four for Ser and Thr; points six and seven for Tyr). A similar representation is obtained for the sulfur-containing residues with a charge close to the S atom and a charge in the neighborhood of the H atom (point four for Cys) or $CH_3$ group (point four for Met). For the negatively charged residues, i.e., Asp and Glu, each carboxylate functional group leads to two negative charges located near the O atoms (points three and four). Positively charged residues, Arg and Lys, present different behaviors. While the side chain of Lys leads to only one positive charge value (point three), the Arg side chain is characterized by a four-point motif (points three to six), wherein each charge is somewhat symmetrically located on the bisectors of each of the three N−C−N angles of the guanidinium group.

Regarding side chain descriptions obtained using Gromos43A1 (Figure 6), alkyl chains such as Ala, Ile, Leu, and Val do not involve any CG. This is also observed for Met. Indeed, for these AAs, the only charged atoms in the united-atom model are the backbone N, H, C, and O atoms. The CG description of Arg differs from the Amber99-based representation, as there is only one positive charge initially located in the neighborhood of the atom Cς. We have simplified the Arg CG model by fixing that CG point exactly on Cς (point three).

Let us additionally mention that for Asp, Glu, and Phe, an identity in the charge values was imposed such as $q_3 = q_4$, $q_3 = q_4$, and $q_4 = q_6$, respectively, as reported in Tables 3 and 4. One also directly notices that some charges are displaced toward the outer part of some AAs, as in Phe, where H atoms seem to be associated with charges located away from the side chain. Even if this, at first, looks unnatural, it was nevertheless decided to keep such original charge distributions, as they correspond to real topological features of the smoothed MEP functions.

A comparison between two models reported in literature and our two MEP-based CG models of the AA side chains generated from the Amber99 and Gromos43A1 sets of charges is reported in Table 5. AA residues are listed according to their properties defined in the MARTINI FF,[21] i.e., hydrophobic residues, mainly classified as apolar, polar residues with or without hydrogen-bond-forming characteristics and charged side chains. Such a description, known to lack an electrostatic contribution, is nevertheless interesting to compare with as it involves the concept of polarity. It is

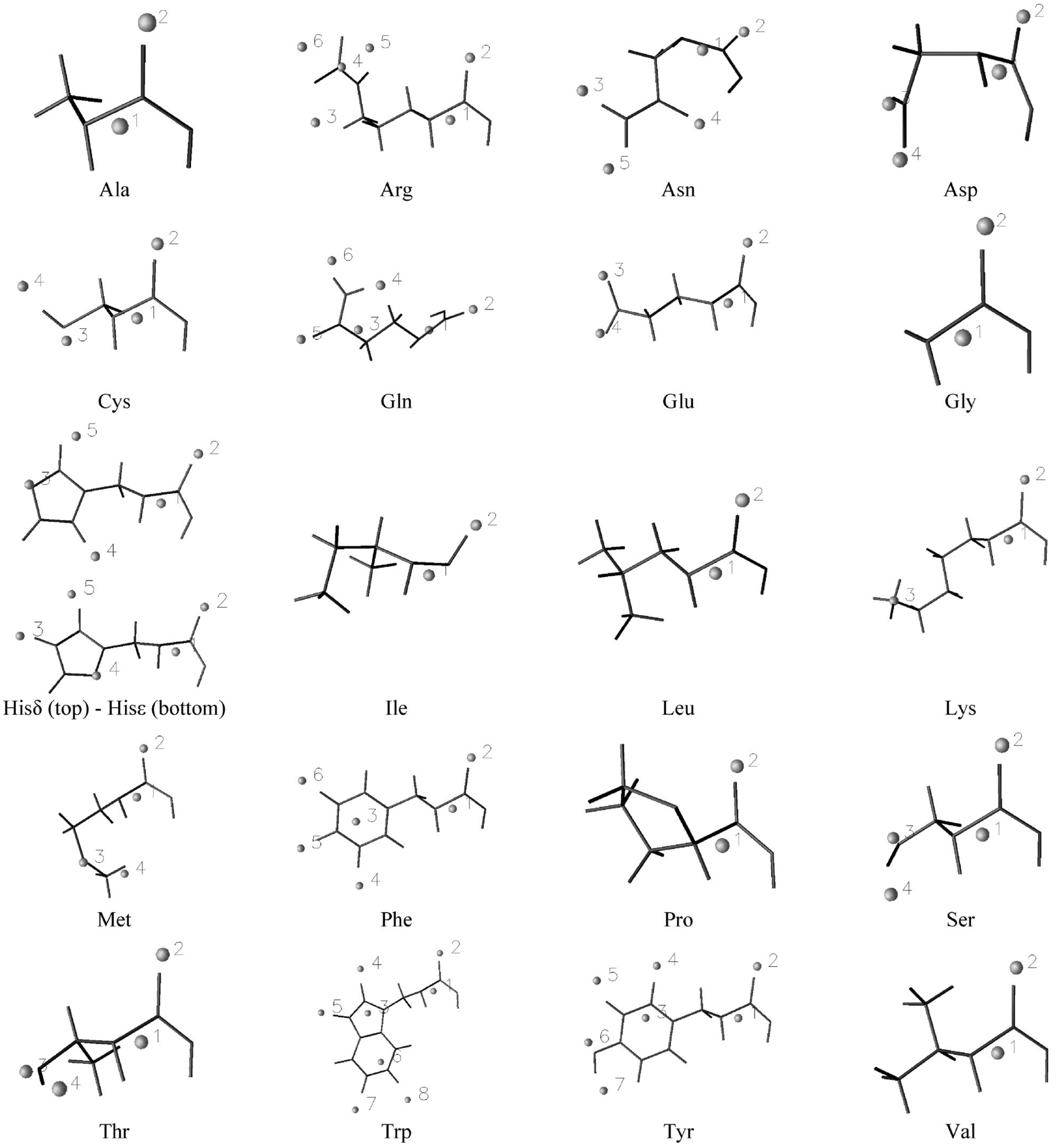


***Figure 5.*** CG model for each of the 20 AA residues as established at $t = 1.25$ bohr$^2$ from the hierarchical merging/clustering algorithm applied to the all-atom Amber99 MEP function. CG points are numbered as in the Supporting Information.

also one of the few descriptions that is easily available in the literature for all AAs. Parallely, we report a description of the Basdevant's model,[6] which is interesting as it is close to representations based on ED maxima described earlier.[36,58] For all residues, the backbone CG representation consists either of one polar bead (MARTINI), one center (Basdevant), or two CGs with opposite charges (our models).

The total number of side chain CGs in each model is variable. The number of grains in the Basdevant's model is strongly dependent on the size of the side chains but does not exceed two. For MARTINI, it is higher than two only for ring-shaped side chains, i.e., Phe, His, Trp, and Tyr. In the case of our MEP-based CG representations, there are up to six CGs for Trp. For all small hydrophobic residues, the MARTINI CG representations involve only one apolar grain. Parallely, in both MEP-based models, there is no side chain CGs. For Phe, our models involve a large number of points, i.e., four for both Amber99 and Gromos43A1. The charge brought by each of the side chain CGs of Phe stays low, with $|q| < 0.10$ e$^-$. Sulfur-containing residues, especially Cys, that are hydrophobic and do not form any H-bond, are however characterized by a dipole moment. In MARTINI, they are, thus, represented by one CG with the intermediate apolar/polar state. For the Amber99- and Gromos43A1-based models, there are two CGs with opposite charges. Regarding Asn and Gln, our MEP-based models provide a finer description of the side chains, with three grains located at the vicinity of the O and H atoms (Figures 5 and 6). In MARTINI, these side chains are represented using one grain characterized by a polar type with a hydrogen-bonding donor and acceptor character. For all residues containing an O−H group, i.e., Ser, Thr, and Tyr, our models include at least two opposite charges located in the neighborhood of O and H; they correspond to one polar group in MARTINI. The side chains of His and Trp not only contain hydrophobic rings but also hydrogen-bonding properties. In the framework of our MEP-based models, they are represented by CGs with a dipole occurring between HN$\delta$ and N$\varepsilon$ in His$\delta$, between

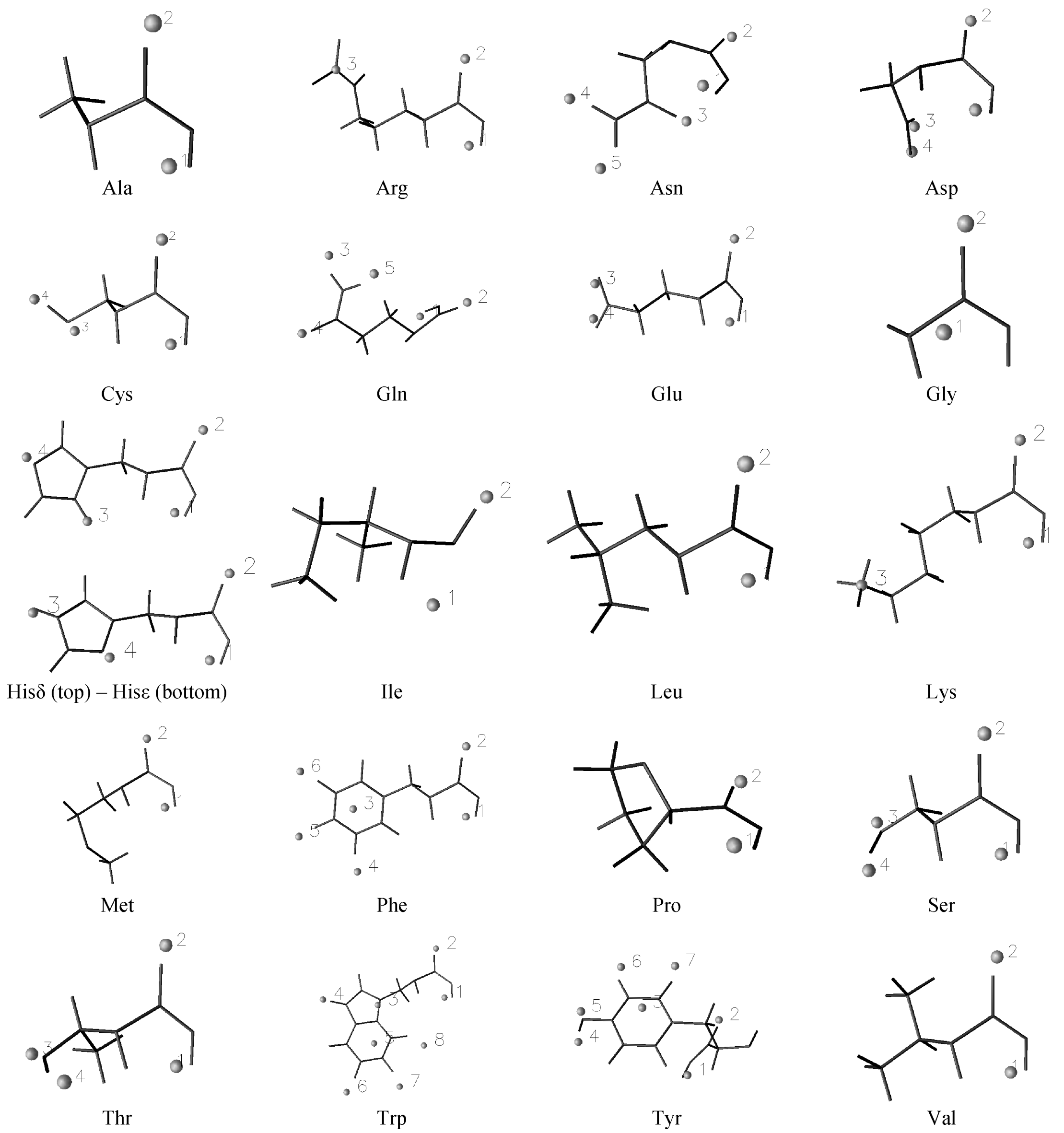


***Figure 6.*** CG model for each of the 20 AA residues as established at $t = 1.3$ bohr$^2$ from the hierarchical merging/clustering algorithm applied to the Gromos43A1 MEP function. CG points are numbered as in the Supporting Information.

HNε and Nδ in Hisε, and between HNε1 and the rings in Trp. Details regarding point charge locations and values are given in the Supporting Information. The polarity property in MARTINI is, thus, expressed as a charge separation in our models. Finally, regarding the residues that are explicitely charged in the MARTINI FF, we observe a finer description of the negative Asp and Glu residues in our models, with two separate negative charges close to the O of the carboxyl group. The Amber99-based CG model of Arg is rather interesting and original as it involves three positive charges almost symmetrically spread around the atom Cς, itself associated with a fourth CG charge (Figure 5). This might be seen as a description that is more consistent with a charge delocalization.

Thus, on the average, one can consider that there is a reduction ratio of about 4.5/1 between the CG and all-atom MEP-based models. Knowing that the calculation time evolves according to $N^2$ and $N\log N$, $N$ being the number of particles for the Coulomb pair potential and the particle mesh Ewald (PME) algorithm, respectively, one expects reduction ratios of CPU times of about 20 for the Coulomb potential and, for example, about 5 for the PME routine in the case of a protein with 100 000 atoms vs the all-atom representation.

**D. Automated CG Generation Procedure.** To study systems that are larger than oligopeptides, an automation stage was developed to avoid the lengthy generation of the CGs for each AA separately, as first carried out.[37] The resulting automated procedure was fully based on the application of a superimposition algorithm of CG motif templates of each AA onto the corresponding AA structures of the protein under study. We used the program QUATFIT[59,60] to, first, superimpose a limited set of atoms from the template on the studied structure and then used the resulting transformation matrix to generate the corresponding CG coordinates.

The templates that were selected in this study are described in Tables 3 and 4 for the Amber99 and Gromos43A1 FFs, respectively. Their size consisted of at least three atoms so

***Table 3.*** Template Coordinates (in Å) and Charges (in e⁻) as Used for the Amber99-Based CG Generation

| | *X* | *Y* | *Z* | charge | | *X* | *Y* | *Z* | charge |
|---|---|---|---|---|---|---|---|---|---|
| backbone | | | | | | | | | |
| C | 22.575 | 13.923 | 2.131 | | | | | | |
| O | 23.021 | 13.167 | 2.993 | | | | | | |
| N | 23.318 | 14.688 | 1.345 | | | | | | |
| PT1 | 21.839 | 14.199 | 1.699 | *q*1[a] | | | | | |
| PT2 | 23.280 | 12.855 | 3.318 | *q*2[a] | | | | | |
| side chain | | | | | | | | | |
| **ARG** | | | | | | | | | |
| Nε | 18.561 | 15.333 | 5.213 | | **HISε** | | | | |
| Cς | 18.087 | 14.123 | 5.542 | | Cγ | 18.921 | 15.161 | 2.698 | |
| NH1 | 17.049 | 13.606 | 4.871 | | Nδ1 | 18.437 | 15.752 | 1.543 | |
| NH2 | 18.651 | 13.432 | 6.542 | | Cε1 | 17.116 | 15.812 | 1.626 | |
| PT3 | 16.526 | 14.616 | 3.638 | 0.2780 | Nε2 | 16.744 | 15.267 | 2.820 | |
| PT4 | 18.195 | 14.221 | 5.594 | 0.0555 | Cδ2 | 17.833 | 14.874 | 3.469 | |
| PT5 | 19.530 | 14.832 | 6.483 | 0.4811 | PT3 | 15.177 | 15.203 | 3.248 | 0.1803 |
| PT6 | 17.112 | 12.282 | 5.831 | 0.2215 | PT4 | 18.512 | 15.718 | 1.511 | −0.2142 |
| **ASN** | | | | | PT5 | 17.525 | 14.157 | 5.209 | 0.0699 |
| Cγ | 21.154 | 16.268 | 3.071 | | **PHE** | | | | |
| Oδ1 | 22.355 | 16.412 | 3.224 | | Cγ | 18.926 | 14.982 | 2.956 | |
| Nδ2 | 20.298 | 17.275 | 2.917 | | Cδ1 | 18.399 | 15.648 | 1.894 | |
| HNδ21 | 20.637 | 18.192 | 2.709 | | Cε1 | 16.993 | 15.816 | 1.788 | |
| HNδ22 | 19.315 | 17.114 | 3.009 | | Cς | 16.173 | 15.310 | 2.748 | |
| PT3 | 18.536 | 16.868 | 2.689 | 0.1470 | Cε2 | 16.700 | 14.643 | 3.810 | |
| PT4 | 22.863 | 16.544 | 3.245 | −0.2340 | Cδ2 | 18.106 | 14.476 | 3.916 | |
| PT5 | 20.335 | 19.138 | 2.987 | 0.0820 | PT3 | 17.237 | 15.111 | 2.852 | −0.0753 |
| **ASP** | | | | | PT4 | 16.307 | 16.902 | 0.127 | 0.0425 |
| Cγ | 21.094 | 16.293 | 3.152 | | PT5 | 14.140 | 15.609 | 2.572 | 0.0093 |
| Oδ1 | 20.959 | 17.047 | 2.164 | | PT6 | 15.393 | 13.922 | 5.289 | 0.0425 |
| Oδ2 | 21.670 | 16.583 | 4.223 | | **SER** | | | | |
| PT3 | 21.800 | 16.833 | 4.393 | −0.4290 | Cβ | 20.443 | 14.916 | 2.987 | |
| PT4 | 21.037 | 17.390 | 2.124 | −0.4290 | Oγ | 19.047 | 15.112 | 2.779 | |
| **CYS** | | | | | Hγ | 18.754 | 14.579 | 1.988 | |
| Cβ | 20.432 | 14.938 | 2.963 | | PT3 | 18.804 | 15.397 | 3.193 | −0.0997 |
| Sγ | 18.650 | 15.161 | 2.610 | | PT4 | 18.739 | 14.433 | 1.036 | 0.1547 |
| Hγ | 18.127 | 14.540 | 3.187 | | **THR** | | | | |
| PT3 | 18.679 | 15.325 | 2.115 | −0.0705 | Cβ | 20.358 | 14.870 | 2.971 | |
| PT4 | 17.875 | 13.550 | 3.798 | 0.0515 | Oγ1 | 19.001 | 14.913 | 2.536 | |
| **GLN** | | | | | Hγ1 | 18.880 | 14.363 | 1.701 | |
| Cγ | 18.930 | 15.114 | 2.699 | | PT3 | 18.660 | 15.285 | 2.856 | −0.1157 |
| Cδ | 18.288 | 16.002 | 3.767 | | PT4 | 19.161 | 13.795 | 1.024 | 0.1682 |
| Oε1 | 17.102 | 16.285 | 3.744 | | **TRP** | | | | |
| Nε2 | 19.135 | 16.423 | 4.701 | | Cγ | 18.914 | 15.167 | 2.725 | |
| HNε21 | 19.656 | 15.761 | 5.240 | | Cδ1 | 17.884 | 14.443 | 3.185 | |
| HNε22 | 19.252 | 17.403 | 4.866 | | Nε1 | 16.676 | 14.963 | 2.769 | |
| PT3 | 18.923 | 15.585 | 3.173 | 0.1768 | Cε2 | 16.950 | 16.087 | 1.998 | |
| PT4 | 20.178 | 15.240 | 5.492 | 0.0958 | Cς2 | 16.065 | 16.958 | 1.352 | |
| PT5 | 16.632 | 16.505 | 3.781 | −0.3215 | CH2 | 16.638 | 18.008 | 0.644 | |
| PT6 | 19.738 | 18.140 | 5.034 | 0.0818 | Cς3 | 18.021 | 18.142 | 0.611 | |
| **GLU** | | | | | Cε3 | 18.919 | 17.280 | 1.251 | |
| Cδ | 18.288 | 16.002 | 3.767 | | Cδ2 | 18.322 | 16.220 | 1.965 | |
| Oε1 | 17.754 | 17.063 | 3.377 | | PT3 | 17.891 | 15.125 | 2.830 | −0.1380 |
| Oε2 | 18.345 | 15.599 | 4.949 | | PT4 | 18.246 | 12.882 | 4.318 | 0.0963 |
| PT3 | 18.270 | 15.634 | 5.269 | −0.4410 | PT5 | 15.128 | 14.269 | 3.159 | 0.1002 |
| PT4 | 17.622 | 17.347 | 3.562 | −0.4410 | PT6 | 17.745 | 17.637 | 0.978 | −0.0640 |
| **HISδ** | | | | | PT7 | 15.309 | 19.357 | −0.386 | 0.0284 |
| Cγ | 18.921 | 15.161 | 2.698 | | PT8 | 18.696 | 19.982 | −0.699 | 0.0068 |
| Nδ1 | 18.437 | 15.752 | 1.543 | | **TYR** | | | | |
| Cε1 | 17.116 | 15.812 | 1.626 | | Cς | 16.164 | 15.242 | 2.786 | |
| Nε2 | 16.744 | 15.267 | 2.820 | | OH | 14.815 | 15.390 | 2.691 | |
| Cδ2 | 17.833 | 14.874 | 3.469 | | HH | 14.590 | 15.887 | 1.852 | |
| PT3 | 16.613 | 15.212 | 2.909 | −0.2306 | Cδ1 | 18.097 | 14.402 | 3.947 | |
| PT4 | 19.443 | 16.282 | 0.260 | 0.1863 | Cε1 | 16.669 | 14.558 | 3.847 | |
| PT5 | 18.531 | 14.300 | 4.885 | 0.0523 | Cε2 | 16.946 | 15.786 | 1.815 | |
| **MET** | | | | | Cδ2 | 18.374 | 15.629 | 1.915 | |
| Cγ | 18.937 | 15.046 | 2.800 | | PT3 | 17.673 | 14.741 | 3.465 | 0.0406 |
| Sδ | 18.639 | 15.869 | 1.245 | | PT4 | 18.866 | 13.489 | 5.441 | 0.0498 |
| Cε | 19.506 | 17.402 | 1.535 | | PT5 | 15.831 | 13.569 | 5.556 | 0.0115 |
| PT3 | 18.579 | 15.903 | 1.151 | −0.0654 | PT6 | 14.650 | 15.117 | 3.194 | −0.1535 |
| PT4 | 20.535 | 18.072 | 2.000 | 0.1154 | PT7 | 14.781 | 16.345 | 1.047 | 0.1610 |

[a] Values of *q*1 and *q*2 depend on the AA type (Table SI1 in the Supporting Information).

***Table 4.*** Template Coordinates (in Å) and Charges (in e⁻) as Used for the Gromos43A1-Based CG Generation

| | *X* | *Y* | *Z* | charge | | *X* | *Y* | *Z* | charge |
|---|---|---|---|---|---|---|---|---|---|
| backbone | | | | | | | | | |
| C | 22.575 | 13.923 | 2.131 | | | | | | |
| O | 23.021 | 13.167 | 2.993 | | | | | | |
| N | 23.318 | 14.688 | 1.345 | | | | | | |
| PT1 | 22.478 | 15.192 | 0.748 | *q*1[a] | | | | | |
| PT2 | 23.357 | 12.902 | 3.371 | *q*2[a] | | | | | |
| side chain | | | | | | | | | |
| **ASN** | | | | | **PHE** | | | | |
| Cγ | 21.154 | 16.268 | 3.071 | | Cδ1 | 18.399 | 15.648 | 1.894 | |
| Oδ1 | 22.355 | 16.412 | 3.224 | | Cε1 | 16.993 | 15.816 | 1.788 | |
| Nδ2 | 20.298 | 17.275 | 2.917 | | Cς | 16.173 | 15.310 | 2.748 | |
| HNδ21 | 20.637 | 18.192 | 2.709 | | Cε2 | 16.700 | 14.643 | 3.810 | |
| HNδ22 | 19.315 | 17.114 | 3.009 | | Cδ2 | 18.106 | 14.476 | 3.916 | |
| PT3 | 22.823 | 16.442 | 3.455 | −0.1910 | PT3 | 17.269 | 15.144 | 2.896 | −0.0936 |
| PT4 | 18.479 | 17.146 | 2.969 | 0.1104 | PT4 | 16.392 | 16.836 | 0.108 | 0.0386 |
| PT5 | 20.388 | 19.047 | 2.427 | 0.0840 | PT5 | 14.145 | 15.597 | 2.521 | 0.0163 |
| **ASP** | | | | | PT6 | 15.379 | 13.928 | 5.206 | 0.0386 |
| Cγ | 21.094 | 16.293 | 3.152 | | **SER** | | | | |
| Oδ1 | 20.959 | 17.047 | 2.164 | | Cβ | 20.443 | 14.916 | 2.987 | |
| Oδ2 | 21.670 | 16.583 | 4.223 | | Oγ | 19.047 | 15.112 | 2.779 | |
| PT3 | 21.645 | 16.781 | 4.010 | −0.5000 | Hγ | 18.568 | 14.235 | 2.826 | |
| PT4 | 21.142 | 17.151 | 2.618 | −0.5000 | PT3 | 20.331 | 14.350 | 4.599 | −0.1466 |
| **CYS** | | | | | PT4 | 18.819 | 12.975 | 3.200 | 0.1466 |
| Cβ | 20.432 | 14.938 | 2.963 | | **THR** | | | | |
| Sγ | 18.650 | 15.161 | 2.610 | | Cβ | 20.358 | 14.870 | 2.971 | |
| Hγ | 18.127 | 14.540 | 3.187 | | Oγ1 | 19.001 | 14.913 | 2.536 | |
| PT3 | 18.642 | 15.763 | 2.474 | −0.0299 | Hγ1 | 18.880 | 14.363 | 1.701 | |
| PT4 | 17.761 | 14.191 | 3.407 | 0.0299 | PT3 | 18.739 | 15.245 | 3.017 | −0.1459 |
| **GLN** | | | | | PT4 | 19.133 | 13.940 | 0.949 | 0.1459 |
| Cδ | 18.288 | 16.002 | 3.767 | | **TRP** | | | | |
| Oε1 | 17.102 | 16.285 | 3.744 | | Cγ | 18.914 | 15.167 | 2.725 | |
| Nε2 | 19.135 | 16.423 | 4.701 | | Cδ1 | 18.147 | 14.690 | 1.736 | |
| HNε21 | 19.656 | 15.761 | 5.240 | | Nε1 | 16.840 | 15.119 | 1.858 | |
| HNε22 | 19.252 | 17.403 | 4.866 | | Cε2 | 16.769 | 15.917 | 2.994 | |
| PT3 | 19.737 | 18.118 | 5.180 | 0.0850 | Cς2 | 15.668 | 16.584 | 3.544 | |
| PT4 | 16.715 | 16.386 | 3.841 | −0.2031 | CH2 | 15.904 | 17.316 | 4.702 | |
| PT5 | 20.244 | 15.453 | 5.740 | 0.1182 | Cς3 | 17.183 | 17.352 | 5.245 | |
| **GLU** | | | | | Cε3 | 18.294 | 16.690 | 4.708 | |
| Cδ | 18.288 | 16.002 | 3.767 | | Cδ2 | 18.038 | 15.953 | 3.534 | |
| Oε1 | 17.754 | 17.063 | 3.377 | | PT3 | 18.602 | 15.283 | 2.908 | −0.1232 |
| Oε2 | 18.345 | 15.599 | 4.949 | | PT4 | 15.743 | 14.810 | 0.939 | 0.1553 |
| PT3 | 18.190 | 15.902 | 4.841 | −0.5000 | PT5 | 17.181 | 16.729 | 4.389 | −0.1409 |
| PT4 | 17.773 | 16.922 | 3.755 | −0.5000 | PT6 | 14.277 | 18.300 | 5.454 | 0.0466 |
| **HISδ** | | | | | PT7 | 17.248 | 18.565 | 6.949 | 0.0319 |
| Cε1 | 17.116 | 15.812 | 1.626 | | PT8 | 19.915 | 17.376 | 5.998 | 0.0303 |
| Nε2 | 16.744 | 15.267 | 2.820 | | **TYR** | | | | |
| Cδ2 | 17.833 | 14.874 | 3.469 | | Cς | 21.930 | 18.777 | 3.647 | |
| PT3 | 18.980 | 16.087 | 0.718 | 0.2623 | OH | 22.401 | 20.040 | 3.824 | |
| PT4 | 16.431 | 15.172 | 3.065 | −0.2623 | HH | 23.280 | 20.139 | 3.355 | |
| **HISε** | | | | | Cδ1 | 20.825 | 17.132 | 2.281 | |
| Cγ | 18.921 | 15.161 | 2.698 | | Cε1 | 21.324 | 18.469 | 2.469 | |
| Nδ1 | 18.437 | 15.752 | 1.543 | | Cε2 | 22.086 | 17.879 | 4.656 | |
| Cε1 | 17.116 | 15.812 | 1.626 | | Cδ2 | 21.587 | 16.541 | 4.468 | |
| PT3 | 15.789 | 15.256 | 2.998 | 0.2729 | PT3 | 21.369 | 17.439 | 4.089 | 0.0148 |
| PT4 | 18.719 | 15.839 | 1.336 | −0.2729 | PT4 | 24.024 | 20.132 | 2.845 | 0.1523 |
| | | | | | PT5 | 21.930 | 20.166 | 4.149 | −0.1988 |
| | | | | | PT6 | 23.204 | 17.862 | 6.329 | 0.0153 |
| | | | | | PT7 | 21.719 | 15.562 | 6.013 | 0.0165 |

[a] Values of *q*1 and *q*2 depend on the AA type (Table SI2 in the Supporting Information).

as to generate unique superposition results, i.e., CG coordinates. For rigid side chains, such as His, Phe, and Trp, more than three atoms were also used to better fit the whole side chain plane. For Arg, more than three atoms were also used to generate, at once, all CGs, within the frame of the Amber99 FF. For Gln, points four and six of the Amber99-based CG representation were generated using the template formed by atoms Nε2, HNε21, and HNε22, while points three and five were determined using atoms Cγ, Cδ, and Nδ2. The Gromos43A1 CG model of the Gln side chain contained only three points. Points three and five were generated using the template formed by Nε2, HNε21, and HNε22, while the location of point four was based on atoms Cγ, Cδ, and Nδ2. Similarly, for Asn, points three and five for Amber99 (or points four and five for Gromos43A1) were generated using the template formed by atoms Nδ2, HNδ21,

***Table 5.*** Descriptions[a] of Protein Side Chain CG Models, as Defined in MARTINI, in Basdevant's model, and As Obtained from the Hierarchical Merging/Clustering of MEP Functions[b]

| | MARTINI[20,21] | Basdevant[6] | Amber99 | Gromos43A1 |
|---|---|---|---|---|
| Gly | – | – | – | – |
| small hydrophic residues | | | | |
| Ala | – | 1 | – | – |
| Ile | 1 apolar | 1 | – | – |
| Leu | 1 apolar | 1 | – | – |
| Pro | 1 apolar | 1 | – | – |
| Val | 1 apolar | 1 | – | – |
| large hydrophobic residue | | | | |
| Phe | 3 apolar | 2 | 4 $\|q\| < 0.08$ e$^-$ | 4 $\|q\| < 0.10$ e$^-$ |
| sulfur-containing residues | | | | |
| Cys | 1 apolar/polar | 1 | 2 | 2 |
| Met | 1 apolar/polar | 2 | 2 | – |
| polar amide-containing residues with H-bond property | | | | |
| Asn | 1 polar | 1 | 3 | 3 |
| Gln | 1 polar | 2 | 4 | 3 |
| small hydrophilic residues with OH group | | | | |
| Ser | 1 polar | 1 | 2 | 2 |
| Thr | 1 polar | 1 | 2 | 2 |
| ring-shape hydrophobic residues with H-bond property | | | | |
| His | 1 apolar, 2 polar | 2 | 3 | 2 |
| Trp | 3 apolar, 1 polar | 2 | 6 | 5 |
| Tyr | 2 apolar, 1 polar | 2 | 5 | 5 |
| charged residues | | | | |
| Arg | 1 apolar/polar, 1 charged | 2 | 4 | 1 |
| Asp | 1 charged | 1 | 2 | 2 |
| Glu | 1 charged | 2 | 2 | 2 |
| Lys | 1 apolar, 1 charged | 2 | 1 | 1 |

[a] Descriptions are in terms of number and property. [b] At $t = 1.25$ bohr$^2$ using Amber99 and $t = 1.3$ bohr$^2$ using Gromos43A1.

and HN$\delta$22, while point four for Amber99 (or point three for Gromos43A1) was obtained using C$\gamma$, O$\delta$1, and O$\varepsilon$1. In the case of Tyr, points three to five were located using the template formed by ring atoms C$\varsigma$, C$\varepsilon$1(or C$\varepsilon$2), and C$\delta$1(or C$\delta$2) in the opposite direction to the O−H bond, while points six and seven were generated using atoms C$\varsigma$, OH, and HH. A similar procedure is valid for the Gromos43A1-based models of Tyr. For the AA residues that are not reported in Tables 3 and 4, the CG coordinates were directly obtained from the side chain atom coordinates as specified in the tables reported in the Supporting Information.

Thus, from Tables 3 and 4, it is clear that CG points and charges can be directly obtained from a high-resolution structure/map of a protein, more precisely, from 3D atomic coordinates. For some AAs, like Asn, Cys, Gln, Ser, Thr, and Tyr, templates involve the knowledge of H atomic coordinates. Presently, these atoms were not defined in the PDB files but were added through the use of a software, such as PDB2PQR[40,41] and SwissPDBViewer.[43,44] In the case of lower crystallographic resolution maps, only a limited number of ED maxima could be located. In previous papers,[36,58] we showed how regular motifs of ED peaks still characterize AA backbone and side chains at resolution values close to 3 Å. A deeper study would be needed to relate the topology-based properties of these ED maxima, i.e., location, main ED curvatures, and local eigenvectors, to the positioning of the CG charges reported in Tables 3 and 4. Indeed, at the location of each ED maximum, a so-called Hessian matrix, built on the second derivatives of the density function vs the position, can be calculated. The diagonalization of such a matrix provides three eigenvalues, which physically define the main curvatures of the density function at the peak location and the three corresponding eigenvectors. Their orientation can help in locating CG charges.

**E. Application to Small Peptides.** Four small peptidic structures with electrostatic properties reported in the literature were selected. The first structure, a 12-residue $\beta$-hairpin HP7 was retrieved from the PDB[56,57] (PDB code 2EVQ) following the work of Basdevant et al.[6] The primary structure of that peptide is Lys−Thr−Trp−Asn−Pro−Ala−Thr−Gly−Lys−Trp−Thr−Glu, with a global net charge of +1. It is an interesting reference structure because a fragment-based description, as well as the corresponding point charges, were provided.[6] In that representation, each pseudoatom is defined as the geometric center of the heavy atoms of a protein fragment. Structure of two other peptides, i.e., the Tgn38 internalization peptide Dyqrln, with sequence Asp−Tyr−Gln−Arg−Leu−Asn, (PDB code 1BXX) and the C-terminal fragment of the chemotaxis receptor, with sequence Asn−Trp−Glu−Thr−Phe, (PDB code 1BC5) were studied following the work of Exner and Mezey.[61] Additionally, we selected the structure of a phospholipase inhibitor, with sequence Leu−Val−Phe−Phe−Ala, (PDB code 2RD4) involved in the A$\beta$7 structure studied by Pizzitutti et al.[7]

For each of those peptides, CG models were obtained by applying the automated procedure specified above. End charges were considered by including two additional charges, one on each of the terminal atoms N and OXT. By default, the corresponding charge values were set equal to ± 1. The

***Table 6.*** Electrostatic Properties of the Amber99-based CG Model of Small Peptides vs Their Corresponding All-Atom Version[a]

| | charge fitting | | | |
|---|---|---|---|---|
| | none[c] | $q_{\mathrm{end}}$ only[d] | all CG[e] | Basdevant[d] |
| 2EVQ | 197 atoms | | | |
| $q$ | 1.0 | | | |
| $\mu$ (all-atom)[b] | 4.78, −2.21, −66.43 | | | |
| no. of CGs | 51 | 51 | 51 | 28 |
| rmsdV | 5.98 | 3.63 | 1.54 | 5.27 |
| rmsd$\mu$ | 9.29 | 2.00 | 0.25 | 1.37 |
| $\mu$[b] | 1.50, 2.46, −73.76 | 3.54, −2.99, −67.81 | 4.67, −2.01, −66.33 | 4.89, −2.49, −65.10 |
| $q_{\mathrm{end}}$ | ±1.0000 | ±0.7660 | | |
| 1BXX | 110 atoms | | | |
| $q$ | 0.0 | | | |
| $\mu$ (all-atom)[b] | 11.34, −0.96, −15.73 | | | |
| no. of CGs | 32 | 32 | 32 | 15 |
| rmsdV | 6.03 | 3.52 | 1.90 | 7.09 |
| rmsd$\mu$ | 12.65 | 1.79 | 0.34 | 2.47 |
| $\mu$[b] | 21.06, 6.64, −12.96 | 12.31, −1.51, −14.33 | 11.62, −0.82, −15.60 | 13.55, −0.14, −14.99 |
| $q_{\mathrm{end}}$ | ±1.0000 | ±0.8532 | | |
| 1BC5 | 90 atoms | | | |
| $q$ | −1.0 | | | |
| $\mu$ (all-atom)[b] | −310.92, −287.53, 7.06 | | | |
| no. of CGs | 29 | 29 | 29 | 13 |
| rmsdV | 5.73 | 3.02 | 1.72 | 8.59 |
| rmsd$\mu$ | 9.02 | 1.49 | 0.09 | 4.82 |
| $\mu$[b] | −308.68, −296.26, 6.90 | −312.11, −286.69, 6.75 | −310.98, −287.55, 7.12 | −311.29, −282.99, 5.50 |
| $q_{\mathrm{end}}$ | ±1.0000 | ±0.8589 | | |
| 2RD4 | 88 atoms | | | |
| $q$ | 0.0 | | | |
| $\mu$ (all-atom)[b] | 35.12, 22.63, −44.04 | | | |
| no. of CGs | 20 | 20 | 20 | 12 |
| rmsdV | 4.40 | 3.25 | 1.62 | 8.74 |
| rmsd$\mu$ | 5.80 | 2.60 | 0.60 | 3.51 |
| $\mu$[b] | 39.38, 22.10, −47.94 | 35.71, 20.10, −43.84 | 34.66, 22.26, −44.14 | 34.98, 22.72, −40.53 |
| $q_{\mathrm{end}}$ | ±1.0000 | ±0.9162 | | |

[a] RmsdV and rmsd$\mu$ are given in kcal/mol and D, respectively. Electric charges are given in $e^-$. [b] $X$, $y$, and $z$ components of $\mu$. [c] No charge-fitting applied. [d] Charge-fitting applied to end charges $q_{\mathrm{end}}$ only. [e] Charge-fitting applied to all CG charges. [d] Charge-fitting applied to Basdevant's model.

quality of the Amber99-based CG model is evaluated vs the all-atom one in Table 6. It is achieved in terms of the rmsdV and rmsd$\mu$ deviation values. When no charge fitting is applied, models 1BC5 and 2RD4 approximate fairly well the all-atom electrostatic properties, while models 2EVQ and 1BXX are less well reproduced, especially at the level of the dipole moment values. For example, the sign of $\mu_y$ is inversed. While keeping all charges constant but the two end ones, we then applied a charge-fitting procedure that led to better models, with end charges $q_{\mathrm{end}}$ lower than unity. Models, and especially their dipole approximation, can, thus, be largely improved by fixing the end charges to absolute values lower than 1 $e^-$. However, we consider this improvement as the reflect only of altered end charges, i.e., modifications that depend on the particular protein structure under study. It is rather artificial to modify two AA models to approximate a global protein property, such as its dipole. The fitting of all CG charges obviously leads to largely better approximations with rmsdV and rmsd$\mu$ values ranging between 1.54−1.90 kcal/mol and 0.09−0.60 D, respectively. For comparison, we also fitted the charge values of the Basdevant's representation, with less efficiency, i.e., rmsdV = 5.27 to 8.74 kcal/mol and rmsd$\mu$ = 1.37 to 4.82 D. A similar discussion is valid for the Gromos43A1-based CG model (results are provided in the Supporting Information). However, let us mention that keeping $q_{\mathrm{end}}$ to unit values is observed to be a good choice for that particular set of charges. Indeed, for each of the four peptides studied, the fitting of the end charges led to absolute $q_{\mathrm{end}}$ values ranging between 1.02 and 1.06 $e^-$. That model, thus, appears to be characterized by more robust transferability properties than that of the Amber99-based one. It is assumed that Amber99-based CG models would provide better approximations of the all-atom representations if established at a less drastic smoothing degree. In conclusion, one observes, from the rmsdV and rmsd$\mu$ values, that the use of our model provides a good approximation of the all-atom MEPs, especially when end charges are fitted. A point charge model based on the com of the AA side chains and backbones is efficient too but requires a charge-fitting step that is not needed in our case. Additionally, the charge values that would be obtained using a full charge-fitting procedure are strongly dependent on the 3D conformation of the molecule. As illustrated in Figure 7 for Amber99, wherein CG-based MEP isocontours and projected values of the MEP onto the 0.0002 $e^-$/bohr$^3$ ED isosurface are compared vs the corresponding all-atom

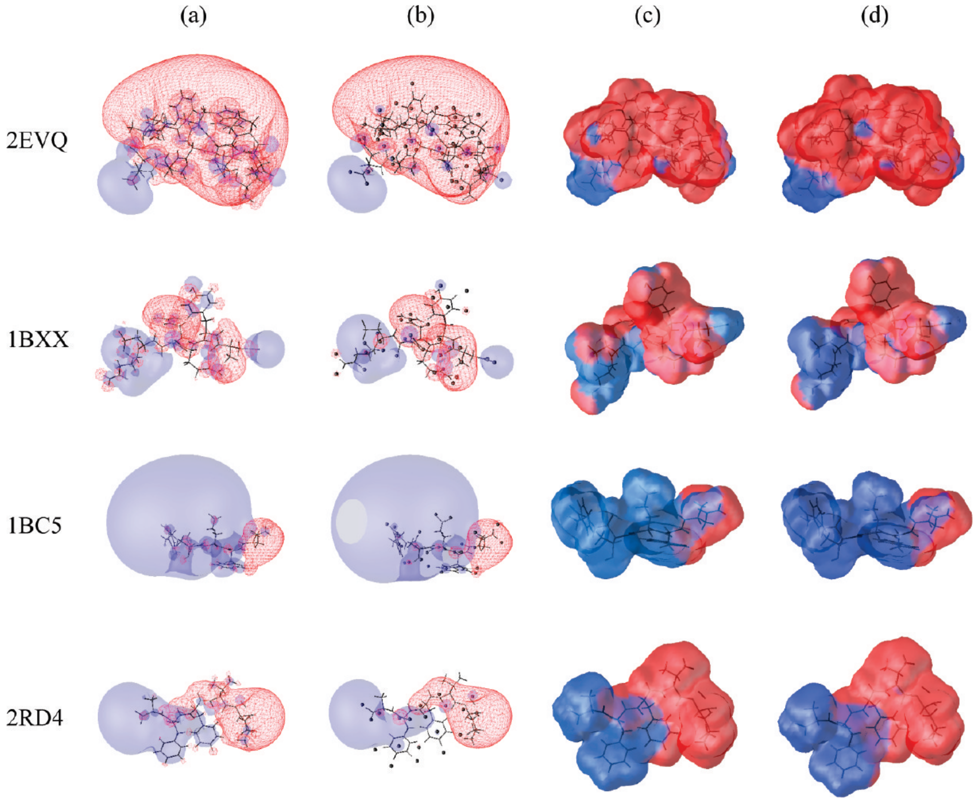


***Figure 7.*** Amber99 MEP isocontours (blue plain surface: −0.07, red mesh: 0.07 e$^-$/bohr) and MEP projected on the ED surface defined at 0.0002 e$^-$/bohr$^3$ (blue: negative, red: positive) of peptides 2EVQ, 1BXX, 1BC5, and 2RD4. (a) Unsmoothed all-atom MEP, (b) CG with fitted $q_{\text{end}}$ MEP with CGs (black spheres), (c) all-atom MEP on ED isocontour, and (d) CG with fitted $q_{\text{end}}$ MEP on ED isocontour.

properties, only local differences are clearly visible, especially at the close proximity of the molecular structure. There is a good correspondence between the CG and all-atom MEP 3D properties. A similar discussion is valid for the Gromos43A1 results provided in the Supporting Information.

**F. Application to the Potassium Ion Channel KcsA.** The protein structure selected to test our automated procedure was the KcsA potassium channel (Figure 8), a transmembrane protein structure that is commonly used to model biological ion channels[22,62−64] as well as to evaluate computational approaches in the study of protein electrostatics.[65−67] It is formed by four identical chains, each chain containing two α-helices connected by a loop located in the channel region (Figure 8). The channel consists of the so-called selectivity filter, that is about 18 Å long, pointing to the extracellular region, a larger cavity of about 10 Å and a 15 Å long narrow gating pore opened toward the intracellular region. The gating pore and the cavity are hydrophobic regions, while the selectivity filter, mainly formed by five residues Thr74−Thr75−Val76−Gly77−Tyr78, is covered by in-line carbonyl O atoms of the protein backbone, which build a structure that is similar to a water solvation shell around a $K^+$ ion.

In the present work, the 3D model of the protein was prepared according to the X-ray crystal structure of the KcsA $K^+$ channel (PDB access code 1BL8) by adding missing side chain atoms using the program SwissPDBViewer.[43,44] The design of the His residues into a Hisε configuration was achieved with the program VEGA ZZ.[68,69] The three $K^+$ ions, labeled K401, K402, and K403 (Figure 8), were not considered. Atom charges were assigned using PDB2PQR.[40,41] From an original structure of 5 888 atoms, the application of our automated procedure, completed by the addition of unit charges on the N and OXT atoms of the end residues of each of the four monomers, led to the generation of 1 284 and 1 204 CGs in the frameworks of the Amber99 and

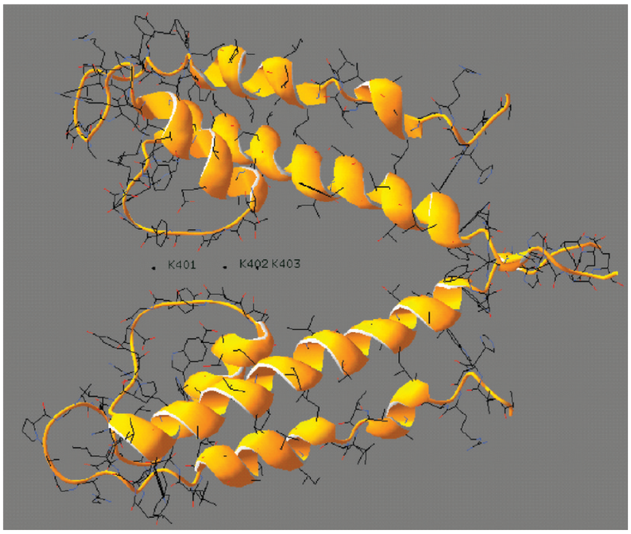


***Figure 8.*** 3D conformation and secondary structure of the potassium channel KcsA (PDB code 1BL8). Two monomers only, chains A and C, are displayed. Figure was generated using SwissPDBViewer.[43,44] Ions K401 and K403 are separated by a distance of 10.62 Å.

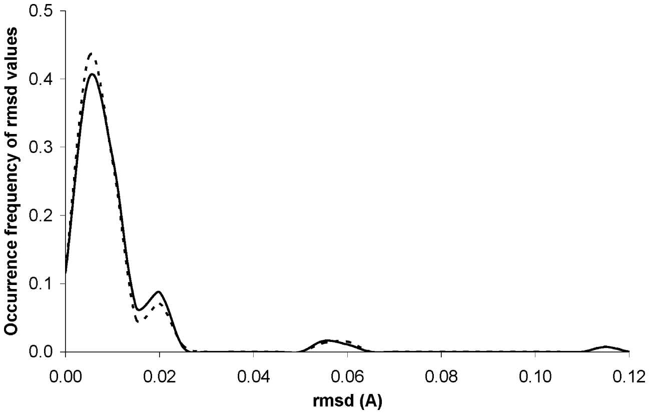


***Figure 9.*** Occurrence frequency of the rmsd values calculated between the atom positions of the AA template motif and the atom positions of the actual AA backbones or side chains, over all superimpositions achieved for the generation of the Amber99-based (plain line) and Gromos43A1-based (dashed line) CGs of protein structure KcsA.

Gromos43A1 FFs, respectively. The obtained reduction ratios, slightly larger than 4.5/1, is close to the 4/1 value reported by Bond and Sansom,[70,71] who studied the interaction of membrane proteins with lipid molecules through MD simulations. A visualization of the rmsd values obtained between the atoms of the AA templates and the corresponding atoms of the protein crystal structure, for each of the superimpositions achieved using QUATFIT[59,60] during the CG generation, is presented in Figure 9. The largest rmsd values, i.e., beyond 0.1 Å, correspond to a less efficient fit of the four end residues Gln119 required to design the Amber99-based CG model due to the terminal OXT atoms (Figure 10, left). The lowest rmsd values, around 0.01 Å, characterize the superimpositions of the backbone templates, while all larger rmsd values, from 0.02 to 0.06 Å, characterize the superimpositions of the side chain templates. Particularly, rmsd values around 0.05−0.06 Å originate from the superimpositions of the Tyr side chains. For example, Tyr82 of chain A that led to rmsd = 0.057 Å is illustrated in Figure 10 (right), where one can see that it nevertheless corresponds to a rather good superimposition of the three template atoms Cς, OH, and HH.

The resulting full KcsA CG models are characterized by dipole moments and total charges that are reported in Table 7, both for the Amber99 and Gromos43A1 FFs. In the case of the Amber99-based model, as the number of CGs is too large to allow any charge fitting procedure, we simply modified the end charge values $q_{\mathrm{end}}$ and observed that $q_{\mathrm{end}} = 0.5\ \mathrm{e}^-$ provided a model characterized by deviation values rmsdV and rmsd$\mu$ that are equal to 8.28 and 0.58 D, respectively. It is to be compared to the values of 7.38 kcal/mol and 81.60 D obtained when $q_{\mathrm{end}} = \pm 1\ \mathrm{e}^-$ is used (Table 7). The original MEP grid values were best approached when $q_{\mathrm{end}} = 0.8\ \mathrm{e}^-$ with a lower rmsdV = 6.13 kcal/mol but this, however, led to rmsd$\mu$ = 48.94 D, a value that is acceptable considering the magnitude of the dipole moment, i.e., 1411.36 D.

Visualizations of 3D MEP isocontours, generated from MEP maps built with a grid step of 0.5 Å (Figure 11), do not permit to clearly differentiate the MEPs calculated using the original sets of charges (Figure 11, left) from those calculated using the CG models (Figure 11, right). Finer and more quantitative comparisons were, thus, achieved. MEP profiles were calculated using the original atom charges along the channel axis, defined by the Cartesian coordinates of ions K401 and K403 (Figure 12). As illustrated, the channel axis region of the selective filter region is characterized by two MEP minima, followed by a large energy barrier which covers the hydrophobic cavity and narrow pore regions. The calculation of the corresponding MEP profiles using the Amber99- and Gromos43A1-based CG models generate similar behaviors, very close to their all-atom version. In that sense, the models presented in this paper led to better approximations than those obtained in a previous approach,[37] wherein AA CG models were generated using pentadecapeptide structures rather than isolated structures. It seems that decoupling backbone and side chain contributions in the elaboration of a CG model is interesting for reproducing all-atom electrostatic properties.

## IV. Conclusions and Perspectives

In this work, we applied a hierarchical merging/clustering algorithm to molecular scalar fields, like molecular electrostatic potential (MEP) functions, to generate coarse point charge representations of proteins. Through the use of such a procedure, the reduction of a molecular structure representation, particularly a protein structure, was achieved by

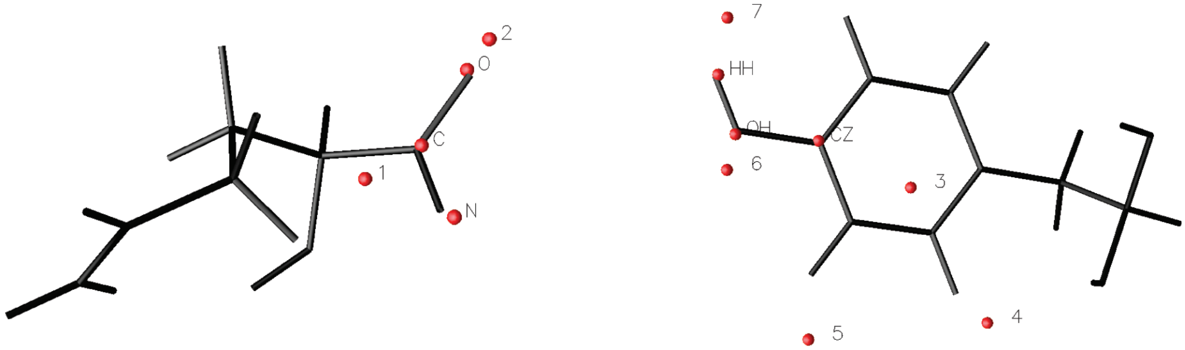


***Figure 10.*** (Left) Amber99-based template motif (red spheres) of the Gln backbone as superimposed on Gln119 of chain A in protein KcsA. The three atoms C, O, and N are used to generate the transformation matrix that is further applied to CGs numbered 1 and 2. (Right) Amber99-based template motif of the Tyr side chain as superimposed on Tyr82 of chain A in protein KcsA. The three atoms Cς, OH, and HH are used to generate the transformation matrix that is further applied to CG points numbered 3 to 7.

***Table 7.*** Electrostatic Properties of the Amber99- and Gromos43A1-Based CG Models of Structure KcsA vs Their Corresponding All-Atom Version[a]

| | | Amber99 | Gromos43A1 |
|---|---|---|---|
| total charge | | 4.0 | 4.0 |
| $\mu$ (D) | | 1411.36 | 1402.76 |
| $\mu$ all-atom (D)[b] | | 1303.10, 511.49, 179.54 | 1295.8, 511.82, 163.31 |
| no. of CG points | | 1 284 | 1 204 |
| reduction factor | | 4.6/1 | 2.1/1 vs charged atoms |
| $q_{\text{end}} = \pm$ 1.0 | $\mu$ CG (D)[b] | 1273.17, 512.15, 103.63 | 1293.14, 512.66, 156.49 |
| | rmsdV (kcal/mol) | 7.38 | 2.78 |
| | rmsd$\mu$ (D) | 81.60 | 7.37 |

[a] RmsdV and rmsd$\mu$ are given in kcal/mol and D, respectively. Electric charges are given in $e^-$. [b] *X*, *y*, and *z* components of $\mu$.

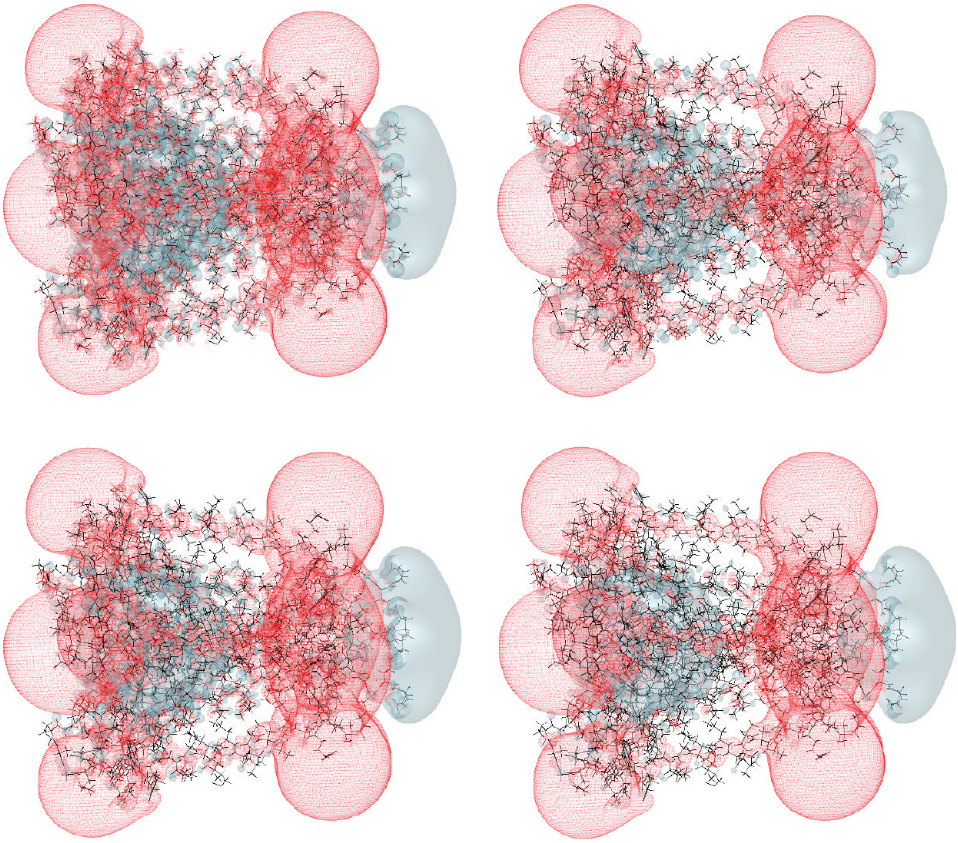


***Figure 11.*** MEP isocontours (blue plain surface: −0.1, red mesh: 0.1 $e^-$/bohr) of (top left) unsmoothed all-atom Amber99, (top right) Amber-based CG with $q_{\text{end}} = \pm$ 0.5 $e^-$, (bottom left) unsmoothed united-atom Gromos43A1, and (bottom right) Gromos43A1-based CG with $q_{\text{end}} = \pm$ 1.0 $e^-$, superimposed on the 3D structure of protein KcsA (sticks).

following the trajectories of its constituting atoms in its progressively smoothed three-dimensional (3D) molecular field. A protein structure can, thus, be described by a limited set of points, which correspond to the local extrema (peaks and pits) of the considered 3D MEP field. The aim of such calculations further consisted in the evaluation of electrostatic properties, such as point charges and dipole moments, of a protein using coarse-grain (CG) descriptions.

The present work especially focused on the use of the sets of charges of the all-atom Amber99 and the united-atom Gromos43A1 force fields (FF) but is readily applicable to other charge sets that are available in the literature. Reduced descriptions were obtained for each of the 20 natural amino acid (AA) residues with the following specific protonation states: Arg(+1), Lys(+1), Asp(−1), and Glu(−1). Each of the 20 AAs was modeled through various rotamers (except for Ala, Asp, Gly, and Pro). The first stage was to apply our merging/clustering algorithm to determine the CG locations of the AA backbone and side chain, separately. In a second stage, charges were assigned to these AA CG representations through a charge-fitting algorithm and were further tabulated as reference values to be used for CG modeling of protein structures. MEP-based CG descriptions were shown to be sensitive to the molecular conformation. Additional studies, achieved at various levels of smoothing, showed that the optimal value of *t* is only slightly dependent on the selected FF charges. It is equal to 1.25 and 1.3 bohr$^2$ for Amber99 and Gromos43A1, respectively.

An automated procedure was implemented and tested on four small peptides (PDB access codes 2EVQ, 1BXX, 1BC5, and 2RD4) and on a larger system KcsA, a tetrameric potassium ion channel made of four 97-residue long monomers (PDB access code 1BL8). The generation of the CG representation of each residue was achieved through a

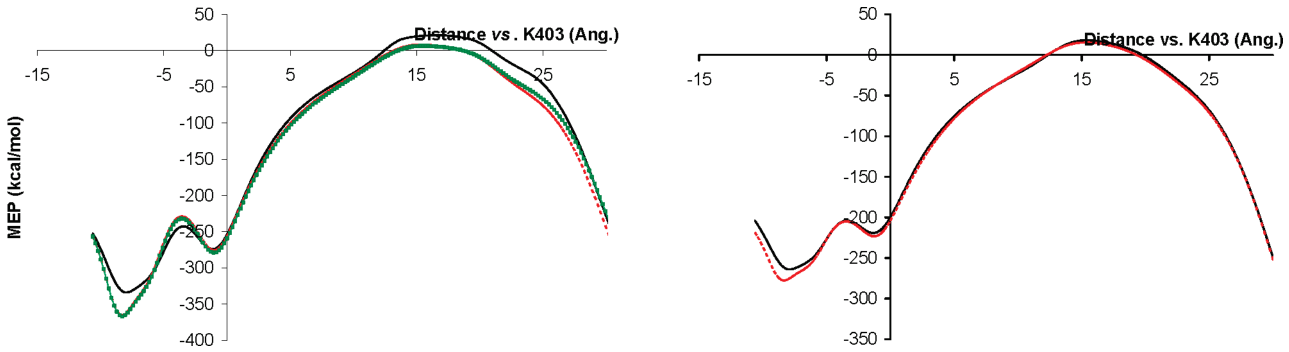


***Figure 12.*** MEP profiles along the central axis of the KcsA potassium channel calculated using (left) the all-atom Amber99 set of charges (black plain line), the Amber99-based CG model with $q_{\mathrm{end}} = \pm 1.0$ e$^-$ (red dashed lines), the Amber99-based CG model with $q_{\mathrm{end}} = 0.5$ e$^-$ (green squares), and (right) the united-atom Gromos43A1 set of charges (black plain line), the Gromos43A1-based CG model with $q_{\mathrm{end}} = \pm 1.0$ e$^-$ (red dashed lines).

superimposition algorithm of CG template motifs on the 3D PDB structure. The study of the four peptides revealed that end charges should be lower than unity for the Amber99-based CG models and equal to unity for the Gromos43A1-based CG models. For Amber99, the variability in the end charge values is assumed to be the reflection of moderate transferability properties. For the larger system KcsA, the CG descriptions, consisting of 1 284 and 1 204 CG points and their tabulated charges, in the frameworks of Amber99 and Gromos43A1, respectively, allowed to well reproduce the trends observed in the unsmoothed all-atom MEP functions.

Our calculations suggest that decoupling backbone and side chain contributions in the elaboration of an AA CG model is interesting for reproducing all-atom electrostatic properties and that the location of CG steric centers, like those defined by ED peaks or by centers-of-mass of specific groups of atoms, differ from the location of CG electrostatic centers. This might be a point to consider in the further development of a CG FF.

During the elaboration of the MEP-based CG models, two points were considered to be important to favor transferability. First, AAs were studied in the isolated state to neglect the protein backbone conformation, and second, for each AA, CG charges were obtained by considering various side chain conformations. Though probably not sufficient to definitely demonstrate transferability robustness of our models, the results are encouraging, and they open an interesting extension to the present work, for example, in the comparison of MEP calculated using the Poisson−Boltzmann formalism.[3] One can also imagine two more direct ways to test transferability of Coulomb potentials built from MEP-based CG models. The first one could consist in applying our procedure to a larger set of protein structures. The other would ask for a detailed comparison between MEP profiles calculated at the all-atom and CG levels, and this, for all possible AA-AA pairs. Finally, one could also elaborate for each AA type, CG representations and/or charges that depend on the rotamer class.

To directly link MEP and experimental ED distribution functions, one could use databases of transferable multipolar ED parameters for evaluating atom charges, as presented by Zarichta et al.[72] and preliminarily applied to the human aldose reductase system,[36] and then calculate MEP functions.

**Acknowledgment.** The authors thank the referees for extremely useful comments that led to an improvement in the CG models obtained for the AAs as well as Prof. E. Clementi and Prof. Cl. Lecomte and the members of his research group for very fruitful discussions. The FNRS-FRFC, the "Loterie Nationale" (convention no. 2.4578.02) and the Facultés Universitaires Notre-Dame de la Paix (FUNDP) are gratefully acknowledged for the use of the Interuniversity Scientific Computing Facility (ISCF) Center.

**Supporting Information Available:** Tables SI1 and SI2 contain a full description of the AA CG models in terms of charges and their rmsdV and rmsd$\mu$ values for Amber99 and Gromos43A1, respectively. Table SI6 for Gromos43A1 is equivalent to Table 6 for Amber99. Figure SI7 for Gromos43A1 is equivalent to Figure 7 for Amber99. 3D structure and charges of the all-atom and coarse point charge models generated for the four peptides (PDB codes 2EVQ, 1BXX, 1BC5, and 2RD4) and structure KcsA (PDB code 1BL8) can be downloaded from http://perso.fundp.ac.be/~lleherte/JCTC_SI/. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) *Coarse-Graining of Condensed Phase and Biomolecular Systems*; Voth, G. A., Ed.; CRC Press: Boca Raton, FL, USA, 2009.

(2) Dong, F.; Olsen, B.; Baker, N. A. Computational Methods for Biomolecular Electrostatics. *Methods Cell Biol.* **2008**, *84*, 843–870.

(3) Schutz, Cl. N.; Warshel, A. What Are the Dielectric "Constants" of Proteins and How To Validate Electrostatic Models? *Proteins* **2001**, *44*, 400–417.

(4) Skepö, M.; Linse, P.; Arnebrant, T. Coarse-Grained Modeling of Proline Rich Protein 1 (PRP-1) in Bulk Solution and Adsorbed to a Negatively Charged Surface. *J. Phys. Chem. B* **2006**, *110*, 12141–12148.

(5) Curcó, D.; Nussinov, R.; Alemán, C. Coarse-Grained Representation of $\beta$-Helical Protein Building Blocks. *J. Phys. Chem. B* **2007**, *111*, 10538–10549.

(6) Basdevant, N.; Borgis, D.; Ha-Duong, T. A. Coarse-Grained Protein-Protein Potential Derived from an All-Atom Force Field. *J. Phys. Chem. B* **2007**, *111*, 9390–9399.

(7) Pizzitutti, F.; Marchi, M.; Borgis, D. Coarse-Graining the Accessible Surface and the Electrostatics of Proteins for Protein-Protein Interactions. *J. Chem. Theory Comput.* **2007**, *3*, 1867–1876.

(8) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules. *Biophys. J.* **2008**, *95*, 5073–5083.

(9) Bereau, T.; Deserno, M. Generic Coarse-Grained Model for Protein Folding and Aggregation. *J. Chem. Phys.* **2009**, *130*, 235106/1–235106/15.

(10) Paramonov, L.; Yaliraki, S. N. The Directional Contact Distance of Two Ellipsoids: Coarse-Grained Potentials for Anisotropic Interactions. *J. Chem. Phys.* **2005**, *123*, 194111/1–194111/11.

(11) Izvekov, S.; Voth, G. A. A Multiscale Coarse-Graining Method for Biomolecular Systems. *J. Phys. Chem. B* **2005**, *109*, 2469–2473.

(12) Liu, P.; Izvekov, S.; Voth, G. A. Multiscale Coarse-Graining of Monosaccharides. *J. Phys. Chem. B* **2007**, *111*, 11566–11575.

(13) Carbone, P.; Varzaneh, H. A. K.; Chen, X. Y.; Müller-Plathe, F. Transferability of Coarse-Grained Force Fields: The Polymer Case. *J. Chem. Phys.* **2008**, *128*, 064904/1–064904/11.

(14) Kondrashov, D. A.; Cui, Q.; Phillips, G. N., Jr. Optimization and Evaluation of a Coarse-Grained Model of Protein Motion Using X-Ray Crystal Data. *Biophys. J.* **2006**, *91*, 2760–2767.

(15) Fukunaga, H.; Aoyagi, T.; Takimoto, J.-I.; Doi, M. Derivation of Coarse-Grained Potential for Polyethylene. *Comput. Phys. Commun.* **2001**, *142*, 224–226.

(16) Lyman, E.; Pfaendtner, J.; Voth, G. A. Systematic Multiscale Parametrization of Heterogeneous Elastic Network Models of Proteins. *Biophys. J.* **2008**, *95*, 4183–4192.

(17) Lyubartsev, A. P.; Laaksonen, A. Calculation of Effective Interaction Potentials from Radial Distribution Functions: A Reverse Monte Carlo Approach. *Phys. Rev. E* **1995**, *52*, 3730–3737.

(18) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. The Multiscale Coarse-Graining Method. I. A Rigorous Bridge between Atomistic and Coarse-Grained Models. *J. Chem. Phys.* **2008**, *128*, 244114/1–244114/11.

(19) Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. The Multiscale Coarse-Graining Method. II. Numerical Implementation for Coarse-Grained Molecular Models. *J. Chem. Phys.* **2008**, *128*, 244115/1–244115/20.

(20) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI Forcefield: Coarse-Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

(21) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI Coarse-Grained Forcefield: Extension to Proteins. *J. Chem. Theory Comput* **2008**, *4*, 819–834.

(22) Treptow, W.; Marrink, S.-J.; Tarek, M. Gating Motions in Voltage-Gated Potassium Channels Revealed by Coarse-Grained Molecular Dynamics Simulations. *J. Phys. Chem. B* **2008**, *112*, 3277–3282.

(23) Liwo, A.; Czaplewski, C.; Oldziej, S.; Rojas, A. V.; Kazmierkiewicz, R.; Makowski, M.; Murarka, R. K.; Sheraga, H. A. Simulation of Protein Structure & Dynamics with the Coarse-Grained UNRES Force Field. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*; Voth, G. A., Ed.; CRC Press: Boca Raton, FL, 2009; Chapter 8, pp 107−122.

(24) Fujitsuka, Y.; Takada, S.; Luthey-Schulten, Z. A.; Wolynes, P. G. Optimizing Physical Energy Functions for Protein Folding. *Proteins* **2004**, *54*, 88–103.

(25) Hori, N.; Chikenji, G.; Berry, R. S.; Takada, S. Folding Energy Landscape and Network Dynamics of Small Globular Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 73–78.

(26) Gabdoulline, R. R.; Wade, R. C. Effective Charges for Macromolecules in Solvent. *J. Phys. Chem.* **1996**, *100*, 3868–3878.

(27) Berardi, R.; Muccioli, L.; Orlandi, S.; Ricci, M. Zannoni, Cl. Mimicking Electrostatic Interactions with a Set of Effective Charges: A Genetic Algorithm. *Chem. Phys. Lett.* **2004**, *389*, 373–378.

(28) Golubkov, P. A.; Ren, P. -Y. Generalized Coarse-Grained Model Based on Point Multipole and Gay-Berne Potentials. *J. Chem. Phys.* **2006**, *125*, 064103/1–064103/11.

(29) Cascella, M.; Neri, M. A.; Carloni, P.; Dal Peraro, M. Topologically Based Multipolar Reconstruction of Electrostatic Interactions in Multiscale Simulations of Proteins. *J. Chem. Theory Comput.* **2008**, *4*, 1378–1385.

(30) Yang, L.-W.; Chng, Ch.-P. Coarse-Grained Models Reveal Functional Dynamics - I. Elastic Network Models - Theories, Comparisons and Perspectives. *Bioinf. Biol. Insights* **2008**, *2*, 25–45.

(31) Chng, Ch.-P.; Yang, L.-W. Coarse-Grained Models Reveal Functional Dynamics - II. Molecular Dynamics Simulation at the Coarse-Grained Level - Theories and Biological Applications. *Bioinform. Biol. Insights* **2008**, *2*, 171–185.

(32) Clementi, C. Coarse-Grained Models of Protein Folding: Toy Models or Predictive Tools? *Curr. Opin. Struct. Biol.* **2008**, *18*, 10–15.

(33) Sherwood, P.; Brooks, B. R.; Sansom, M. S. P. Multiscale Methods for Macromolecular Simulations. *Curr. Opin. Struct. Biol.* **2008**, *18*, 630–640.

(34) Eyal, E.; Bahar, I. Toward a Molecular Understanding of the Anisotropic Response of Proteins to External Forces: Insights from Elastic Network Models. *Biophys. J.* **2008**, *94*, 3424–3435.

(35) Zhou, L.; Siegelbaum, S. A. Effects of Surface Water on Protein Dynamics Studied by a Novel Coarse-Grained Normal Mode Approach. *Biophys. J.* **2008**, *94*, 3461–3474.

(36) Leherte, L.; Guillot, B.; Vercauteren, D. P.; Pichon-Pesme, V.; Jelsch, Ch.; Lagoutte, A.; Lecomte, C. Topological Analysis of Proteins as Derived from Medium and High-Resolution Electron Density: Applications to Electrostatic Properties. In *The Quantum Theory of Atoms in Molecules - From Solid State to DNA and Drug Design*; Matta, C. F., Boyd, R. J., Eds.; Wiley-VCH: Weinheim, Germany, 2007; Chapter 11, pp 285−316.

(37) Leherte, L.; Vercauteren, D. P. Determination of Protein Coarse-Grain Charges from Smoothed Electron Density Distribution Functions and Molecular Electrostatic Potentials.

In *Handbook of Computational Chemistry Research*; Collett, C. T.; Robson, C. D., Eds.; NovaScience Publishers, in press.

(38) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(39) 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.

(40) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: An Automated Pipeline for the Setup of Poisson-Boltzmann Electrostatics Calculations. *Nucleic Acids Res.* **2004**, *32*, W665–W667.

(41) An Automated Pipeline for the Setup, Execution, and Analysis of Poisson-Boltzmann Electrostatics Calculations. *PDB2PQR*; SourceForge: Mountain View, CA, 2007; http://pdb2pqr.sourceforge.net/. Accessed August 31, 2009.

(42) Scott, W. R. P.; Hunenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Kruger, P.; van Gunsteren, W. F. The GROMOS Biomolecular Simulation Program Package. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.

(43) Guex, N.; Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer. An Environment for Comparative Protein Modeling. *Electrophoresis* **1997**, *18*, 2714–2723.

(44) Guex, N.; Diemand, A.; Peitsch, M. C.; Schwede, T. *Swiss-PdbViewer DeepView*, Version 4.0, 2008; http://spdbv.Vital-it.ch/. Accessed August 26, 2009.

(45) *Open Source Software Project*; Based on IBM's Visualization Data Explorer; IBM: Armonk, New York, 2007; http://www.opendx.org/. Accessed August 26, 2009.

(46) Leung, Y.; Zhang, J.-S.; Xu, Z.-B. Clustering by Scale-Space Filtering. *IEEE T. Pattern Anal* **2000**, *22*, 1396–1410.

(47) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*; Abramowitz, M., Stegun, I. A., Eds.; Dover Publications: New York, 1970.

(48) Hart, R. K.; Pappu, R. V.; Ponder, J. W. Exploring the Similarities between Potential Smoothing and Simulated Annealing. *J. Comput. Chem.* **2000**, *21*, 531–552.

(49) Borodin, O.; Smith, G. D. *Force Field Fitting Toolkit*; University of Utah: Salt Lake City, UT; http://www.eng.utah.edu/~gdsmith/fff.html. Accessed August 26, 2009.

(50) Singh, U. C.; Kollman, P. A. An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.* **1984**, *5*, 129–145.

(51) Eisenmenger, F.; Hansmann, U. H. E.; Hayryan, S.; Hu, C.-K. An Enhanced Version of SMMP-Open-Source Software Package for Simulation of Proteins. *Comput. Phys. Commun.* **2006**, *174*, 422–429.

(52) *Simple Molecular Mechanics for Proteins*; Michigan Technological University: Houghton, MI; http://www.smmp05.net/. Accessed August 26, 2009.

(53) Nemethy, G.; Gibson, K. D.; Palmer, K. A.; Yoon, C. N.; Paterlini, G.; Zagari, A.; Rumsey, S.; Scheraga, H. A. Energy Parameters in Polypeptides. 10. Improved Geometrical Parameters and Nonbonded Interactions for Use in the ECEPP/3 Algorithm, with Application to Proline-Containing Peptides. *J. Phys. Chem.* **1992**, *96*, 6472–6484.

(54) Simms, A. M.; Toofanny, R. D.; Kehl, C.; Benson, N. C.; Daggett, V. Dynameomics: Design of a Computational Lab Workflow and Scientific Data Repository for Protein Simulations. *Prot. Eng. Des. Sel.* **2008**, *21*, 369–377.

(55) *DYNAMEOMICS*; University of Washington: Seattle, WA, 2007; http://www.dynameomics.org/. Accessed August 26, 2009.

(56) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(57) RCSB PDB Protein Data Bank; Rutgers, the State University of New Jersey: Piscataway, NJ and San Diego Supercomputer Center (SDSC) and Skaggs School of Pharmacy and Pharmaceutical Sciences: La Jolla, CA, 2009; http://www.rcsb.org/pdb. Accessed August 26, 2009.

(58) Leherte, L. Hierarchical Analysis of Promolecular Full Electron Density Distributions: Description of Protein Structure Fragments. *Acta Crystallogr, Sect. D: Biol. Crystallogr.* **2004**, *60*, 1254–1265.

(59) Heisterberg, D. J. Technical report. *Translation from FORTRAN to C and Input/Output*; Labanowski, J., Ed.; Ohio Supercomputer Center: Columbus, OH, 1990.

(60) *CCL quaternion-mol-fit*; Computational Chemistry List, Ltd: Columbus, OH; http://www.ccl.net/cca/software/SOURCES/C/quaternion-mol-fit/. Accessed August 26, 2009.

(61) Exner, Th. E.; Mezey, P. G. Ab Initio-Quality Electrostatic Potentials for Proteins: An Application of the ADMA Approach. *J. Phys. Chem. A* **2002**, *106*, 11791–11800.

(62) Boiteux, C.; Kraszewski, S.; Ramseyer, Ch.; Girardet, Cl. Ion Conductance vs. Pore Gating and Selectivity in KcsA Channel: Modeling Achievements and Perspectives. *J. Mol. Model.* **2007**, *13*, 699–713.

(63) Noskov, S. Y.; Roux, B. Importance of Hydration and Dynamics on the Selectivity of the KcsA and NaK Channels. *J. Gen. Physiol.* **2007**, *129*, 135–143.

(64) Khavrutskii, I. V.; Fajer, M.; McCammon, J. A. Intrinsic Free Energy of the Conformational Transition of the KcsA Signature Peptide from Conducting to Nonconducting State. *J. Chem. Theory Comput.* **2008**, *4*, 1541–1554.

(65) Gascon, J. A.; Leung, S. S. F.; Batista, E. R.; Batista, V. S. A Self-Consistent Space-Domain Decomposition Method for QM/MM Computations of Protein Electrostatic Potentials. *J. Chem. Theory Comput.* **2006**, *2*, 175–186.

(66) Warshel, A.; Kato, M.; Pisliakov, A. V. Polarizable Force Fields: History, Test Cases, and Prospects. *J. Chem. Theory Comput.* **2007**, *3*, 2034–2045.

(67) Piccinini, E.; Ceccarelli, M.; Affinito, F.; Brunetti, R.; Jacoboni, C. Biased Molecular Simulations for Free-Energy Mapping: A Comparison on the KcsA Channel as a Test Case. *J. Chem. Theory Comput.* **2008**, *4*, 173–183.

(68) Pedretti, A.; Villa, L.; Vistoli, G. VEGA - An Open Platform To Develop Chemo-Bio-Informatics Applications using Plug-In Architecture and Script Programming. *J. Computer. Aided Mol. Des.* **2004**, *18*, 167–173.

(69) VEGA ZZ, 2009. Drug Design Laboratory-University of Milan. http://www.ddl.unimi.it/ Accessed August 26, 2009.

(70) Bond, P. J.; Sansom, M. S. P. Insertion and Assembly of Membrane Proteins via Simulation. *J. Am. Chem. Soc.* **2006**, *128*, 2697–2704.

(71) Bond, P. J.; Holyoake, J.; Ivetac, A.; Khalid, S.; Sansom, M. S. P. Coarse-Grained Molecular Dynamics Simulations of Membrane Proteins and Peptides. *J. Struct. Biol.* **2007**, *157*, 593–605.

(72) Zarychta, B.; Pichon-Pesme, V.; Guillot, B.; Lecomte, C.; Jelsch, C. On the Application of an Experimental Multipolar Pseudo-Atom Library for Accurate Refinement of Small-Molecule and Protein Crystal Structures. *Acta Crystallogr. A* **2007**, *63*, 108–125.

CT900193M

JCTC Journal of Chemical Theory and Computation

# Molecular Dynamics Simulations of *BcZBP*, A Deacetylase from *Bacillus cereus*: Active Site Loops Determine Substrate Accessibility and Specificity

Vasiliki E. Fadouloglou,[†,‡,⊥] Athanassios Stavrakoudis,[§] Vassilis Bouriotis,[‡,||] Michael Kokkinidis,[‡,||] and Nicholas M. Glykos*[,†]

*Department of Molecular Biology and Genetics, Democritus University of Thrace, University campus, GR-68100, Alexandroupolis, Greece, Department of Biology, University of Crete, PO Box 2208, GR-71409, Heraklion, Crete, Greece, Department of Economics, University of Ioannina, GR-45110, Ioannina, Greece, and Institute of Molecular Biology and Biotechnology (IMBB), PO Box 1527, GR-71110, Heraklion, Crete, Greece*

Received May 11, 2009

**Abstract:** *BcZBP* is an LmbE-like, homohexameric, zinc-dependent deacetylase from the opportunistic pathogen *Bacillus cereus* with three, thus far uncharacterized, homologues in *B. anthracis*. Although its specific substrate is still unknown, the enzyme has been shown to preferentially deacetylate *N*-acetylglucosamine and diacetylchitobiose via an active site based on a zinc-binding motif of the type HXDDX$_n$H. In the crystal structure, the active site is located at a deep and partially blocked cleft formed at the interface between monomers related by the molecular 3-fold axis, although the major, in structural terms, building block of the enzyme is not the trimer, but the intertwined dimer. Here, we report results from a 50 ns molecular dynamics simulation of *BcZBP* in explicit solvent with full electrostatics and show that (i) the view of the intertwined dimer as the major structural and functional building block of this class of hexameric enzymes is possibly an oversimplification of the rather complex dynamics observed in the simulation, (ii) the most mobile (with respect to their atomic fluctuations) parts of the structure coincide with three surface loops surrounding the active site, and (iii) these mobile loops define the active site's accessibility, and may be implicated in the determination of the enzyme's specificity.

## 1. Introduction

*Bacillus anthracis* has recently attracted significant interest, mainly because of its putative usage as a biological weapon.[1] Part of this interest was subsequently transferred to more benign, but still closely related to *B. anthracis*, species like *B. cereus*, an opportunistic bacterium causing food poisoning.[2] In a drive to characterize, both functionally and structurally, deacetylases that are shared between these two organisms (and which may be implicated in metabolic pathways of biotechnological and pharmaceutical interest), we have recently reported the characterization, purification, crystallization and crystal structure determination of *BcZBP*, a homohexameric, LmbE-like, zinc-dependent deacetylase from *B. cereus*.[3−5] *BcZBP* is the product of the *bc1534* gene, with three, thus far uncharacterized, homologues in *B. anthracis* sharing sequence identities (at the protein level) of 96%, 28%, and 24%, respectively. Functional studies[5]

* To whom correspondence should be addressed. Tel. +302551030620. Fax. +302551030620, glykos@mbg.duth.gr. URL: http://www.mbg.duth.gr/~glykos/.

[†] Democritus University of Thrace.

[‡] University of Crete.

[§] University of Ioannina.

[||] Institute of Molecular Biology and Biotechnology (IMBB).

[⊥] Present address: Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, CB2 1GA, Cambridge, England (UK).

10.1021/ct9002338 CCC: $40.75
Published on Web 11/04/2009

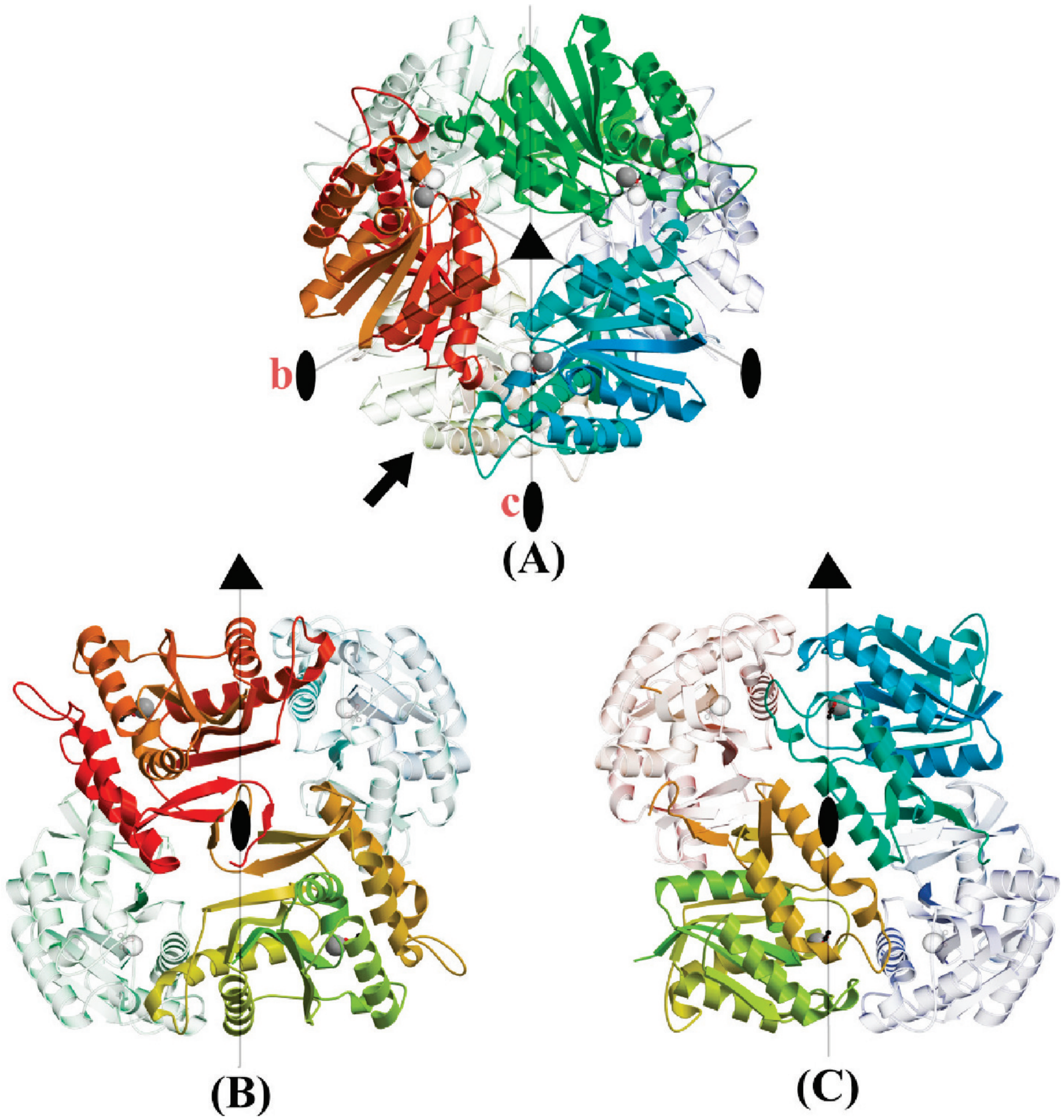


***Figure 1.*** Crystal structure of *BcZBP*. Panel (A) is a schematic diagram of the hexamer viewed down the molecular 3-fold axis (which coincides with a crystallographic 3-fold of the *R*32 space group). The 3-fold passes through the geometric center of the molecule, is perpendicular to the plane of the paper, and its position is marked with a filled triangle. The 2-fold axes of the hexamer are on the plane of the paper, intersect the 3-fold at the molecular center, and their position is indicated by thin lines marked with filled ovals. The trimer that is farthest from the viewer (and below the plane of the paper) is drawn using transparency to reduce clutter. The arrow points to the monomer (of the lower trimer) that is colored green-orange in panels B and C. The position of the active site in each monomer is marked by the presence of a space-filling representation of the zinc atom. Panels B and C are views of the hexamer along the molecular 2-fold axes which are marked as b and c in panel A. In both cases, the 2-folds are perpendicular to the plane of the paper, pass through the molecular center and their position is marked with a filled oval. The 3-fold axis is on the plane of the paper and is marked with a line ending in a filled triangle. The trimer that was drawn transparent in panel A is now located toward the lower end of the page. To reduce clutter, only four monomers are shown in panels B and C with two of them drawn semitransparent. The dimers shown in panels B and C are referred to in the text as type I and type II dimers, respectively.

showed that *BcZBP* preferentially deacetylates *N*-acetylglucosamine and diacetylchitobiose, although its specific substrate remains unknown. The overall structure of the enzyme is shown in Figure 1A. It is a symmetric 32 ($D_3$ in the Schoenflies notation) homohexamer with a molecular mass of 163 kDa and 234 residues in each monomer (we will hereafter refer to these six chains using the letters A−F). Their arrangement in the hexamer is such that chains A,C,E form the first trimer and chains B,D,F the second. In the crystal, the molecular 3-fold coincides with a crystallographic 3-fold axis of the *R*32 space group, leaving the equivalent of two monomers (chains A and B, or equivalently C−D or E−F) per crystallographic asymmetric unit. Each monomer folds as a single α/*β* domain in the form of a four-layered α/α/*β*/α sandwich (most easily seen in the lower monomer of Figure 1B). The two trimers (of the hexamer) associate strongly, mainly via contacts located around the 2-fold axes. The first of these contacts is shown in Figure 1B and involves an exchange of two short $\beta$-strands between the monomers A−B (and equivalently, C−D and E−F). The second contact (between monomers B−C, D−E, and A−F) is shown in Figure 1C and corresponds to a typical ridges-into-grooves $\alpha$-helical association with an interhelix angle of approximately 25° and a mean helix−helix distance of 8 Å. We will hereafter refer to these two types of dimers as "type I" (shown in Figure 1B) and "type II" (shown in Figure 1C) dimers, respectively.

The close, strand-exchange-based association seen in Figure 1B led to the conclusion[5] that the major structural building block of *BcZBP* is the type I dimer, and not, for example, the trimer or the type II dimer shown in Figure 1C. The view of *BcZBP* as a trimer of type I dimers was further reinforced by the relatively loose association of the monomers in the trimers as can be inferred from Figure 1A. Still, this view of *BcZBP* as a trimer of dimers could not be easily reconciled with two pronounced features of the crystal structure: The first feature was a systematic difference between the mean atomic temperature factors of the two trimers, with one trimer having significantly higher thermal parameters than the other. This observation was not consistent with the view of the type I dimer as the major structural and functional building block, mainly because such a dimer comprises monomers belonging to different trimers (it should be noted, however, that the presence of crystallographic symmetry relating the trimers' monomers, forces any deviations in the overall atomic temperature factors to be at the trimer's level). The second feature was that the accessibility to the active site appears to be mainly determined from loops originating from neighboring monomers not involved in the formation of the type I dimer. Furthermore, the crystal structure's active sites were partially blocked from three surface loops of neighboring (in the trimer) monomers, making it difficult to imagine the type I dimer as the enzyme's functional unit.

Here we present results from a 50 ns molecular dynamics simulation on the *BcZBP* hexamer in explicit solvent and with full electrostatics which was undertaken to characterize the structural and dynamical properties of this enzyme with emphasis on the properties of its hexameric association and its active sites' accessibility.

## 2. Computational Methods

**2.1. System Preparation.** Starting from the crystallographically determined coordinates of the *BcZBP* hexamer (PDB entry 2IXD) missing side-chain and hydrogen atoms were built with the program PSFGEN from the NAMD distribution[6] and assuming a neutral pH. The histidine residues protonation state was determined according to their chemical environment in the crystal structure. An explicit solvent hexagonal periodic boundary system was prepared using VMD.[7] The unit cell basis vectors (projections along the orthogonal axes) of the periodic cell were (111,0,0), (0,90,52), and (0,0,103) Å with a shortest (initial) solute−solute distance of 30 Å. The solvation system comprised 25938 pre-equilibrated TIP3 water molecules, with the crystallographically determined waters retained throughout the procedure, while those water molecules lying closer than 1.8 Å from the protein surface (or the crystallographic waters) were removed. The net charge of the solute was neutralized through the addition of sodium and chloride ions to a final concentration of ~100 mM corresponding to the addition of 33 sodium and 15 chloride ions. The final system comprised a total 99744 atoms, of which 21834 protein atoms, 6 zinc ions, 42 acetate atoms (located at the active site), and 77814 water atoms. The topology and parameter files used throughout the system preparation were those of the CHARMM27 force field.[8] The zinc ions were modeled using the nonbonded representation[9] implemented by the CHARMM27 force field.

**2.2. Molecular Dynamics Simulation Protocol.** A 50 ns molecular dynamics simulation was performed with the program NAMD[6] using the CHARMM27 force field[8] as follows. The system was first energy minimized for 2000 conjugate gradient steps with the positions of the backbone atoms fixed, and then for another 2000 steps without positional restraints. It was then slowly heated-up to a final temperature of 298 K (with a temperature step $\Delta T = 20$ K) over a period of 66 ps with the positions of the $C_\alpha$ atoms harmonically restrained about their energy-minimized positions. Subsequently the system was equilibrated for 200 ps under NpT conditions without any restraints. This was followed by a 50 ns production NpT run with the temperature and pressure controlled using the Nosé-Hoover Langevin dynamics and Langevin piston barostat control methods as implemented by the NAMD program (and maintained at 298 K and 1 atm). The production run was performed with the impulse Verlet-I multiple time step integration algorithm as implemented by NAMD. The inner time step was 2 fs, short-range nonbonded interactions were calculated every one step, and long-range electrostatics interactions were calculated every two timesteps using the particle mesh Ewald method.[10] A cutoff for the van der Waals interactions was applied through a switching function, and SHAKE was used to restrain all bonds involving hydrogen atoms. Trajectories were obtained by saving the atomic coordinates of the whole system every 0.4 ps.

**2.3. Trajectory Analysis.** Generation of modified PSF files was performed with X-PLOR.[11] Calculation of the anisotropic fluctuations was performed with the program g_rmsf from the GROMACS suite of programs.[12] Removal of global rotations-translations, calculation of rms deviations from the experimental structure, calculation of the average trajectory structures, of the rms deviation from the average structures, of the radius of gyration, of the atomic rms fluctuations, the $C_\alpha-C_\alpha$ distance map (and the corresponding rms deviation from it), the cross-correlation matrix, and the Cartesian[13,14] and dihedral-angle[15,16] principal component analysis were performed with the program *Carma*,[17] available via http://www.mbg.duth.gr/~glykos/.

## 3. Results

**3.1. The *BcZBP* Trajectory is Stable.** The simulation was stable, both with respect to its state variables and the structure of the enzyme, with the notable exception of three

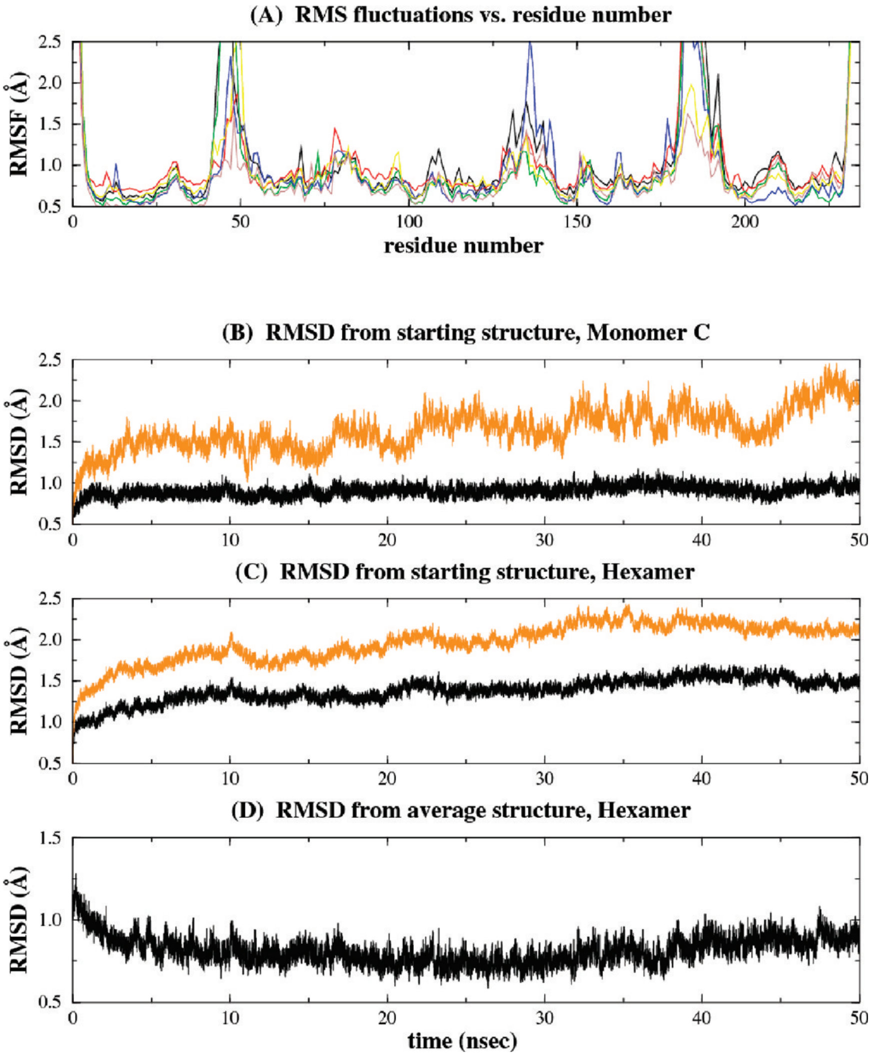


***Figure 2.*** Evolution of structure-dependent quantities during the molecular dynamics simulation. Panel A shows the root-mean-squared fluctuations (in Å) of the $C_\alpha$ atoms around their average positions during the whole length of the simulation. The graph is a superposition of six curves (corresponding to the six monomers) with the horizontal axis corresponding to residue numbers. The three highly mobile active-site loops are centered around residues 48, 135, and 186 (see text for details). For clarity this diagram is truncated to 2.5 Å along the vertical axis, with the fluctuation values reaching out to 4 and 5 Å for the first and third loop respectively. Panel B shows the rms deviation of the $C_\alpha$ atoms of monomer C versus simulation time. The two graphs shown in this panel were calculated either with the mobile loops included in the calculation (upper curve), or excluded (lower curve). Panel C is the same calculation as for panel B, but using the $C_\alpha$ atoms of the whole hexamer. Finally, panel D shows the rms deviation of the $C_\alpha$ atoms of the whole hexamer from their average positions during the length of the simulation (the three mobile loops, see text for details, have been excluded from this calculation).

surface loops which -as will be discussed later- surround the enzyme's active sites. This is shown in Figure 2A, which depicts the root-mean-squared (rms) fluctuations of the $C_\alpha$ atoms' positions (from their average) for the six monomers and for the whole length of the trajectory. As can be seen from this figure, the majority of atomic fluctuations are significantly less than 1.0 Å with the exception of (a) the first two and last three (N- and C-terminal) residues of each monomer and (b) three surface loops extending from residues 42−51, 129−140, and 180−192. The same conclusions can be drawn from the lower half of Figure 3, which depicts (using a grayscale representation) the rms deviations of the $C_\alpha$−$C_\alpha$ distances from their average during the length of the trajectory. The pronounced horizontal and vertical dark lines, which are apparent in this diagram correspond to residues with higher than average mobility with respect to the rest of the structure, and, not unexpectedly, match closely the loop regions identified from Figure 2A. The agreement between the simulation-derived atomic fluctuations and the crystallographically determined atomic temperature factors (allowing for the fact that only two monomers are crystallographically independent) is quite high, with an average value for the linear correlation coefficient (over all possible pairs of monomers) of 0.70. The highly correlated values of the experimental and simulation-derived atomic fluctuations, together with the stability of the crystallographically determined structure (discussed below), provide further indications for the validity and quality of the simulation protocol.

Examination of the linear correlation coefficients between the per-residue atomic fluctuations of the various monomers (Figure 2A), shows a notable pattern: the correlations between monomers of type I dimers are relatively low at

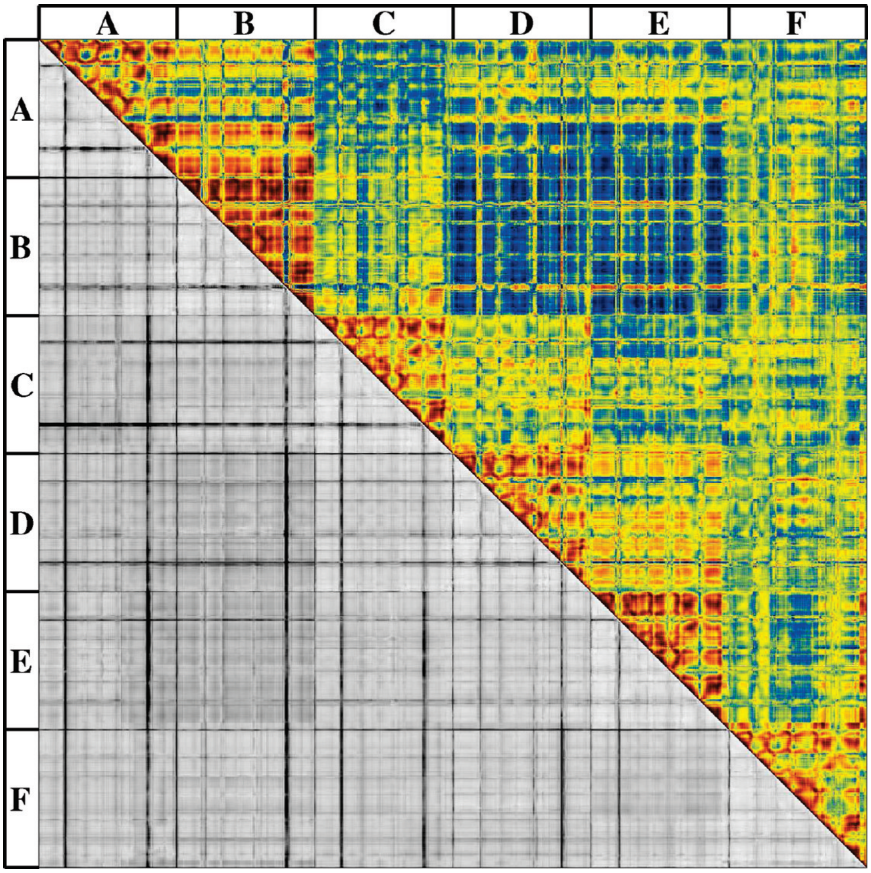


***Figure 3.*** Dynamics of the *BcZBP* hexamer. The upper triangle of the diagram is the normalized variance-covariance (cross-correlation) matrix of the $C_\alpha$ atoms. The color representation ranges from dark red, through yellow, to dark blue corresponding to correlation values from +1.0 (fully correlated), to 0.0 (uncorrelated), to −1.0 (fully anticorrelated). To increase the contrast of this diagram, a sigmoidal function of the form $\sigma(x) = 2/[1 + e^{(-3*x)}] - 1$ has been applied to the raw data. The areas of the diagram corresponding to the various monomers of the hexamer are indicated with the letters A−F at the top and left-hand-side of the matrix. The lower half of the matrix is a grayscale representation of the rms deviation of the $C_\alpha - C_\alpha$ distances (and for all possible pairs) from their average distances observed during the length of the trajectory (in other words, it is the rms deviation map of the average $C_\alpha$ -$C_\alpha$ distance map). The grayscale gradient ranges from white (corresponding to an rms deviation of 0.0 Å) to black (corresponding to an rms deviation of 3.0 Å or more). The limits in terms of the individual monomers of the hexamer are shown on the top and left-hand-side of the matrix.

0.70, 0.64, and 0.42, lower than the correlations between type II dimers (at 0.89, 0.75, and 0.46). Moreover, they are both lower than the values observed for monomers belonging to the same trimer (0.97, 0.78, 0.85 for the A−C−E trimer, 0.84, 0.68, 0.82 for the B−D−F trimer). The higher values of the linear correlation coefficient between monomers belonging to the same trimer is consistent with the results discussed in the next section.

Ignoring the three highly mobile loops, the structures of both the individual monomers and of their relative arrangement on the hexamer are well preserved during the simulation. As shown in the upper curve of Figure 2B for a representative monomer (monomer C), the rms deviation from the starting (crystal) structure increases steadily throughout the simulation, reaching values close to 2.5 Å. If the three surface loops surrounding the active site are excluded from the calculation, the results are significantly different (Figure 2B, lower curve): the rms deviation from the crystal structure quickly stabilizes to a value of approximately 0.8 Å and remains stable throughout the simulation. Similarly, Figure 2C compares the behavior of the rms deviation from the crystal structure with or without the active site loops, but this time considering the $C_\alpha$ atoms of the whole hexamer. As can be seen from this figure, the effect of excluding the active site loops from the calculation is again pronounced, though less dramatic when compared with the monomer-derived results. This indicates that there is a contribution to the rms deviation arising from the intermonomer association. Again, this is in agreement with the $C_\alpha - C_\alpha$ distance deviation map (lower half of Figure 3) which clearly shows whole areas with higher than average rms deviations (note, for example, the darker area corresponding to vectors between the $C_\alpha$ atoms of the B monomer and those of monomers E and D).

In agreement with the results presented above, Figure 2D shows the evolution (as a function of simulation time) of the rms deviation between the trajectory's average structure and each and every of the structures observed during the simulation (considering $C_\alpha$ atoms only). As can be seen, the structure quickly (within 5 ns) relaxes from the initial crystal structure, and then remains stably close to its average with deviations of approximately 0.6 Å going up to 0.8 Å near the end of the trajectory. The major differences between the starting (crystal structure) and the trajectory's average structure can be accounted by a concerted relaxation of the intermonomer association as indicated by an increase of the mass-weighted radius of gyration from the starting value of 31.4 Å to a value of 32.1 Å after only 5 ns of simulation time. Nevertheless, the rms deviation between the $C_\alpha$ atoms' positions in the average and crystal structures remains quite low at 1.2 Å (excluding the active site loops).

**3.2. The Hexameric Association Dynamics Are Complex.** The upper half of Figure 3 shows a pseudocolor representation of the hexamer's normalized variance-covariance (cross-correlation) matrix. Apart from the (expected) trend of strong positive correlations between atoms belonging to the same monomer, the cross-correlation pattern appears to be rather complex and inconclusive, especially about the mode of intermonomer association within the hexamer. If we compare the cross-correlation patterns observed for the type I dimers (A−B, C−D, E−F) with those observed for the type II dimers (B−C, D−E, A−F), we find that none of these two association models is conclusively supported by the simulation: there is strong positive correlation for the (type I) A−B dimer, but negative for the (also type I) E−F dimer. Similarly, there is positive correlation for the type II D−E dimer but negative for the (also type II) B−C dimer.

The case of the E−F dimer warrants additional discussion with respect to the $\beta$-strand-exchange dimerization motif: focusing in the area of the matrix corresponding to cross-correlations between the E and F monomers, note the thin band of positive correlation connecting the C-terminus of the F monomer with the whole of E. Similarly, there is a thin band of positive correlation connecting the C-terminus of the E monomer with the whole of F. Apart from these two bands, the rest of the matrix in this area shows either uncorrelated or even anticorrelated motion. What this implies is that, at least for the case examined here, the exchanged strands became integral parts of the monomers that receive them, and that they do not affect the independence of dynamics of the associating monomers. Clearly, even a dimerization motif as strong and explicit as a $\beta$-strand-exchange, can be surprisingly malleable with respect to protein dynamics.

Turning our attention to the trimers, we note what is possibly the most persistent characteristic of the matrix: cross-correlations between monomers related by the molecular 3-fold axis are mostly negative (see the matrix areas defined by the monomer pairs A−C, A−E, and C−E for the first trimer, B−D, B−F, and D−F for the second trimer). This motif of anticorrelations is consistent with a 'breathing' motion of the trimers about the molecular 3-fold axis, in agreement with the results from the principal component analysis discussed in the next section.

For completeness we should note that the hexamer dynamics model that best agrees with the variance-covariance matrix is a rather unexpected one, as it involves the interpretation of the data in terms of two dimers (of different types) and of two independent monomers: Referring to Figure 3, the most prominent feature of the cross-correlation matrix is the band of negative correlations connecting the A−B pair of monomers with the D−E pair. Additionally, the cross-correlations between the monomers A and B on one hand, and monomers D and E one the other, are among the most strongly positive of the matrix. Lastly, monomers C and F appear to be mostly uncorrelated with all other monomers of the hexamer. Taking these indications together, they appear to suggest a 1 + 2 + 2 + 1 model for the hexamer dynamics: four monomers (A, B, D, E) form a dimer of dimers (A−B + D−E), which is capped on either side by two independent monomers (C and F). This arrangement can be indirectly visualized from Figures 1B and C if it is assumed that these two views are related by a rotation of 180° about the 3-fold axis: the two semitransparent monomers correspond to monomers C and F, which cap the two (colored) dimers, one dimer of type I (Figure 1B) and one of the other of type II (Figure 1C). Although this model appears to explain most of the features of the cross-correlation matrix, it is difficult (if not fundamentally impossible) to reconcile with the intramolecular 32 symmetry. Indeed, a trajectory of a stable, symmetric, homo-hexameric protein at equilibrium should, if sufficiently sampled, give a cross-correlation matrix obeying the intramolecular symmetry. We attribute the absence of symmetry from the matrix to the insufficient sampling of our trajectory as discussed below.

**3.3. Principal Component Analysis and Sufficient Sampling.** The relatively low and rather stable rms deviations shown in Figure 2 (both in terms of the starting structures and of the trajectory-average structures) may leave the impression that the protein dynamics (especially at the monomer level, Figure 2B) have been sufficiently sampled during the simulation. As Figure 4 clearly indicates, this is definitely not so: the projections of the $C_\alpha$ atoms' fluctuations on the planes of their principal components deviate significantly from two-dimensional Gaussians centered at the origin (which is what we would expect from the trajectory of a sufficiently sampled single-state protein structure at equilibrium). Although there is a clear and significant difference between the extend of fluctuations of the monomers and of the higher-order oligomers, even the monomers' principal component projections show fine structure inconsistent with sufficiently sampled dynamics. It could be argued that this fine structure may correspond to functionally important discrete conformational states of the *BcZBP* monomers (with this line of argument being easily expandable to the whole hexamer). But, if this were indeed the case and because of the presence of intramolecular symmetry, we would expect these substates to be correlated between the different monomers. A cursory examination of the monomer diagrams in Figure 4 indicates that this is probably not the case. To resolve the matter in a quantitative way, we calculated the overlap between the eigenvector-defined subspaces for all possible monomer combinations and for the three eigenvectors corresponding to the three largest eigenvalues. The overlap between the subspaces defined by two sets **v** and **w** of *n* eigenvectors is defined as

$$\text{overlap}(\mathbf{v}, \mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{v}_i\cdot\mathbf{w}_j)^2$$

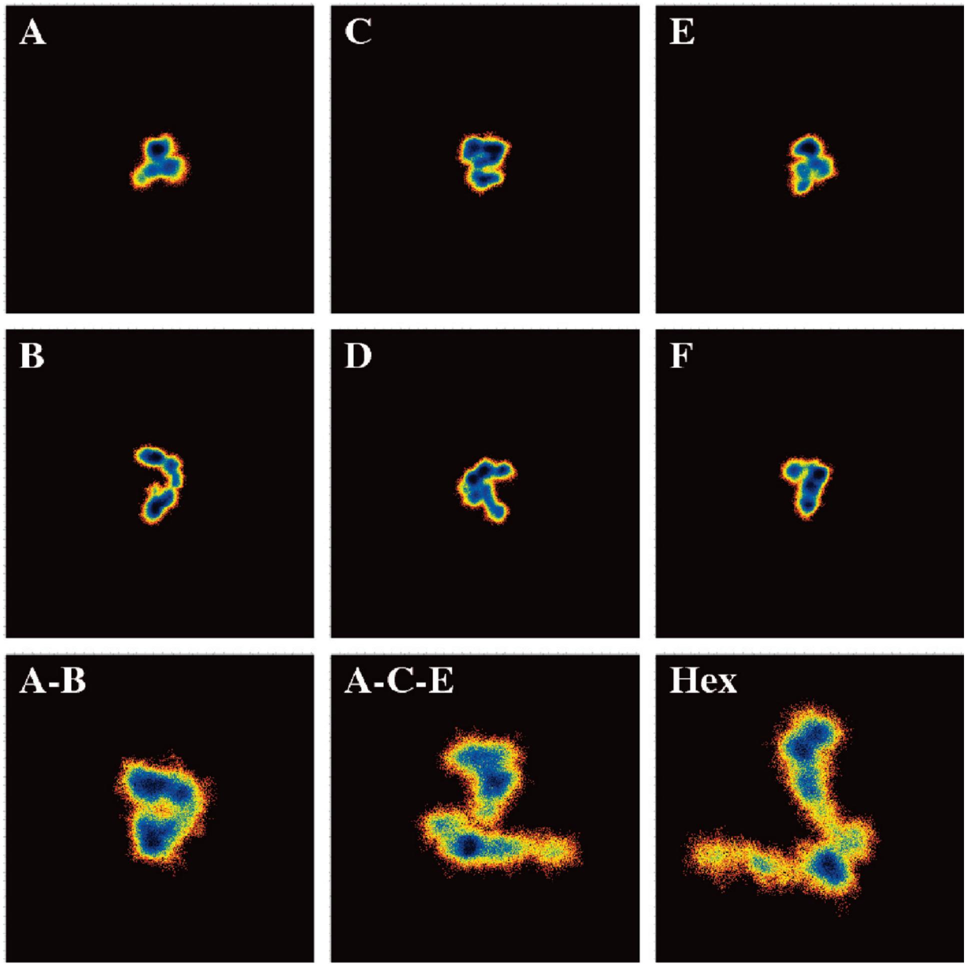


***Figure 4.*** Cartesian principal component analysis: first vs second principal component plots for the *BcZBP* monomers (top two rows), the hexamer (lower right), and a representative dimer and trimer (last row). All diagrams shown in this figure are pseudocolor representations of density functions corresponding to the projections of the fluctuations of the $C_\alpha$ motion (excluding the active-site loops) on the planes of the top two eigenvectors (of the respective molecular species indicated in the figure). For all graphs the origin is on the upper, left-hand side corner, values on all axes range from −35 to 35 Å and the eigenvectors corresponding to the largest eigenvalues are along the vertical axes. The density function shown is $\Delta G = -k_B T \ln (p/p_{max})$ where $k_B$ is Boltzmann's constant, $T$ is the temperature in Kelvin, and $p$ and $p_{max}$ are probabilities obtained from the distribution of the principal components for each structure (frame) from the corresponding trajectory. As a result of applying this function, the diagrams have units in kcal/mol with corresponding values for the minimum of the A monomer of −3.51 kcal/mol, for B −3.30, C −3.19, D −3.27, E −3.52, F −3.35, for the A−B dimer −2.56, for the A−C−E trimer −2.47, and, finally, for the hexamer −2.37 kcal/mol. Note that these values can only be compared between trajectories of the same molecular species (in this case, only between monomers). Please also note that the $\Delta G$ values obtained from this procedure are on an arbitrary scale in the sense that they depend on the binning procedure used for calculating the $p$ and $p_{max}$ values. For all diagrams of this figure, the raw data were binned on a square matrix of size $\sqrt{N}/2$ where $N$ is the number of frames of the corresponding trajectory.

and takes values from zero (signifying no convergence of the corresponding subspaces) to one (for full overlap of the subspaces). Over all possible monomer combinations, the average overlap value (for the top three eigenvectors) was only 0.05 (with a standard deviation of 0.02). The highest value observed was 0.10 between monomers C and D. What these results clearly show is that the length of the simulation has been inadequate to sample sufficiently even the monomers' dynamics, let alone the whole of the *BcZBP* hexamer. This is in agreement with the indications obtained from the cross-correlation matrix (see last paragraph of the previous section). Similar results to those discussed above have been obtained from a principal component analysis performed in dihedral ($\phi$, $\psi$) angle space (which is not sensitive to rigid-body-like motion of protein domains or subdomains).

With the precautions necessitated by the lack of sufficient sampling discussed above, we note the systematic difference between the projections of fluctuations for monomers belonging to different trimers (Figure 4, first row of graphs vs second row): the extent of the atomic fluctuations on the eigenvector planes are correlated at the level of the two trimers, but not at the level of the type I or II dimers (the fluctuations are systematically lower for the A−C−E trimer compared with the B−D−F trimer). Such a systematic difference at the trimers' level was also observed when considering the fluctuations from the average structures in

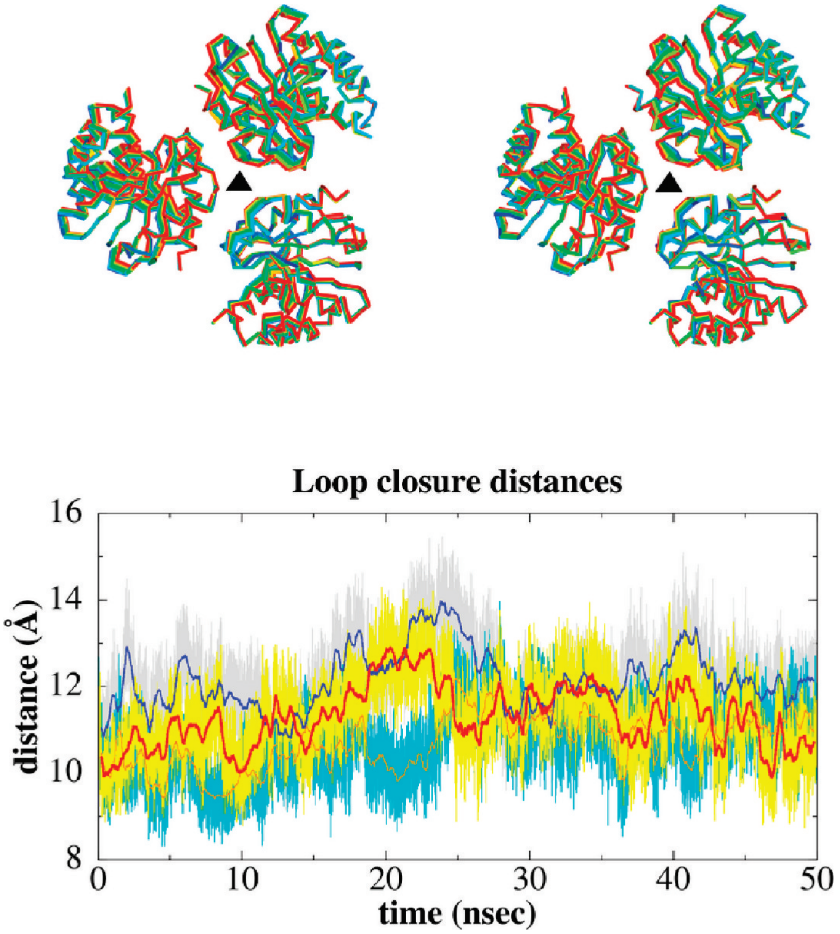


***Figure 5.*** Trimer dynamics: the upper panel stereodiagram (wall-eyed) is a smooth representation of the trimer's $C_\alpha$ motion (excluding the active-site loops) as calculated from the first principal component only. The intramolecular 3-fold axis is perpendicular to the plane of the paper and its position is noted by the filled triangle. The various structures shown superimposed are color-coded from blue, via green and yellow, to red, and correspond to the structures obtained by applying (on the average structure) the fluctuations corresponding to the first eigenvector weighted by a smoothly varying amplitude (obtained from the principal component analysis, and ranging for this diagram from −19 to 19 Å). The graphs in the lower panel show the variation of the distances between the tips of the three loops closest-to and surrounding the 3-fold axis (the loops immediately next to the filled triangle in the upper panel). The actual distances (as observed in the trajectory) are shown as light-colored backgrounds. The solid lines are averages which were calculated using a 40 ps window.

section 3.1 above, and also during the analysis of the variance-covariance matrix in section 3.2.

These indications concerning correlated dynamics at the trimers' level, prompted us to examine in more detail the molecular motion associated with the trimers' principal components. The eigenvector with the largest contribution to the intermonomeric association dynamics is the first one (data not shown). The top panel of Figure 5 shows a smooth representation of the ACE trimer's motion due to the first eigenvector considering only the $C_\alpha$ atoms and ignoring the flexible active-site loops. As can be seen from this figure, there are indeed some indications of an anticorrelated motion of the monomers about the molecular 3-fold axis. Such a breathing-like motion of the monomers was also observed when discussing the differences between the crystal structure and the trajectory-average structure (section §3.1). To quantify this statement, we have calculated —as a function of simulation time— the distances between the $C_\alpha$ atoms of the residues located at the tips of the loops which are closest to the molecular 3-fold axis [Note that these distances were obtained directly from the molecular dynamics trajectory and not from the principal component-derived motion shown in the top panel of Figure 5]. If the notion of a breathing-like motion was supported by the raw simulation data, then these distances should be correlated. The results from this calculation are shown in the lower graph of Figure 5. As can be seen from this graph, the variation of the loop-closure distances is indeed correlated, but not uniformly: the linear correlation coefficient between the A−C and C−E distances is 0.40, but is reduced to 0.20 for the A−C and A−E monomers, and to 0.12 for the A−E, C−E combination. Taking the results from these calculations together, they seem to be consistent, at least within the limitations posed by the lack of sufficient sampling, with the notion of a breathing-like motion of the monomers about the molecular 3-fold.

The discussion above, together with the absence of symmetry from the cross-correlation matrix shown in Figure 3, may create the impression that the symmetry of the hexamer is not well preserved during the simulation. To unequivocally show that this is not the case, we examined the symmetry and structural conservation of the whole enzyme using for our calculations its average structure calculated over the length of the trajectory. This we did as follows: In the first step, the average structure of monomer A was least-squares superimposed on the average structures of the other five monomers and the rms deviations between their $C_\alpha$ atoms (excluding the flexible loops) were recorded. The values we obtained were 0.57 Å for the A−B monomers, 0.31 Å for A−C, 0.51 Å (A−D), 0.39 Å (A−E) and 0.44 Å between the A and F monomers. This clearly shows that the average structures of the six monomers are practically identical, making unnecessary to calculate superpositions for all their possible pairwise combinations. To examine the preservation of symmetry, we converted the rotation matrices (obtained from the least-squares superposition) to their equivalent sets of polar angles ($\omega$, $\phi$, $\kappa$) and examined the deviations of these angles from the values expected for a hexamer of 32 symmetry (for, example, rotations with $\Delta\kappa = 120°$ for the 3-fold axes or $\Delta\kappa = 180°$ for the two-folds). The average deviations of these polar angles from their ideal values (for a perfect 32 hexamer) were as low as 0.9 degrees, showing that the hexameric symmetry is almost perfectly conserved during the simulation. This, however, creates a conceptual problem: if both the symmetry and the structure of the enzyme are so highly conserved, why the variance-covariance matrix is not symmetrical (or, equivalently, why its derived principal components have not converged) ? The answer, we believe, is that the variance-covariance matrix (and its principal components) are dominated by the small-scale fluctuations of a very stable (but large) structure, making convergence difficult to achieve within the time scale of our simulation. To show that it is indeed the small-scale fluctuations that dominate the PCA calculation, we recalculated the variance−covariance matrix, but this time we did not normalize it, keeping its units in Å$^2$. Taking the average of the absolute values of the matrix (and excluding all intramonomer correlations), we obtained a value of only 0.106 Å$^2$ with a standard deviation of 0.166 Å$^2$. If the flexible loops are excluded from the calculation, then we obtain an average value of 0.077 Å$^2$ with a standard deviation of only 0.069 Å$^2$. These results clearly show that the major contribution to the variance−covariance matrix (and its principal components) is not large-scale correlated motion (which, if present, would mask the absence of convergence for the smaller-scale motions), but small-scale fluctuations about an otherwise stable average structure.

**3.4. Active-Site Loops: Mobility and Accessibility.** The stereodiagram shown in Figure 6B illustrates in a direct and immediate way what has already been mentioned on several occasions in the previous sections: the *BcZBP* structure, both at the monomeric and oligomeric levels, is very well preserved during the simulation with the exception of the three loops that surround the enzyme's active sites. As can be inferred from Figure 6B, both the structural core of the monomers and their relative orientations in the oligomer are very stable as indicated by the excellent superposition of the structures and the low rms fluctuations (indicated by the dark blue color). In contrast, the three loops surrounding the active sites, which are marked by the space-filling model of the zinc atom, are exceedingly mobile as evidenced both by the divergence of the various structures and the high rms fluctuation values (indicated by the dark red color). For each active site, the three loops surrounding it are contributed by two neighboring (at the trimers' level) monomers: the monomer to which the active site belongs contributes two loops, the first loop comprising residues 42−51 and the second residues 180−192. The neighboring monomer contributes the loop extending from residue 129 to residue 140. We will hereafter refer to these three loops as $L_{46}$, $L_{185}$, and $L_{135}$. Comparison of the loop mobility for the three active sites shown in Figure 6B shows a notable pattern: loops $L_{46}$ and $L_{185}$ have consistently higher mobility than $L_{135}$ (this can also be inferred from the graphs shown in Figure 2A). Additionally, the amount of mobility observed for $L_{46}$ and $L_{185}$ varies significantly between the various active sites (compare, for example, the two active sites that are located in the upper part of Figure 6B). Although the presence of this variability in the atomic fluctuations may be connected with the limited sampling discussed in section 3.3, the molecular dynamics trajectory per se is remarkably self-consistent with respect to the presence and the amplitude of fluctuations of the hypermobile loops. To quantify this statement we proceeded as follows: The rms atomic fluctuations of all 1386 $C_\alpha$ atoms of the protein were calculated for two disjoined trajectory segments extending from 10 to 30 ns (for the first segment), and from 30 to 50 ns (for the second). The value of the linear correlation coefficient between the atomic fluctuations obtained from these two segments was as high as 0.852, clearly indicating that the molecular dynamics trajectory is internally consistent with respect to the presence of the hypermobile loops. Additionally, the fluctuations obtained from the two segments are in excellent agreement with the results obtained from the whole trajectory (and shown in Figure 2A) with corresponding values of the linear correlation coefficient of 0.954 and 0.916.

It could be argued that the amount of loop mobility observed in the trajectory is not the result of the protein dynamics per se, but arises as an artifact of the nonbonded representation of the zinc ions used for modeling the enzyme's active sites.[18] This is clearly not the case for two reasons. The first reason is that the protein residues involved in zinc coordination (residues 12, 15 and 113) are outside the limits of the mobile loops as described above. The second and more important reason is that the geometries of all six active sites are very highly conserved. To quantify this statement, we calculated the rms deviation from the starting (crystal) structure for all non-hydrogen atoms of all protein residues that are involved in the zinc ion coordination. In the case, for example, of monomer C, the mean rms deviation (and for the whole length of the trajectory) was only 0.47 Å with a standard deviation of 0.05 Å, comparable with the expected coordinate error of the crystal structure. Such low rms deviations for the active site residues clearly indicate that the presence of hypermobile loops is not in any way connected with the model chosen for the representation of the zinc ions.

The amount of loop mobility seen on Figure 6B, immediately possess the question of whether the motion of the

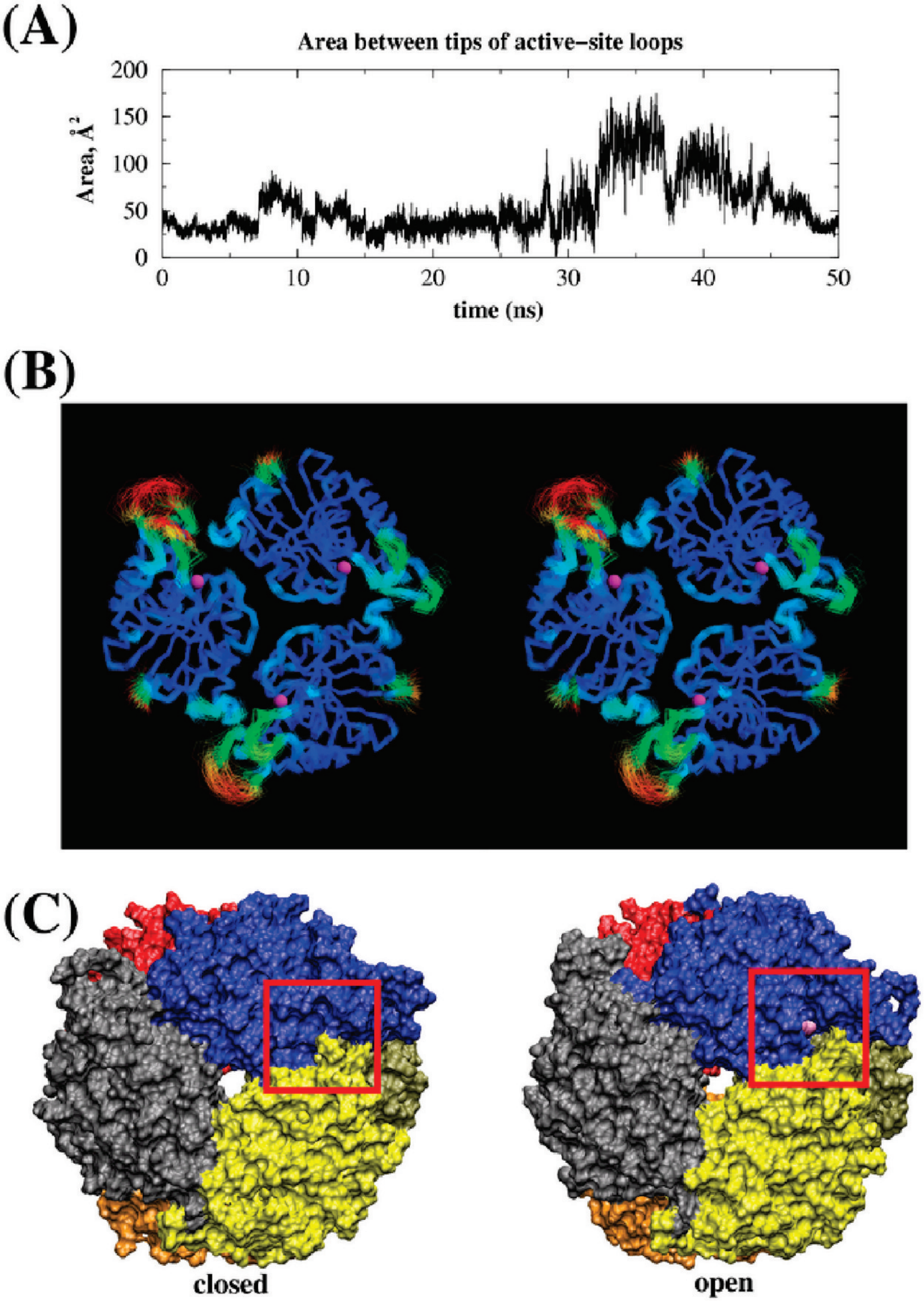


***Figure 6.*** Active-sites' loop mobility and accessibility. Panel (A) shows the variation (as a function of simulation time) of the area of a triangle defined by the $C_\alpha$ atoms of Asp184 and Ser46 from monomer A, and Lys131 from monomer E. These three residues lie at the tips of the three loops surrounding one of the enzyme's active sites. Panel (B) is a stereodiagram (wall-eyed) illustrating the mobility of the loops surrounding the active sites. The diagram corresponds to a superposition of structures obtained directly from the molecular dynamics trajectory (after removal of overall rotations-translations). The view is down the molecular 3-fold axis, and to reduce clutter only one trimer is shown. The position of the active sites is marked by the space-filling models (colored magenta) of the zinc atoms. The structures are colored according to their atomic (per $C_\alpha$) rms fluctuations using a linear gradient ranging from dark blue to dark red. Finally, panel (C) shows two space filling models of the whole hexamer (taken directly from the trajectory) illustrating the loop-dependent opening and closing of one of the active sites (see boxed area of the diagrams).

active-site loops is correlated or not. This would have functionally important implications since the presence of correlated motion would suggest that the loops undergo a concerted movement (as expected, for example, from a periodic opening and closing of the active sites). To tackle this question we once again resorted to the cross-correlation matrix shown in Figure 3, this time examining only those parts of the matrix that correspond to the $L_{46}$, $L_{185}$ and $L_{135}$ loops. To avoid qualitative assessments we proceeded as follows: In the first step, the entries of the matrix corresponding to the cross-correlation values for the $C_\alpha$ atoms of the three loops were isolated. In the second step, we selected only those entries that corresponded to cross-correlation values between $C_\alpha$ atoms of different loops (that is, we excluded values between atoms belonging to the same loop). In the final step, we grouped these values with respect to the active sites that the loops belonged to. The net result of this procedure is a set of cross-correlation values between the $C_\alpha$ atoms of each loop, with each and every $C_\alpha$ atom of the two other loops that surround the same active site. The numerical results obtained from these calculations were conclusive: the average value of the cross-correlation coefficient between atoms surrounding, for example, the active site located between the A and E monomers is only 0.03 with a standard deviation of 0.14. Similar results have been obtained for all the other active sites of *BcZBP*. The implication of the preceding analysis is that the opening and closing of the enzyme's active sites is a stochastic process, dependent on the random conformational changes of the three loops.

Consistent with the notion of a stochastic process, Figure 6A shows the variation of the area of a triangle defined by the tips of the three loops surrounding the active site located between the A and E monomers. The area defined by the three loops remains more or less stable at ~45 Å$^2$ for the first 30 ns of the trajectory, and then an opening event is recorded lasting for approximately 17 ns. During this opening event, the area defined by the three loops more than triples, reaching values as high as ~170 Å$^2$. To show unequivocally the extend of this significant change in the active sites' accessibility, Figure 6C compares snapshots (recorded directly from the trajectory) taken at 5 ns (diagram on the left, active site closed) and at 37 ns (diagram on the right, active site open). This very notable change in the active sites' accessibility as seen in Figure 6C is amplified even further if it is considered that the plane defined by the tips of the three active site loops is not perpendicular to the viewing axis (and so, the observed amount of opening is only a projection of the real difference in active-site accessibility).

**3.5. Active-Site Loops: Conservation and Specificity.** Figure 7 shows a multiple sequence alignment of *BcZBP* with its three homologues from *B. anthracis* together with a per residue conservation and quality score as calculated by the program Jalview[20] (note that the alignment shown is the unedited result of a default run of the program T-coffee[19]). The correspondence between the three hypermobile active-site loops and the parts of the alignment with accumulated gaps and low conservation score is striking, especially for the $L_{46}$ and $L_{135}$ loops. Indeed, if the proteins' termini are excluded from consideration, then almost all low conservation regions from the alignment match exactly the three active-site loops. An exception to this observation is the area centered around *BcZBP*'s Lys154 which also shows very low conservation and quality score. Although this region is also close to the active site, examination of the *BcZBP* structure suggests that it is rather unlikely for this area to be directly involved with the active site accessibility and/or specificity.

Given the observed accumulation of insertions/deletions and the low conservation score of the loops surrounding the active site, it is tempting to speculate that these loops not only contribute to defining the accessibility to the active sites (see previous section and Figure 6A and C), but may also be implicated in determining the enzymes' substrate specific-

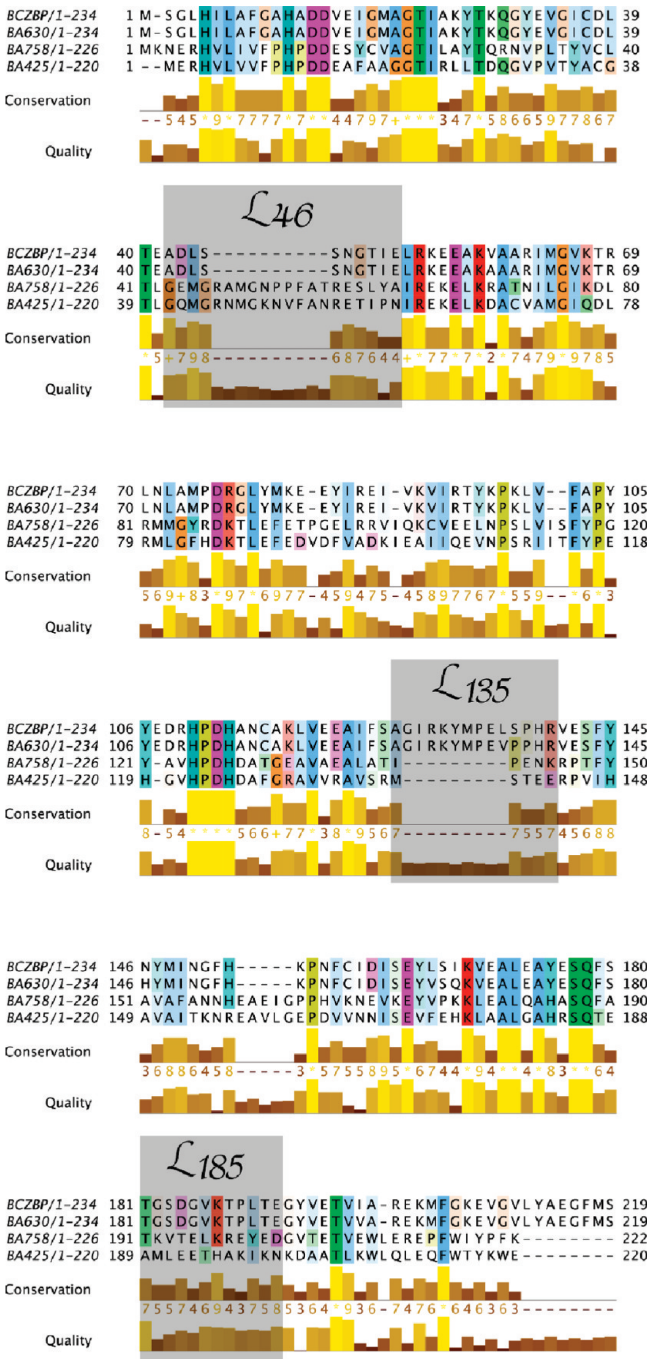

BCZBP/1-234 1 M-SGLHILAFGAHADDVEIGMAGTIAKYTKQGYEVGICDL 39
BA630/1-234 1 M-SGLHILAFGAHADDVEIGMAGTIAKYTKQGYEVGICDL 39
BA758/1-226 1 MKNERHVLIVFPHPDDESYCVAGTILAYTQRNVPLTYVCL 40
BA425/1-220 1 --MERHVLVVFPHPDDEAFAAGGTIRLLTDQGVPVTYACG 38
Conservation
--545*9*7777*7**44797+***347*58665977867
Quality
L46
BCZBP/1-234 40 TEADLS----------SNGTIELRKEEAKVAARIMGVKTR 69
BA630/1-234 40 TEADLS----------SNGTIELRKEEAKAAARIMGVKTR 69
BA758/1-226 41 TLGEMGRAMGNPPFATRESLYAIREKELKRATNILGIKDL 80
BA425/1-220 39 TLGQMGRNMGKNVFANRETIPNIREKELKDACVAMGIQDL 78
Conservation
*5+798----------687644+*77*7*2*7479*9785
Quality
BCZBP/1-234 70 LNLAMPDRGLYMKE-EYIREI-VKVIRTYKPKLV--FAPY 105
BA630/1-234 70 LNLAMPDRGLYMKE-EYIREI-VKVIRTYKPKLV--FAPY 105
BA758/1-226 81 RMMGYRDKTLEFETPGELRRVIQKCVEELNPSLVISFYPG 120
BA425/1-220 79 RMLGFHDKTLEFEDVDFVADKIEAIIQEVNPSRIITFYPE 118
Conservation
569+83*97*6977-459475-45897767*559--*6*3
Quality
L135
BCZBP/1-234 106 YEDRHPDHANCAKLVEEAIFSAGIRKYMPELSPHRVESFY 145
BA630/1-234 106 YEDRHPDHANCAKLVEEAIFSAGIRKYMPEVPPHRVESFY 145
BA758/1-226 121 Y-AVHPDHDATGEAVAEALATI---------PENKRPTFY 150
BA425/1-220 119 H-GVHPDHDAFGRAVVRAVSRM---------STEERPVIH 148
Conservation
8-54****566+77*38*9567---------755745688
Quality
BCZBP/1-234 146 NYMINGFH-----KPNFCIDISEYLSIKVEALEAYESQFS 180
BA630/1-234 146 HYMINGFH-----KPNFCIDISEYVSQKVEALEAYESQFS 180
BA758/1-226 151 AVAFANNHEAEIGPPHVKNEVKEYVPKKLEALQAHASQFA 190
BA425/1-220 149 AVAITKNREAVLGEPDVVNNISEVFEHKLAALGAHRSQTE 188
Conservation
36886458-----3*5755895*6744*94**4*83**64
Quality
L185
BCZBP/1-234 181 TGSDGVKTPLTEGYVETVIA-REKMFGKEVGVLYAEGFMS 219
BA630/1-234 181 TGSDGVKTPLTEGYVETVVA-REKMFGKEVGVLYAEGFMS 219
BA758/1-226 191 TKVTELKREYEDGVTETVEWLEREPFWIYPFK-------- 222
BA425/1-220 189 AMLEETHAKIKNKDAAILKWLQLEQFWTYKWE-------- 220
Conservation
7557469437585364*936-7476*646363--------
Quality

***Figure 7.*** Multiple sequence alignment of *BcZBP* with its three *B. anthracis* homologues. The three hypermobile active site loops are noted with the shaded boxes and are marked as $L_{46}$, $L_{185}$, and $L_{135}$ (see text for details). The *B. anthracis* proteins are denoted as: BA630 corresponding to NP_844007 (gi: 30261630), BA758 corresponding to NP_846135 (gi: 30263758), and BA425 corresponding to NP_845802 (gi: 30263425). The multiple sequence alignment shown is the unedited result from a default run of the program T-coffee.[19] The conservation and quality scores are as produced by the program Jalview.[20] The coloring of the amino acids corresponds to the clustal color scheme as implemented by Jalview. The portion of the alignment extending beyond the end of the shortest sequence (BA425) is not shown for clarity.

ity. Clearly, the specificity-related clause of the previous sentence may appear as an overinterpretation of the data, especially when it is partly based on something as inherently inconclusive as a multiple sequence alignment. It could be argued, for example, that accumulation of insertions/deletions (and the corresponding low conservation scores) is exactly what we would expect from surface exposed loops with no functional or structural importance (and, thus, with low pressure from natural selection). Although this would be an otherwise valid argument, we find it hard to reconcile this view with the image of the dynamics of these loops as seen in Figure 6B and C. Indeed, it appears highly unlikely that of all surface exposed loops present in the *BcZBP* structure, only those loops surrounding the active sites have no functional importance and thus escape the pressure of natural selection. It should be noted, however, that in the absence of solid experimental evidence in the form of a crystallographically determined structure of an enzyme−substrate complex, it is impossible to take this analysis much further. This is more so given the absence of firm knowledge concerning the specific substrate of *BcZBP*.

## 4. Discussion

We have performed a state-of-the-art molecular dynamics simulation of *BcZBP* in explicit solvent and with full electrostatics. Analysis of the resulting trajectory showed not only that the simulation per se was very stable but also that the overall structure of the enzyme was very well preserved with an average rms deviation at the monomers' level of approximately 0.80 Å. Analysis of the variance-covariance matrix showed that the crystal structure-based view of the enzyme as a trimer of dimers does not convey the complex dynamics observed in the simulation and indicated that even a dimerization motif as strong and explicit as a $\beta$-strand-exchange, can show surprising plasticity with respect to protein dynamics. Analysis of the pattern of atomic mobility and fluctuations identified three hyper-mobile regions of the *BcZBP* structure corresponding to the three loops surrounding each of the hexameric enzyme's active sites. Examination of their mobility clearly indicated that at least in the case of the apoenzyme, these loops are directly implicated in determining the active site accessibility. Comparison of the molecular dynamics results with the indications obtained from a multiple sequence alignment with the *B. anthracis* homologues, led to the hypothesis that these three active-site loops may be implicated in determining the enzyme's substrate specificity.

**Acknowledgment.** We should like to thank the Research Center of Scientific Simulation of the University of Ioannina for provision of computational resources.

### References

(1) Read, T. D.; Peterson, S. N.; Tourasse, N.; Baillie, L. W.; Paulsen, I. T.; Nelson, K. E.; Tettelin, H.; Fouts, D. E.; Eisen, J. A.; Gill, S. R.; Holtzapple, E. K.; Okstad, O. A.; Helgason, E.; Rilstone, J.; Wu, M.; Kolonay, J. F.; Beanan, M. J.; Dodson, R. J.; Brinkac, L. M.; Gwinn, M.; DeBoy, R. T.; Madpu, R.; Daugherty, S. C.; Durkin, A. S.; Haft, D. H.; Nelson, W. C.; Peterson, J. D.; Pop, M.; Khouri, H. M.; Radune, D.; Benton, J. L.; Mahamoud, Y.; Jiang, L.; Hance, I. R.; Weidman, J. F.; Berry, K. J.; Plaut, R. D.; Wolf, A. M.; Watkins, K. L.; Nierman, W. C.; Hazen, A.; Cline, R.; Redmond, C.; Thwaite, J. E.; White, O.; Salzberg, S. L.; Thomason, B.; Friedlander, A. M.; Koehler, T. M.; Hanna, P. C.; Kolsto, A. B.; Fraser, C. M. The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* **2003**, *423*, 81–86.

(2) Ivanova, N.; Sorokin, A.; Anderson, I.; Galleron, N.; Candelon, B.; Kapatral, V.; Bhattacharyya, A.; Reznik, G.; Mikhailova, N.; Lapidus, A.; Chu, L.; Mazur, M.; Goltsman, E.; Larsen, N.; D'Souza, M.; Walunas, T.; Grechkin, Y.; Pusch, G.; Haselkorn, R.; Fonstein, M.; Ehrlich, S. D.; Overbeek, R.; Kyrpides, N. Genome sequence of Bacillus cereus and comparative analysis with *Bacillus anthracis*. *Nature.* **2003**, *423*, 87–91.

(3) Psylinakis, E.; Boneca, I. G.; Mavromatis, K.; Deli, A.; Hayhurst, E.; Foster, S. J.; Varum, K. M.; Bouriotis, V. Peptidoglycan *N*-acetylglucosamine deacetylases from *Bacillus cereus*, highly conserved proteins in *Bacillus anthracis*. *J. Biol. Chem.* **2005**, *280*, 30856–30863.

(4) Fadouloglou, V. E.; Kotsifaki, D.; Gazi, A. D.; Fellas, G.; Meramveliotaki, C.; Deli, A.; Psylinakis, E.; Bouriotis, V.; Kokkinidis, M. Purication, crystallization and preliminary characterization of a putative LmbE-like deacetylase from *Bacillus cereus*. *Acta Crystallogr.* **2006**, *F62*, 261–264.

(5) Fadouloglou, V. E.; Deli, A.; Glykos, N. M.; Psylinakis, E.; Bouriotis, V.; Kokkinidis, M. Crystal structure of the BcZBP, a zinc-binding protein from Bacillus cereus: Functional insights from structural data. *FEBS J.* **2007**, *274*, 3044–3054.

(6) Kale, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. NAMD2: Greater scalability for parallel molecular dynamics. *J. Comput. Phys.* **1999**, *151*, 283–312.

(7) Humphrey, W.; Dalke, A.; Schulten, K. VMD−Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.

(8) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics Studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(9) Stote, R. H.; Karplus, M. Zinc binding in proteins and solution: A simple but accurate nonbonded representation. *Proteins* **1995**, *23*, 12–31.

(10) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald. An Nlog (N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(11) Brünger, A. T. Management of Trajectories. In *X-PLOR*; version 3.1, Yale University Press: New Haven, CT, 1992; pp 143−158.

(12) Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001**, *7*, 306–317.

(13) Ichiye, T.; Karplus, M. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins* **1991**, *11*, 205–217.

(14) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential dynamics of proteins. *Proteins* **1993**, *17*, 412–425.

(15) Mu, Y.; Nguyen, P. H.; Stock, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins* **2005**, *58*, 45–52.

(16) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chem. Phys.* **2007**, *126*, 244111.

(17) Glykos, N. M. Carma: A molecular dynamics analysis program. *J. Comput. Chem.* **2006**, *27*, 1765–1768.

(18) Dal Peraro, M.; Spiegel, K.; Lamoureux, G.; Vivo, M. D.; DeGrado, W. F.; Klein, M. L. Modeling the charge distribution at metal sites in proteins for molecular dynamics simulations. *J. Struct. Biol.* **2007**, *157*, 444–453.

(19) Notredame, C.; Higgins, D. G.; Heringa, J. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **2000**, *302*, 205–217.

(20) Clamp, M.; Cuff, J.; Searle, S. M.; Barton, G. J. The Jalview Java alignment editor. *Bioinformatics* **2004**, *20*, 426–427.

CT9002338

JCTC Journal of Chemical Theory and Computation

# Solvent Effects on Donor−Acceptor Couplings in Peptides. A Combined QM and MD Study

Frank Wallrapp,[†] Alexander Voityuk,[*,‡] and Victor Guallar[*,†]

*Catalan Institució Catalana de Recerca i Estudis Avançats, 08010 barcelona, and Life Science Program, Barcelona Supercomputing Center, Nexus II Building, 08028 Barcelona, Spain, and Institució Catalana de Recerca i Estudis Avançats, 08010 Barcelona, and Institut de Química Computational, Universitat de Girona, 17071 Girona, Spain; Catalan Institution for Research and Advanced Studies, Institute of Computational Chemistry, Department of Chemistry, University of Girona, 17071 Girona, Spain*

Received July 20, 2009

**Abstract:** We present a combined Quantum Chemical/Molecular Dynamics study on electronic coupling between tryptophan-based donor and acceptor in oligopeptides of variable length. Molecular dynamics was performed on Trp-(Pro)n-Trp ($n = 1$ to 6) molecules in gas phase and aqueous solvent and the electronic coupling matrix element was computed for thermal hole transfer applying semiempirical INDO/S together with the generalized Mulliken-Hush approach. For comparison, we also computed coupling values of 40 000 snapshots applying ab initio Hartree−Fock, showing good agreement with the INDO/S results. We demonstrate that the coupling values strongly fluctuate throughout the molecular dynamic trajectory and the mechanism of electron transfer is affected by the presence of solvent through restriction of the conformational space. Gas-phase calculations show gated electron transfer dominated by direct through-space coupling due to strong conformational changes bringing donor and acceptor in close vicinity. Solvent calculations establish a nongated mechanism dominated by bridge-mediated coupling. In agreement with experimental data, our results point to a donor−acceptor distance of ~20 Å as a possible point for transition from superexchange to hopping electron transfer mechanism.

## I. Introduction

Protein-mediated electron transfer (ET) between separated local donor and acceptor sites plays a central role in biochemistry.[1−5] For example, long-range electron transfer over a distance of 10 to 30 Å is a process of major importance in photosynthesis and respiration. In general, electron transfer can be considered as a transition between electronic states. Its rate is determined both by the coupling between those electronic states and by the reorganization energy needed by the system to adapt to its new state. Although the motion of the electron is in fact instantaneously on the time scale of nuclear motion, the rate of electron transfer is generally slow when compared to that. Rate constants can vary over many orders of magnitude due to the roughly exponential dependence of the electronic-state overlap on the donor−acceptor distance.[6] Variations of this distance dependence has been the subject of a dynamic discussion between experimentalists[7−9] and theoreticians[10,11] in the last years and is still ongoing. Many of these studies have focused on test systems of single-chained short oligopeptides to get an easier understanding of the underlying ET mechanism. Despite bridge-mediated superexchange between donor and acceptor, electron transfer can also occur through a process of incoherent hopping between localized electronic states on the bridge.[12] The respective contributions

[*] Corresponding authors: Victor Guallar: victor.guallar@bsc.es; Alexander Voityuk: alexander.voityuk@icrea.es.

[†] Barcelona Supercomputing Center.

[‡] University of Girona.

10.1021/ct900377j CCC: $40.75 © 2009 American Chemical Society
Published on Web 10/29/2009

of the two mechanisms are dependent on the energy difference between the donor- and bridge-localized electronic states. The total rate constant of the system is then the composite of both bridge-mediated superexchange and sequential hopping mechanism. Experimental work on oligoproline between Ruthenium based donors and acceptors has shown that electron transfer mechanism can also change from a predominantly electron superexchange to a predominantly electron hopping mechanism when the peptide spacer distance exceeds about 20 Å.[13] Furthermore, recent studies on protein electron transfer emphasize the importance of tryptophans within these multistep electron transfers.[14,15]

Next to the distance, dynamical flexibility of the protein and solvent has also a strong impact on the electronic coupling between donor and acceptor. Recently there have been several studies describing how conformational dynamics of proteins strongly influence the donor−acceptor coupling,[6,16−24] showing that these observed fluctuations can be large and sometimes within only the femtosecond time scale. Also, a recent study on through-bond electron transfer in Ru-modified azurin indicates the central role of valence angle fluctuations in coupling dephasing.[25] Furthermore, there exist recent studies on U-shaped electron transfer systems where a highly curved bridge imparts a vacant cleft along the line-of-sight between the electron donor and acceptor.[26−28] Semiempirical calculations show that electronic coupling between donor and acceptor within these model systems results mainly from direct coupling or coupling through bridging solvent molecules. This feature differs strongly from craned model systems in which the coupling is mainly due to through-bond contributions resulting in a linear decrease of direct coupling with larger donor−acceptor distance.[29−31] To capture the dynamical behavior of the electronic coupling, it is necessary to calculate the average over many snapshots of a long molecular dynamics trajectory to cover full conformational space of flexible model systems.[16]

The electronic coupling for the hole transfer process can be calculated using the one-electron or Koopmans' theorem approximation.[32,33] Within this scheme, the desired properties of the adiabatic states for a radical cation can be approximated using one-electron energies and occupied molecular orbitals of the corresponding neutral (close-shell) system.[34] Computing the electronic coupling for large systems is a challenge due to the amount of atoms involved in the electron transfer. Avoiding the computational effort to calculate donor−acceptor coupling values from ab initio calculations,[35] there exist several semiempirical approaches, which allow analyzing the coupling of large proteins over many snapshots. Among them are Extended-Hückel related methods[21,36,37] and the neglect of differential overlap (INDO/S).[38] Unlike the standard semiempirical schemes based on the NDDO approximation (MNDO, AM1 and PM3), the INDO/S method provides surprisingly good results for electronic couplings[39] and thus is widely accepted as a feasible approximation for ET calculations, being subject of much comparison against ab initio calculations.[6,40,41]

The aim of this article is to study the mechanism of thermal hole transfer in oligopeptides in gas phase as well as aqueous solution. Therefore, we calculate coupling values from electronic properties derived from semiempirical INDO/S calculations as well as more sophisticated Hartree−Fock (HF) calculations to compare both levels of theory against each other. The results show that the mechanism is highly affected by the presence of solvent by means of restricting the conformational space within the dynamics. On the basis of recent studies on protein electron transfer indicating the importance of tryptophans,[14,15] we used a set of oligoproline peptides of variable length linking tryptophan based electron donors (D) and acceptors (A) as our test system.

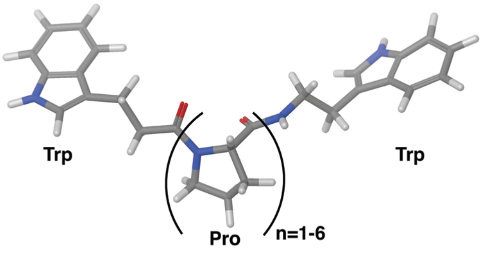


***Figure 1.*** Trp-(Pro)n-Trp in gas phase.

## II. Methods

**Modeling Oligopeptide Structure and Dynamics.** Oligopeptide Trp-(Pro)n-Trp with $n$ from 1 to 6 was assembled using Schrödinger's Maestro[42] build utility, having the amino group of the donor Trp and the acetyl group of the acceptor Trp replaced by hydrogen atoms. For each peptide, we prepared two different setups for the molecular dynamics (MD): in gas phase and in aqueous solvent. Figure 1 shows Trp-(Pro)n-Trp in gas phase as an example. For all of the systems, we performed a 10 ns NVT trajectory at 298.15 K, following a truncated Newton minimization and a short equilibration (10 ps for vacuum and 100 ps for solvent). For the $n = 4$ case, we performed an additional 30 ns trajectory in solvent to compare its mean donor−acceptor distance and mean coupling value against the 10 ns trajectory. Snapshots for the electronic coupling calculation were saved every picosecond based on a donor−acceptor autocorrelation analysis of the highly mobile system Trp-Pro2-Trp in gas phase. Further explanation to this as well as the distance autocorrelation plot is given in the Supporting Information. Vacuum minimization, equilibration, and MD simulations were performed with *Impact*[43] using a nonbonded cutoff value of 12 Å. Solvent simulation applied the SPC water model[44] in a cubic box of 10 Å buffer region, periodic boundary condition, and the Ewald summations. Solvent MD was performed with *Desmond*.[45] The applied force field throughout all the MD simulations was OPLS-2005.[46] From the production run, we extracted 10 000 snapshots, which in the solvent cases include only a layer of 4 Å of waters around the backbone of the oligopeptide to keep computational time of the electronic properties calculations as low as possible.

**Electronic Couplings.** In biochemistry, many electron transfer reaction haves only weak electronic coupling between donors and acceptors. In this case, the ET rate can be described by Marcus theory by the following high-temperature nonadiabatic expression:[1]

$$k_{\mathrm{ET}} = \frac{2\pi}{\hbar} V_{\mathrm{DA}}^2 \frac{1}{\sqrt{4\pi\lambda k_{\mathrm{B}}T}} \exp\left(\frac{-(\lambda + \Delta G^{\circ})^2}{4\lambda k_{\mathrm{B}}T}\right) \qquad (1)$$

Here, $V_{\mathrm{DA}}$ is the electronic coupling between the diabatic donor and acceptor states, $\hbar$ is Planck's constant, $k_{\mathrm{B}}$ is Boltzmann's constant, $\lambda$ is the reorganization energy, $T$ is the temperature, and $\Delta G^{\circ}$ is the overall Gibbs free energy change of the electron transfer reaction. The coupling can be derived by applying the generalized Mulliken-Hush method (GMH)[29,30] using electronic properties of the adiabatic states of the system. How to apply the GMH scheme within the one-electron picture of hole transfer is considered in detail elsewhere.[47] For simple systems like the one to which we have applied a two-state model is a good approximation, whereas for more sophisticated systems the multistate model with bridge states has to be considered. Applying the two-state model, the bridge-mediated electronic coupling can be calculated through the following formula:

$$V_{\mathrm{DA}} = \frac{\Delta E_{12}|\mu_{12}|}{|\mu_{\mathrm{D}} - \mu_{\mathrm{A}}|} \qquad (2)$$

where $\Delta E_{12} = E_1 - E_2$ is the vertical excitation energy with $E_1$ and $E_2$ being the energies of the two relevant adiabatic states. $\mu_{12}$ is the transition dipole moment and $|\mu_{\mathrm{D}} - \mu_{\mathrm{A}}|$ is the difference of the diabatic dipole moments. Here, we can estimate $|\mu_{\mathrm{D}} - \mu_{\mathrm{A}}|$ as $ed_{\mathrm{DA}}$ or as $((\mu_1 - \mu_2)^2 + 4\,\mu_{12}{}^2)^{1/2}$.[29,30] Here, it is important to use only the projection of the transition and dipole moments onto the axis between the donor and acceptor rather than the length of the vectors. Furthermore, it is assumed that the two-state system is only weakly coupled, meaning $V_{\mathrm{DA}} < k_{\mathrm{B}}T$. Koopmans' theorem states that the energies of occupied molecular orbitals for a closed-shell system approximate the (negative) vertical ionization potentials ($-IP$).[32,33,48] On the basis of this theorem, the adiabatic splitting, $E_{12}$, is computed from the energy difference of the highest occupied molecular orbital (HOMO) and the next-highest occupied molecular orbital (HOMO-1) of the neutral system approximating the two quasi-degenerate electronic eigenstates of the ET system. We are aware of the approximate description of the hole transfer process in the systems considered. In particular, MD treatment of radical cation states of oligopeptides (explicit presence of a hole) can lead to structures that somewhat deviate from the generated structures within this study. However, we believe that averaging of the couplings over many thousand conformations make our estimates quite robust.

**Quantum Mechanical Calculations.** We carried out two different levels of theory to derive the electron transfer parameters, the semiempirical method INDO/S[38] and ab initio HF. The INDO/S calculations on the neutral oligopeptides are carried out on all 10 000 snapshots of every trajectory of Trp-(Pro)n-Trp with $n$ from 1 to 6. We further distinguish systems simulated in gas phase, systems simulated in explicit solvent but solvent molecules omitted from electronic property calculations (latterly denoted as *solvated-conformation-only*), and finally systems simulated in water keeping a layer of 4 Å of waters around backbone of oligopeptide (latterly denoted as *solvated*). Within the calculations, an average of 2.75% of the models had to be omitted due to the rare case of HOMO and HOMO-1 orbitals localizing into the same site and thus making the coupling meaningless in the sense of electron transfer between the two tryptophans. All ab initio Hartree−Fock calculations were performed with *Jaguar*.[49] We carried out single point energy calculations on all snapshots on solvated Trp-(Pro)3-Trp and Trp(Pro)6-Trp with and without the water molecules included as point charges in the quantum chemical calculations. We applied the 6-31G* basis set on all atoms resulting in 774 basis functions for Trp-(Pro)3-Trp and 1131 basis functions for Trp-(Pro)6-Trp.

**Statistical Analysis.** We performed all data analyses with the open source software package R.[50] For one of these, we applied the multivariate regression statistical method Partial Least Squares regression (PLS-R),[51] which is used to find the fundamental relations between two matrices X and Y. The method works by finding the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space, describing both X and Y by a few latent variables also known as principal components (PC). PLS-R models are usually built by extracting successive PCs, each one increasing the total percentage of Y variance explained given by cumulative Q2, until the predictive ability of the model is optimized. PLS-R is particularly suited when there is multicollinearity among the X values, where, by contrast, standard regression methods will fail. A further advantage is the ability to extract information about both the objects and the variables. The objects can be represented graphically according to the PC values, obtaining highly informative score plots in which similar objects appear as points closely situated in the space. In addition, the original variables can also be represented according to their overall contribution to the model by coefficient plots. These plots provide information about which variables exhibit a relevant association with the dependent variable, the direction of it, and if this association is statistically significant.

## III. Results and Discussion

**Molecular Dynamics.** The classical molecular dynamics trajectories of the oligopeptide reveal a substantial mobility of the peptide chains. Figure 2 shows the donor−acceptor distance $d_{\mathrm{DA}}$, measured between the middle of bond CD2-CE2 of the respective tryptophans, of all systems in gas phase as well as in solvent. We checked the reliability of the average values of the 10 ns trajectories by calculating $d_{\mathrm{DA}}$ from a 30 ns trajectory for Trp-(Pro)4-Trp in solvent. The resulting average for 30 ns (Pro4*) has no significant difference to 10 ns, indicating that the applied 10 ns trajectories are representative for the behavior of the systems.

There are several insights derived from this plot. First, mean $d_{\mathrm{DA}}$ of every oligopeptide system is smaller in gas phase than in solvent and second the variance of the distance is higher in gas phase than in solvent, indicated by the box size in Figure 2. Also, the minimum separation of the two tryptophans in gas phase is a constant of about 5 Å. This shows that the oligopeptides in gas phase undergo strong

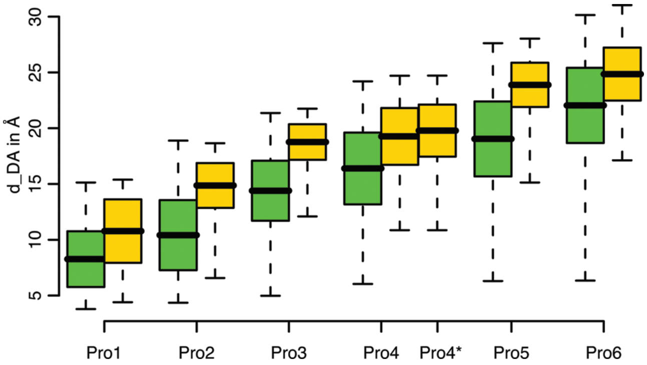


***Figure 2.*** Donor−acceptor distance $d_{DA}$ in Å for 10 ns trajectory of oligopeptides in gas phase (green) and solvent (yellow). $d_{DA}$ of Pro4 in solvent for 30 ns trajectory is indicated with an asterisk. Black bar shows mean value, box captures values between $-\sigma$ and $+\sigma$ and dashed lines indicate minimum and maximum values of respective distribution.

conformational changes and even collapse, bringing the two tryptophans in very close vicinity to each other during the MD. The oligopeptides simulated in solvent also undergo conformational changes but not as extreme and quick as those in gas phase, also they stay more craned and never collapse.

**Electronic Couplings.** We calculated the bridge-mediated electronic coupling between the donor and acceptor for every single snapshot as denoted above in eq 2 with parameters derived from INDO/S calculations on the three different trajectory types mentioned above: gas phase, solvated-conformation-only, and solvated. For direct comparison we also calculated coupling values in HF level of theory for solvated Trp-(Pro)n-Trp with $n = 3$ and 6 with and without the waters included as point charges. All rms $V_{DA}$ values are shown in Table 1. The data shows that there is no significant difference between INDO/S and HF indicating the high accuracy of the semiempirical INDO/S method in the calculation of electronic couplings. We also included coupling values calculated for the 30 ns solvated-conformation-only trajectory on Trp-(Pro)4-Trp, having a mean coupling value of $2.42 \times 10^{-5}$ eV (GMH). This value differs only marginally from the mean coupling value derived from the 10 ns trajectory, being $3.02 \times 10^{-5}$ eV, indicating that a sample size of 10 000 snapshots is sufficient for electronic coupling calculations of the applied oligoproline systems. Further cross checking also revealed excellent agreement between $V_{DA}$ computed by GMH and $V_{DA}$ computed with charge fragment difference method (FCM).[47] Figure 3 shows the rms $V_{DA}$ (GMH) plotted against mean $d_{DA}$ of the respective trajectories. It can be seen that there is an increasing gap between the coupling values calculated for the oligopeptides simulated in gas phase and those simulated in solvent. Figure 3 points out that these differences are not only based upon longer mean $d_{DA}$ within the solvated trajectories but also rely on further conformational effects of the solvent, which becomes larger with increasing $n$. There is no significant difference between rms $V_{DA}$ computed from solvated oligopeptides including or excluding the waters within INDO/S calculations, which points out that the electronic properties of the water molecules do not contribute to the coupling of the two tryptophans. Therefore, we will only focus on oligopeptides simulated in gas phase as well as in solvent, omitting the waters within the INDO/S calculations (solvated-conformation-only) in the following analyses.

**Gated/Nongated Mechanism.** It is known from the literature that electronic coupling can strongly depend on the conformation of the system.[52] In the so-called conformationally gated electron transfer, the charge is not gradually transferred from the donor to the acceptor instead there is a sudden jump enhanced by favorable conformations with high electronic coupling between donor and acceptor. Tracking the rate of conformational gating of the electron transfer in the different oligoproline systems, we plotted the distributions of the data by histograms of $V^2_{DA}$. The data is sorted and clustered by number of standard deviations above the total mean value of $V^2_{DA}$, called z-score. For each cluster, we then plotted the fraction of mean $V^2_{DA}$ given by its data points against its z-score. Values of coupling being higher than 100 times the standard deviation are accumulated into the last cluster. The results for gas-phase trajectories are given in Figure 4. Here, the plots show a trend of increasing conformational gating with increasing number of prolines. As already discussed above, with growing number of bridge prolines, the mean distance between donor and acceptor $d_{DA}$ becomes larger as well (shown in Figure 2). Thus, the coupling values associated with the conformations around the mean distance value decrease significantly. The dashed lines in Figure 2 also indicate the large mobility for the gas-phase systems. For any number of bridge prolines, we find few snapshots with very short $d_{DA}$, associated with high coupling values. The results indicate that the contribution of these few snapshots to the mean value increases with the number of prolines, establishing a gated mechanism for $n > 2$. A closer inspection at this $n = 3$ transition point indicates a critical mean distance around ∼15 Å. If a system with this mean value is capable of visiting conformations with $d_{DA}$ ∼7 Å, then the overall ET mechanism might be gated.

The results for the solvated trajectories are given in Figure 5. As expected, the overall results for the electron transfer in presence of water shows a nongated mechanism; only Pro2 lies on the border of conformational gating. This result can be rationalized in terms of the mean and extreme donor−acceptor distance observations deduced in the gas-phase studies. As seen in Figure 2, in the presence of water the extreme (lower) donor−acceptor distances are much higher than in gas phase for each trajectory. Only in the case of Pro2 we observe mean distances ∼15 Å together with low extreme $d_{DA}$ values close to 7 Å.

**Direct/Bridge-Mediated Coupling.** We also calculated the direct coupling between donor and acceptor, skipping the bridging prolines within the INDO/S calculations. From estimating $|\mu_D - \mu_A|$ through $ed_{DA}$ in eq 2, we know that direct coupling is inversely proportional to $d_{DA}$. Hence, we expect decreasing direct coupling values for oligopeptide systems with increasing $n$ due to increasing mean $d_{DA}$. Figure 6 shows both bridge-mediated coupling (solid lines) as well as direct coupling (dashed lines) from gas phase (black) and solvated-conformation-only (red) plotted against respective $d_{DA}$. In gas phase, direct couplings are essentially the same

***Table 1.*** Rms $V_{DA}$ Values in eV Calculated for Oligopeptide Systems with the INDO/S and HF Method

| | | gas phase | | solvated-conf.-only | | solvated | |
|---|---|---|---|---|---|---|---|
| prolines | method | GMH | FCM | GHM | FCM | GHM | FCM |
| 1 | INDO/S | $2.18 \times 10^{-2}$ | $2.17 \times 10^{-2}$ | $1.95 \times 10^{-2}$ | $1.94 \times 10^{-2}$ | $1.95 \times 10^{-2}$ | $1.94 \times 10^{-2}$ |
| 2 | INDO/S | $9.74 \times 10^{-3}$ | $9.68 \times 10^{-3}$ | $1.47 \times 10^{-3}$ | $1.46 \times 10^{-3}$ | $1.51 \times 10^{-3}$ | $1.46 \times 10^{-3}$ |
| 3 | INDO/S | $1.95 \times 10^{-3}$ | $1.94 \times 10^{-3}$ | $1.19 \times 10^{-4}$ | $1.19 \times 10^{-4}$ | $1.04 \times 10^{-4}$ | $1.19 \times 10^{-4}$ |
| | HF | | | $1.61 \times 10^{-4}$ | $1.61 \times 10^{-4}$ | $1.49 \times 10^{-4}$ | $1.51 \times 10^{-4}$ |
| 4 | INDO/S | $1.16 \times 10^{-3}$ | $1.16 \times 10^{-3}$ | $3.02 \times 10^{-5}$ | $3.03 \times 10^{-5}$ | $2.78 \times 10^{-5}$ | $2.79 \times 10^{-5}$ |
| 4 (30 ns) | INDO/S | | | $2.42 \times 10^{-5}$ | $2.44 \times 10^{-5}$ | | |
| 5 | INDO/S | $3.79 \times 10^{-4}$ | $3.79 \times 10^{-4}$ | $4.36 \times 10^{-8}$ | $4.37 \times 10^{-8}$ | $4.39 \times 10^{-8}$ | |
| 6 | INDO/S | $1.25 \times 10^{-4}$ | $1.25 \times 10^{-4}$ | $7.36 \times 10^{-8}$ | $7.37 \times 10^{-8}$ | $6.58 \times 10^{-8}$ | |
| | HF | | | $2.48 \times 10^{-7}$ | $2.32 \times 10^{-7}$ | $7.25 \times 10^{-8}$ | |

as bridge-mediated couplings, which can be explained by conformational gating. As described above, in gas phase only conformations of very short $d_{DA}$ contribute to the mean coupling $V_{DA}$. Here, direct couplings between donor and acceptor are also high, consequently summing up to equal mean $V_{DA}$. Direct coupling values calculated on solvent trajectories show expected behavior, being much lower than bridge-mediated coupling values. Here, in contrast to the gas phase, mean $d_{DA}$ and also lower extreme $d_{DA}$ are increasing with $n$ and thus direct couplings drop significantly because of their inverse proportionality to distance. Equal direct and bridge-mediated couplings of trajectories with $n = 1$ or 2 are based on few snapshots of low $d_{DA}$ having also high direct coupling between donor and acceptor hence contributing most of mean $d_{DA}$.

**Conformational Analysis.** We extracted additional conformational parameters from the trajectories for further analysis on their influence on the coupling. Therefore, we split the oligopeptides into separate groups for every tryptophan and $\pi$-system within the oligopeptide chain. For the tryptophans, we take the middle of bond CD2-CE2 as center M of the plane given by the aromatic ring system. For the $\pi$-systems, we take atom C as the center of the plane given by the carboxyl group and its neighbored atoms. Between two planes, we define the distance $d$ as the Euclidean distance between their centers, the angle $p$ as the inner angle between the two normal vectors tracking the planarity of the two planes, and finally position angle $r$ as the minimum of the angle between normal vector of the first plane and center of the second plane and vice versa to distinguish planes lying on top of each other from planes being next to each other. Figure 7 depicts $d$, $p$, and $r$ between the two tryptophans in Trp-(Pro)2-Trp. We extracted $d$, $p$, and $r$ between all adjacent groups as well as between both tryptophans. Here, donor is denoted by D, the acceptor is denoted by A, and the $\pi$-systems are counted in ascending order starting at 1 being next to the donor.

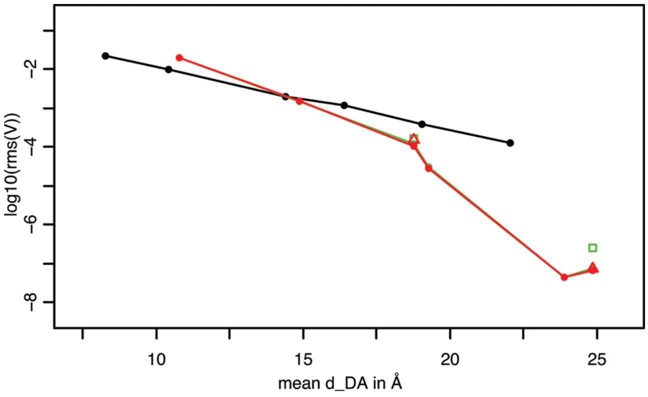


***Figure 3.*** Rms $V_{DA}$ (GMH) values plotted against mean donor−acceptor distance $d_{DA}$ in trajectories of type gas phase (black), solvated-conformation-only (red) and solvated (green). Rms $V_{DA}$ from HF calculations are given for Trp-(Pro)3-Trp and Trp-(Pro)6-Trp for solvated-conformation-only (red triangle) and solvated (green square) trajectories.

We applied PLS-R on all snapshots in gas phase as well as solvated-conformation-only trajectories to analyze the impact of the different conformational parameters on the coupling. Data matrix X is given by derived conformational parameters $d$, $p$, and $r$ and Y is $V^2_{DA}$ for each respective trajectory. For normalization, we applied UV scaling on X, assuming a normal distribution, and $\log_{10}$ on Y due to its large range. In general, the PLS-R models for the oligopeptide systems in gas phase have good quality with cumulative Q2 values ranging from 0.34 for Trp-(Pro)6-Trp to 0.53 for Trp-(Pro)2-Trp. The cumulative Q2 of the PLS-R models for oligopeptides in solvated-conformation-only trajectory are lower, ranging from only 0.05 for Trp-(Pro5)-Trp to 0.51 for Trp-(Pro)1-Trp. All models have a maximum of only three principle components (PCs).

The analysis of PLS-R results is based on outcomes of all models from which we show resulting score and loading plots of Trp-(Pro)3-Trp in gas phase (figure 8) as well as Trp-(Pro)3-Trp in solvated-conformation-only (figure 9) as examples. All other PLS-R results can be found in the Supporting Information. Within the score plots, green indicates snapshots with low, yellow medium, and red high coupling values. In all PCs of all models in gas phase, the donor−acceptor distance $d_{DA}$ clearly has the highest loading coefficient, indicated by red bar in Figure 8, meaning that it has the highest influence on the coupling. Results of PLS-R analyses on oligopeptide systems modeled in solvent are not as predictive as those from models in gas phase, indicated by their lower cumulative Q2 values. Overall, they show different behavior from those modeled in gas phase. Within the solvated models, the parameter of highest influence on the coupling $V^2_{DA}$ is not only $d_{DA}$ but, with equally high impact, also $d_{D\text{-}\pi 1}$ and $r_{D\text{-}\pi 1}$ (red bars in Figure 9). Additionally, their loading stays high in PC2 and PC3. Planarity angles $p$ and distances $d$ between the bridging $\pi$-systems do not show any influence on the coupling values within the solvent models.

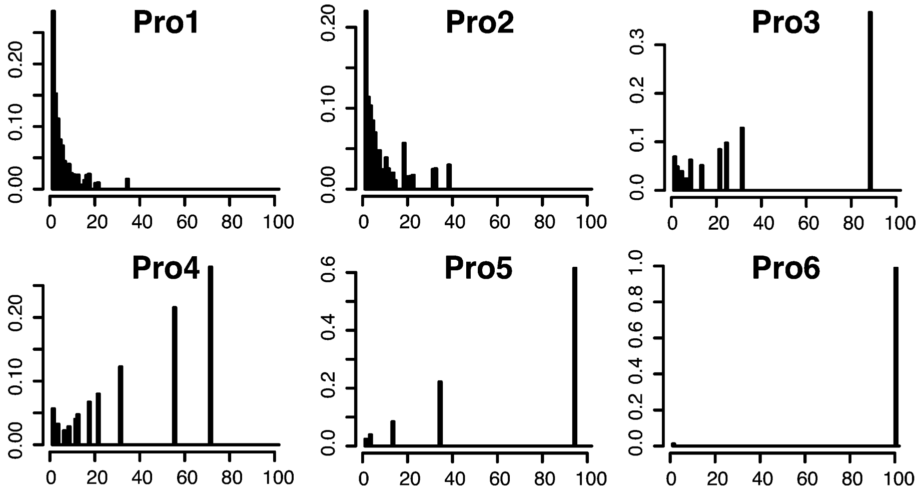


*Figure 4.* Fraction of mean $V^2_{DA}$ given by data cluster *i* (*y* axes) against z-score of *i* on $V^2_{DA}$ (*x* axes) for trajectories in gas phase. Coupling values higher than 100 z-scores are accumulated into the last cluster.

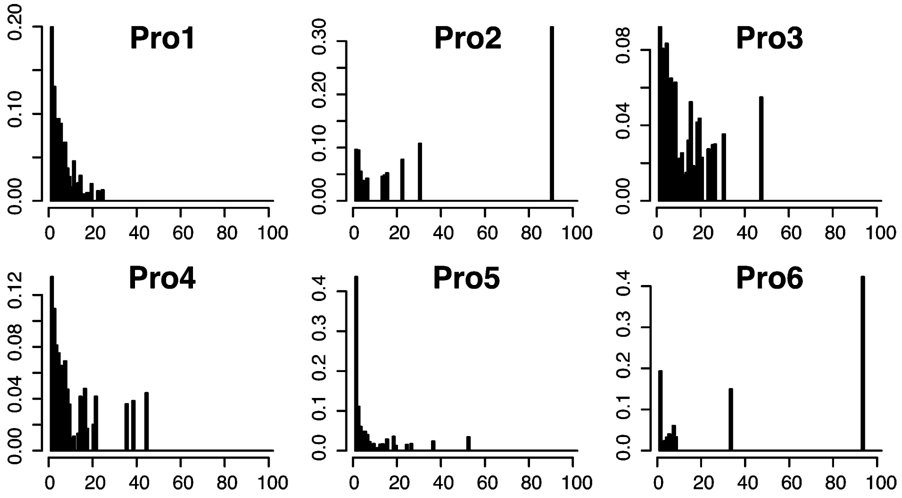


*Figure 5.* Fraction of mean $V^2_{DA}$ given by data cluster *i* (*y* axes) against z-score of *i* on $V^2_{DA}$ (*x* axes) for solvated-conformation-only trajectories. Coupling values higher than 100 z-scores are accumulated into last cluster.

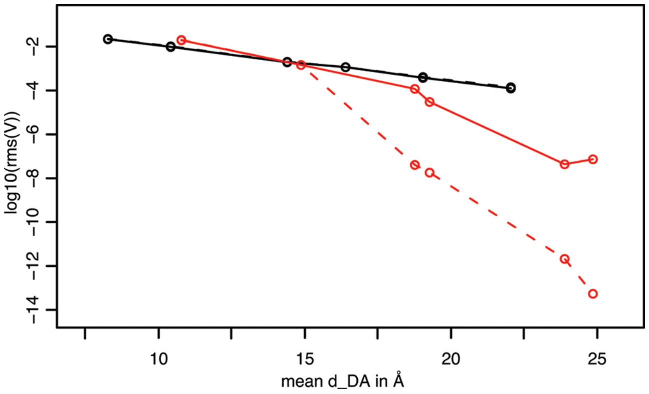


*Figure 6.* Rms $V_{DA}$ values plotted against mean donor−acceptor distance $d_{DA}$ in gas phase (black) and solvated-conformation-only (red) trajectories. Solid lines indicate bridge mediated couplings and dashed lines only direct couplings between donor and acceptor.

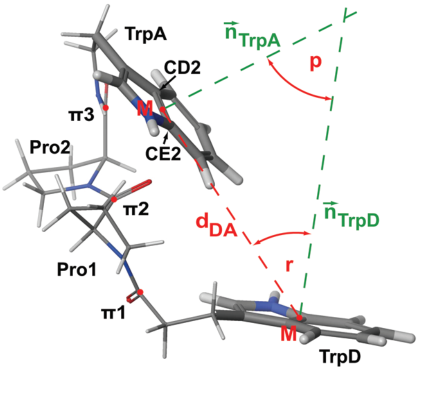


*Figure 7.* Scheme of conformational parameters distance *d* and angles *p* and *r*, derived from 3D oligopeptide structure. See text for more details.

Translating the influence of the previously detected conformational parameters on the coupling into concrete values derived from the actual snapshots, we calculated the mean values of these parameters extracted from two groups. The first group is snapshots having high coupling, meaning $V^2(x) >$ mean $V^2_{DA}$, and the second group is the 1000 snapshots lowest in coupling. Figure 10 gives the box plot of significant conformational parameters derived from both groups (high and low coupling) for gas phase as well as solvated-conformation-only trajectories.

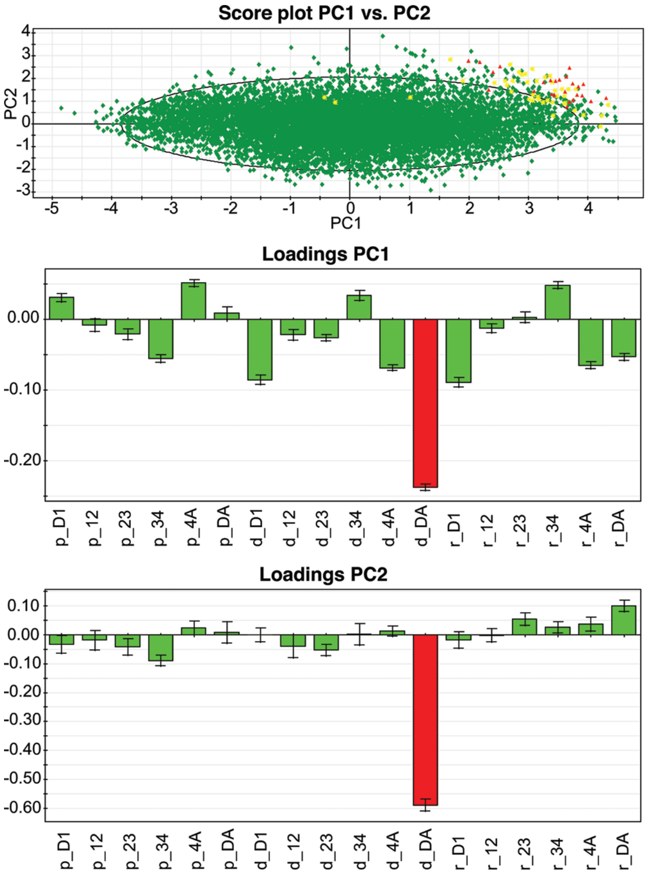


***Figure 8.*** Score and loading plots of PLS-R analysis on Trp-(Pro)3-Trp in gas phase trajectory.

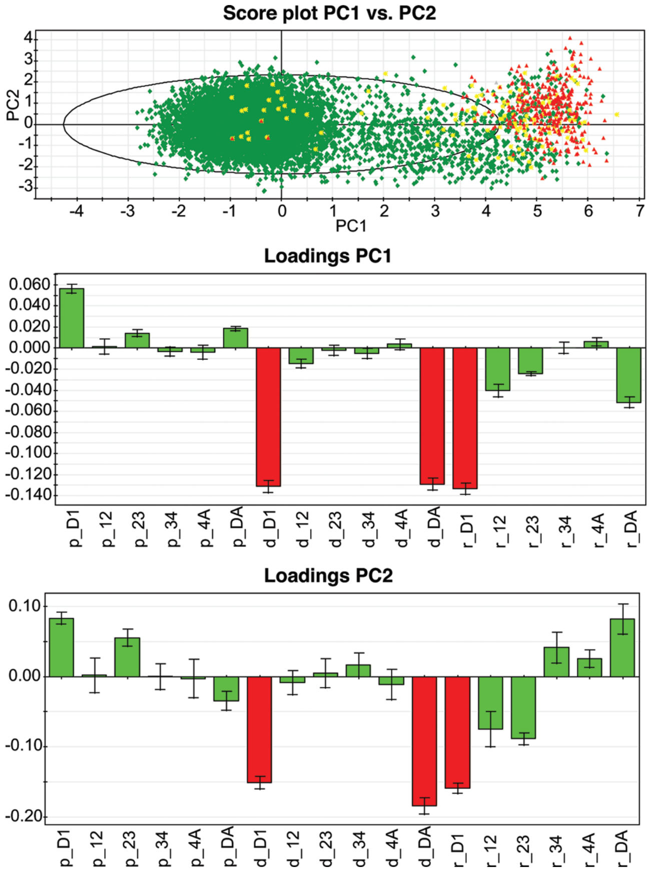


***Figure 9.*** Score and loading plots of PLS-R analysis on Trp-(Pro)3-Trp in solvated-conformation-only trajectory.

In general, there is a clear correlation of shorter distances with high coupling in systems derived from gas-phase simulations as well as in solvated systems. This effect is very strong for the donor−acceptor distance $d_{DA}$ in gas phase. Here, the average values are 5.99 Å for high coupling against 17.48 Å for low coupling. This trend is less clear but still discriminative for the solvated systems where the average $d_{DA}$ for high coupling is 16.43 Å and the average $d_{DA}$ for low coupling is 19.60 Å. The strong loading coefficient of stacking angles $r_{D\text{-}\pi 1}$ in the PLS-R analysis becomes clear when we examine the average values of the two groups. At this point, higher coupling correlates with a smaller angle.

**Rate Comparison with Experiments.** Isied et al., working on oligoproline peptides with Ruthenium based donors and acceptors, showed that the electron transfer mechanism changes from a predominantly electron superexchange to a predominantly electron hopping when $d_{DA}$ exceeds about 20 Å.[13] Using the parameters from Isied et al. for the reorganization energy, we have computed the ET rates, shown in Figure 11. Analyzing the solvated-conformation-only (red) and solvated (green) plots it seems clear that the linear behavior is broken at $d_{DA}$ = ~20 Å. At this point, the superexchange ET rate is drastically reduced. These results confirm a possible change in mechanism from superexchange to electron hopping. Future work will address the electron hopping mechanism to confirm this point as the presented study exclusively investigates the superexchange electron transfer mechanism.

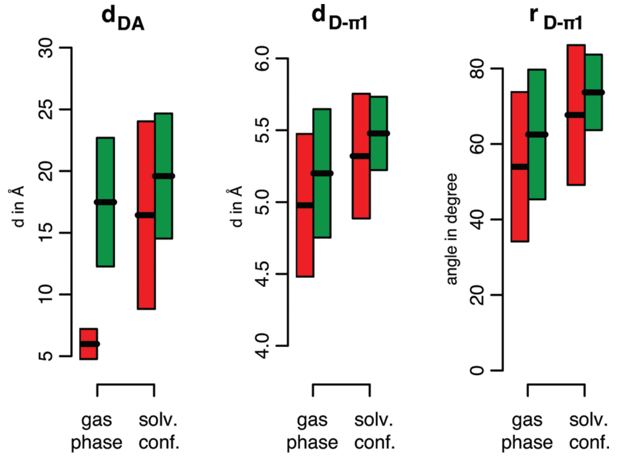


***Figure 10.*** Box plot of conformational parameters extracted from high and low coupling groups of gas-phase and solvated-conformation-only trajectories. Distances *d* measured in Å, angle *r* in degrees. Color code: red = high coupling and green = low coupling.

## IV. Conclusions

We have produced a comprehensive study of hole electron transfer in oligoproline peptides of variable length with tryptophan-based donors and acceptors. We have performed extensive molecular dynamics studies and computed the electronic coupling by means of INDO/S semiempirical method and ab initio Hartree−Fock methods (for $n$ = 3 and 6). The HF calculations sum up to 40 000 single-point

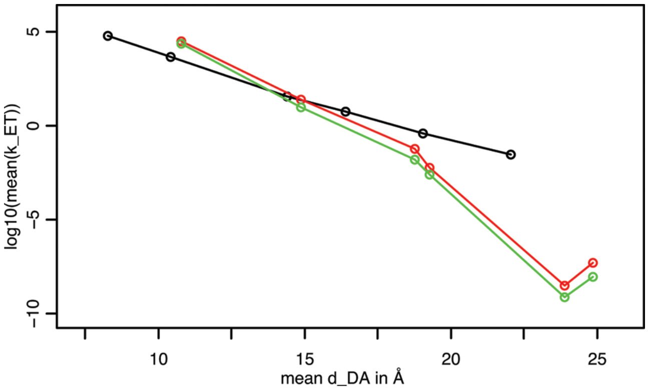


***Figure 11.*** Mean rate constant $k_{ET}$ plotted versus average donor−acceptor distance $d_{DA}$ of oligopeptide in gas-phase (black), solvated-conformation-only (red), and solvated (green) trajectories. Reorganization energy given by ref 13 through $\lambda = \lambda_{in} + \lambda_{out}$, where $\lambda_{in} = 0.085$ eV, $\lambda_{out} = 7.91(1/2a_1 + 1/2a_2 - 1/d_{DA})$ eV and $a_1 = a_2 = 3.3$ Å.

calculations from which we found good agreement between results of semiempirical INDO/S and ab initio HF in the oligopeptides we have investigated. In gas phase, all oligopeptide systems undergo strong conformational changes bringing the two tryptophans in close vicinity, which results in a gated electron transfer mechanism. The dynamics of the oligopeptides in water do not allow for such a close proximity between the donor and acceptor as the number of bridge tryptophans increases, establishing a nongated mechanism where the bridging prolines play a major role in mediating the electronic coupling.

The agreement between the solvated-conformation-only and the solvated results indicate that the water effects are mainly in restricting the conformational space rather than in electronic effects. Whereas the excitation energy $E_{12}$ differs when explicitly adding the water point charges in the one electron Hamiltonian, there is not a large effect of $E_{12}$ on the electronic coupling, as only (diabiatic) orbital overlap accounts. Finally, in agreement with experimental data, our results point to a $d_{DA}$ ~20 Å as a possible point for superexchange to hopping mechanism transition.

**Acknowledgment.** This work was supported by a grant from the Spanish Ministry of Education and Science to V.G. through the project CTQ2007-62122. Computational resources were provided by the Barcelona Supercomputing Center.

**Supporting Information Available:** Autocorrelation plot of donor−acceptor distance in 10 ps trajectory of Trp-(Pro)2-Trp in gas phase with 1 ps time steps as well as score and loading plots of all PLS-R analyses on conformational parameters. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Marcus, R. A.; Sutin, N. *Biochim. Biophys. Acta* **1985**, *811*, 265–322.

(2) Newton, M. D. *Chem. Rev.* **1991**, *91*, 767–792.

(3) Beratan, D. N.; Onuchic, J. N.; Winkler, J. R.; Gray, H. B. *Science* **1992**, *258*, 1740–1741.

(4) Gray, H. B.; Winkler, J. R. *Annu. Rev. Biochem.* **1996**, *65*, 537–61.

(5) Balzani, V.; Piotrowiak, P.; Rodgers, M. A. J.; Mattay, J.; Astruc, D.; Gray, H. B.; Fukuzumi, S.; Mallouk, T. E.; Haas, Y.; de Silva, A. P.; Gould, I. R. *Electron Transfer in Chemistry*; Wiley-VCH: Weinheim, Germany, 2001; Vol. I-V.

(6) Ungar, L. W.; Newton, M. D.; Voth, G. A. *J. Phys. Chem. B* **1999**, *103*, 7367–7382.

(7) Isied, S. S.; Ogawa, M. Y.; Wishart, J. F. *Chem. Rev.* **1992**, *92*, 381–394.

(8) Bixon, M.; Giese, B.; Wessely, S.; Langenbacher, T.; Michel-Beyerle, M. E.; Jortner, J. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 11713–11716.

(9) Ogawa, M. Y.; Wishart, J. F.; Young, Z.; Miller, J. R.; Isied, S. S. *J. Phys. Chem.* **1993**, *97*, 11456–11463.

(10) Felts, A. K.; Pollard, W. T.; Friesner, R. A. *J. Phys. Chem.* **1995**, *99*, 2929–2940.

(11) Petrov, E. G.; May, V. *J. Phys. Chem. A* **2001**, *105*, 10176–10186.

(12) Berlin, Y. A.; Ratner, M. A. *Radiat. Phys. Chem.* **2005**, *74*, 124–131.

(13) Malak, R. A.; Gao, Z.; Wishart, J. F.; Isied, S. S. *J. Am. Chem. Soc.* **2004**, *126*, 13888–13889.

(14) Shih, C.; Museth, A. K.; Abrahamsson, M.; Blanco-Rodriguez, A. M.; Di Bilio, A. J.; Sudhamsu, J.; Crane, B. R.; Ronayne, K. L.; Towrie, M.; Vlcek, A.; Richards, J. H.; Winkler, J. R.; Gray, H. B. *Science* **2008**, *320*, 1760–1762.

(15) Wittekindt, C.; Schwarz, M.; Friedrich, T.; Koslowski, T. *J. Am. Chem. Soc.* **2009**, *131*, 8134–8140.

(16) Wolfgang, J.; Risser, S. M.; Priyadarshy, S.; Beratan, D. N. *J. Phys. Chem. B* **1997**, *101*, 2986–2991.

(17) Miller, N. E.; Wander, M. C.; Cave, R. J. *J. Phys. Chem. A* **1999**, *103*, 1084–1093.

(18) Castner, E. W.; Kennedy, D.; Cave, R. J. *J. Phys. Chem. A* **2000**, *104*, 2869–2885.

(19) Xie, Q.; Archontis, G.; Skourtis, S. S. *Chem. Phys. Lett.* **1999**, *312*, 237–246.

(20) Skourtis, S. S.; Archontis, G.; Xie, Q. *J. Chem. Phys.* **2001**, *115*, 9444–9462.

(21) Balabin, I. A.; Onuchic, J. *Science* **2000**, *290*, 114–117.

(22) Newton, M. D. *Int. J. Quantum Chem.* **2000**, *77*, 255–263.

(23) Kawatsu, T.; Kakitani, T.; Yamato, T. *J. Phys. Chem. B* **2002**, *106*, 11356–11366.

(24) Balabin, I. A.; Beratan, D. N.; Skourtis, S. S. *Phys. Rev. Lett.* **2008**, *101*, 158102–4.

(25) Skourtis, S. S.; Balabin, I. A.; Kawatsu, T.; Beratan, D. N. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 3552–3557.

(26) Zimmt, M. B.; Waldeck, D. H. *J. Phys. Chem. A* **2003**, *107*, 3580–3597.

(27) Troisi, A.; Ratner, M. A.; Zimmt, M. B. *J. Am. Chem. Soc.* **2004**, *126*, 2215–2224.

(28) Lu, S.-Z.; Li, X.-Y.; Liu, W. *Chem. Phys. Lett.* **2005**, *414*, 71–75.

(29) Cave, R. J.; Newton, M. D. *Chem. Phys. Lett.* **1996**, *249*, 15–19.

(30) Cave, R. J.; Newton, M. D. *J. Chem. Phys.* **1997**, *106*, 9213–9226.

(31) Shin, Y.-g. K.; Newton, M. D.; Isied, S. S. *J. Am. Chem. Soc.* **2003**, *125*, 3722–3732.

(32) Liang, C.; Newton, M. D. *J. Phys. Chem.* **1992**, *96*, 2855–2866.

(33) Onuchic, J. N.; Beratan, D. N.; Hopfield, J. J. *J. Phys. Chem.* **1986**, *90*, 3707–3721.

(34) Blancafort, L.; Voityuk, A. A. *J. Phys. Chem. A* **2006**, *110*, 6426–6432.

(35) Zhang, L. Y.; Friesner, R. A. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 13603–13605.

(36) Balabin, I. A.; Onuchic, J. N. *J. Phys. Chem.* **1996**, *100*, 11573–11580.

(37) Gehlen, J. N.; Daizadeh, I.; Stuchebrukhov, A. A.; Marcus, R. A. *Inorg. Chim. Acta* **1996**, *243*, 271–282.

(38) Ridley, J.; Zerner, M. *Theor. Chem. Acc.* **1973**, *32*, 111–134.

(39) Voityuk, A. *Chem. Phys. Lett.* **2006**, *427*, 177–180.

(40) Prytkova, T. R.; Kurnikov, I. V.; Beratan, D. N. *J. Phys. Chem. B* **2005**, *109*, 1618–1625.

(41) Lambert, C.; Amthor, S.; Schelter, J. *J. Phys. Chem. A* **2004**, *108*, 6474–6486.

(42) *Maestro, version 8.5*; Schrödinger, LCC: New York, NY, 2008.

(43) *Impact, version 5.0*; Schrödinger, LCC: New York, NY, 2008.

(44) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Hermans, J. *Intermolecular Forces*, First ed.; Reidel, Dordrecht: 1981; Vol. 331.

(45) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*; ACM: Tampa, FL, 2006, 84.

(46) Jorgensen, W. L.; Tirado-Rives, J. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6665–6670.

(47) Voityuk, A. A.; Rosch, N. *J. Chem. Phys.* **2002**, *117*, 5607–5616.

(48) Jordan, K. D.; Paddon-Row, M. N. *J. Phys. Chem.* **1992**, *96*, 1188–1196.

(49) *Jaguar, version 7.5*; Schrödinger, LCC: New York, NY, 2008.

(50) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing: Vienna, Austria, 2007.

(51) Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.

(52) Hoffman, B. M.; Ratner, M. A. *J. Am. Chem. Soc.* **1987**, *109*, 6237–6243.

CT900377J

JCTC Journal of Chemical Theory and Computation

# Discrete Optimization of Electronic Hyperpolarizabilities in a Chemical Subspace

B. Christopher Rinderspacher,*[†,‡] Jan Andzelm,[†] Adam Rawlett,[†] Joseph Dougherty,[†]
David N. Beratan,[‡] and Weitao Yang[‡]

*Army Research Laboratory, Aberdeen Proving Ground, Aberdeen, Maryland 21005,
and Department of Chemistry, Duke University, 124 Science Dr,
Durham, North Carolina 27708*

Received June 26, 2009

**Abstract:** We introduce a general optimization algorithm based on an interpolation of property values on a hypercube. Each vertex of the hypercube represents a molecule, while the interior of the interpolation represents a virtual superposition ("alchemical" mutation) of molecules. The resultant algorithm is similar to branch-and-bound/tree-search methods. We apply the algorithm to the optimization of the first electronic hyperpolarizability for several tolane libraries. The search includes structural and conformational information. Geometries were optimized using the AM1 Hamiltonian, and first hyperpolarizabilities were computed using the INDO/S method. Even for small libraries, a significant improvement of the hyperpolarizability, up to a factor of ca. 4, was achieved. The algorithm was validated for efficiency and reproduced known experimental results. The algorithm converges to a local optimum at a computational cost on the order of the logarithm of the library size, making large libraries accessible. For larger libraries, the improvement was accomplished by performing electronic structure calculations on less than 0.01% of the compounds in the larger libraries. Alternation of electron donating and accepting groups in the tolane scaffold was found to produce candidates with large hyperpolarizabilities consistently.

## 1. Introduction

In recent years, organic molecules have garnered increasing attention as components of high-hyperpolarizability materials, partly due to the variety of synthetically accessible compounds, cost, and ease of processing.[1,2] Applications for materials with high hyperpolarizabilities are found in telecommunication and optics.[3] The dominant nonlinear response of organic molecules often finds its origin in the conjugated $\pi$-system, which facilitates the electronic polarizability. The design of such molecules *in silico* is complicated by the fact that chemical space, even constrained to smaller organic compounds, is combinatorially complex. The number of organic molecules of medium size is estimated[4] to be on the order of $10^{200}$. Enumeration is therefore unfeasibly costly, and other methods for property optimization need to be developed. Including conformational searching further complicates molecular design.

Methods for optimization in discrete spaces have been studied extensively and recently reviewed.[5] Optimization methods include integer programming, as in branch-and-bound techniques (including dead-end elimination[6]), simulated annealing,[7] and genetic algorithms.[8] These algorithms have found renewed interest and application in molecular and materials design.[9–12] Recently, new approaches have been explored to embed discrete chemical space in continuous spaces to take advantage of continuous optimization techniques. These include, in particular, activities in our group on the linear combination of atomic potentials (LCAP)[13–15] method and the approach of von Lilienfeld,[16–19] using a grand-canonical ensemble strategy. Here, we further employ continuous optimization methods aimed at discovering structures with optimal properties.

* Corresponding author e-mail: berend.rinderspacher@arl.army.mil.

[†] Army Research Laboratory.

[‡] Duke University.

10.1021/ct900325p CCC: $40.75 © 2009 American Chemical Society
Published on Web 11/04/2009

The problem of discrete optimization in chemical space can be tackled by embedding the discrete space in a virtual continuous space, parametrized by a set of continuous variables. This strategy establishes a continuous path from one molecule to another. Such a space can be constructed by defining molecules as a succession of gradual replacements of an atom or molecular fragment by another. These fragment or atom placements may be arbitrary, but the satisfaction of valency rules may be desirable. For example, a hydrogen in $CH_4$ might be replaced by a halogen or a methyl group, each corresponding to a specific geometry (or ensemble of geometries), energy(ies), and property value(s). It is possible to construct a continuous transition between Hamiltonians for the chemical structures as was done for LCAP.[13] Equation 1 illustrates the procedure.

$$H(\boldsymbol{\lambda}) = \sum_i \lambda_i H_i, \qquad \sum_i \lambda_i = 1, \qquad 0 \le \lambda_i \le 1 \tag{1}$$

Each Hamiltonian $H_i$ acts only on its own molecular subspace $\Omega_i$, and $H$ acts on the direct sum of these spaces $\oplus_i \Omega_i$. In eq 1, the summation constraint implies the mutual exclusivity of the groups in the library (e.g., in the above example, as the hydrogen component increases toward 1, the halogen component decreases toward 0). In this approach, the groups are still linked through the wave function. Therefore, it is possible that all optima are at nonphysical configurations (e.g., half hydrogen and half halogen in the same location). Starting with allowed values of $\lambda_i$ ($0 \le \lambda_i \le 1$), it is possible to compute the numerical derivative of a property $P$. We now explore the application of this idea for discrete optimization of the first hyperpolarizability using differences of property values to replace continuous gradients.

## 2. Methods

**Linear Interpolation of Discrete Spaces.** Analogous to LCAP optimization, any property can in principle be interpolated in a virtual continuous space. We call the interpolated space "virtual" since noninteger $\lambda_i$-values correspond to intermediate or "alchemical" species. In general, given a library with $N$ molecules with property values $P_s$ for molecule $s$, $\lceil \log_2 N \rceil$ (the smallest integer larger than $\log_2 N$) variables may be used to embed the discrete library in the continuous space. In LCAP, intermediate species contain contributions of each subspecies as well as cross-terms that arise from coupling *via* the wave function, which is one source of "virtual" optima. The values for intermediate species in this scheme are not contaminated by these cross-terms and depend only on the values at the real molecules. For example, assume a library consisting of methane, ethane, propane, and butane in exactly that order (see Figure 1). It is possible to interpolate among the 4 molecules using the parameters $\lambda_0$ and $\lambda_1$. A (quadratic) polynomial interpolating the ground state energies (for example) is the following:

$$E(\lambda_0, \lambda_1) = E_0(1-\lambda_0)(1-\lambda_1) + E_1\lambda_0(1-\lambda_1) + E_2(1-\lambda_0)\lambda_1 + E_3\lambda_0\lambda_1 \tag{2}$$

| Molecule | s | $\lambda_1\lambda_0$ |
|---|---|---|
| $CH_4$ | 0 | 0,0 |
| $C_2H_6$ | 1 | 0,1 |
| $C_3H_8$ | 2 | 1,0 |
| $C_4H_{10}$ | 3 | 1,1 |

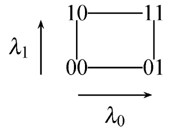


***Figure 1.*** Simple example for interpolation. The bits $\lambda_1\lambda_0$ represent the molecule number $s = 2\lambda_1 + \lambda_0$ in the binary system.

This energy equation has a well-defined minimum when constrained to the square domain. Due to the domain constraints (only values on the square are allowed), the components of the gradient at the vertices pointing outside the square have to be dropped in pure gradient methods. At the minimum the gradient $g$ points outside the square of definition for all components and thus is zero in the function's domain. Similarly, the Hessian $J$ decomposes into normals, which point into the square or away. Again the constraints mandate that components of $\Delta x = -J^{-1}g$ pointing out of the square are dropped for methods depending on $\Delta x$. Interpolation using a single variable for this set of compounds would produce a third degree polynomial, but homogeneous solutions to third order polynomials are not trivial, and the optimum is not guaranteed to correspond to a true molecule, i.e., $\lambda \in \{0, 1, 2, 3\}$.

The preceding example highlights the dependence of the property polynomial on the ordering of the molecules. Generalization of the example to a library $\mathscr{L}$ of size $N$ leads to eqs 3 and 4. Equation 3 describes the bit-string (binary) representation of a number $s$ with bit $s(i)$ at the $i$th position.

$$s = \sum_{i=0}^{\lceil \log_2 N \rceil} s(i) \times 2^i, \quad s(i) \in \{0, 1\} \tag{3}$$

$$\tilde{P}(\boldsymbol{\lambda}) = \sum_{s=0}^{N-1} P_s \prod_{i=0}^{\lceil \log_2 N-1 \rceil} ((1-\lambda_i)^{s(i)} \lambda_i^{1-s(i)}) \tag{4}$$

Equation 4 defines the property interpolation $\tilde{P}$ based on the bit-strings. We differentiate between the interpolation function $\tilde{P}$ and the set of discrete property values $P_s$ to emphasize the domain of definition. The former is defined on the "virtual", continuous hypercube ($[0, 1]^{\lceil \log_2 N \rceil}$), while the latter is defined on the discrete space $\mathscr{L}$. This polynomial is continuous on the hypercube and has order $\lceil \log_2 N \rceil$ and $\lceil \log_2 N \rceil$ variables.

**Derivatives of $\tilde{P}$.** In order to use conventional optimization algorithms on continuous spaces, it is necessary to find the derivatives of $\tilde{P}$.

$$\frac{\partial \tilde{P}}{\partial \lambda_j}(\boldsymbol{\lambda}) = \sum_{s=0}^{N-1} P_s (-1)^{s(j)} \prod_{b \neq j}^{\lceil \log_2 N \rceil} ((1-\lambda_b)^{s(b)} \lambda_b^{1-s(b)}) \tag{5}$$

$$\frac{\partial^2 \tilde{P}}{\partial \lambda_k \partial \lambda_l}(\boldsymbol{\lambda}) = \sum_{s=0}^{N-1} P_s (-1)^{s(k)+s(l)} \prod_{b \notin \{k,l\}}^{\lceil \log_2 N \rceil} ((1-\lambda_b)^{s(b)} \lambda_b^{1-s(b)}) \tag{6}$$

Equations 5 and 6 show first and second order analytical derivatives of $\tilde{P}$. The derivative of $\tilde{P}$ at $\boldsymbol{\lambda}$ corresponding to the molecule with number $s$ in the library $\mathscr{L}$ can be computed from nearest bit-string neighbors (see eqs 7−10). $s^{(j)}$ denotes the neighbor which differs only by the $j$th bit, while $s^{(k,l)}$ identifies the neighbor which differs only in the $k$th and $l$th bits.

$$s^{(j)} = s + (-1)^{s(j)} \times 2^j \qquad (7)$$

$$s^{(k,l)} = s + (-1)^{s(k)} \times 2^k + (-1)^{s(l)} \times 2^l, \qquad k \neq l \qquad (8)$$

$$\lambda_i = s(i), \qquad \frac{\partial \tilde{P}}{\partial \lambda_j}(\boldsymbol{\lambda}) = (-1)^{s(j)}(P_s - P_{s^{(j)}}) \qquad (9)$$

$$\frac{\partial^2 \tilde{P}}{\partial \lambda_k \partial \lambda_l}(\boldsymbol{\lambda}) = (-1)^{s(k)}(-1)^{s(l)}(P_s - P_{s^{(k)}} - P_{s^{(l)}} + P_{s^{(k,l)}}), \qquad l \neq k, \quad \lambda_i = s(i) \qquad (10)$$

The highly nonlinear, but continuous, function $\tilde{P}$ allows the development of optimization methods by substituting derivatives by finite differences in continuous optimization methods. In this case, the analytical property derivatives for a molecule (i.e., at the vertices of the hypercube, where $\lambda_i = s(i)$ for vertex $s$) are simple (finite) property value differences, unlike in LCAP. The derivatives of LCAP need not be on straight lines pointing from one physical (non-"alchemical") molecule to another, although the property values of each real molecule are the same for either optimization scheme. Formally, $\tilde{P}$ is very similar to the Bayesian clustering approach, but no stochastic interpretation is needed in this case.[30] This framework also unifies some previous approaches.[15,20] Balamurugan et al.[20] have applied a best-first approach (BFA) to chemical optimization, which chooses the first substituent at a substitution site that improves the property. This method resembles the optimization algorithm employed in the latter sections in that the property improves at every step, but BFA uses the property value instead of the derivative. Keinan et al.[15] have used an algorithm which represents the steepest-descent method applied to $\tilde{P}$, as well as a line-search in which the direction of largest change is exclusively used. Unlike the other algorithms described, the steepest-descent method potentially jumps through the hypercube. While the following algorithm and Keinan's line-search both traverse the edges of the hypercube constantly improving the property value, Keinan computes all single substitutions at every step.

**Comparison with Dead-End Elimination.** To compare our approach (eq 4) with dead-end-elimination algorithms (DEE), we consider the minimization of a pairwise additive property function comprised of single-parameter contributions $P_i^{(\mu)}$ acting on site $i$ with occupation $\mu$ and double-parameter contributions $P_{ij}^{(\mu,\nu)}$ acting on sites $i$ and $j$ with occupation $\mu$ and $\nu$ (eq 11).

$$P_s = \sum_i P_i^{(s(i))} + \sum_{i<j} P_{ij}^{(s(i),s(j))} \qquad (11)$$

$$\begin{aligned}\tilde{P}(\boldsymbol{\lambda}) = \sum_i (P_i^{(0)}\lambda_i + P_i^{(1)}(1-\lambda_i)) + \\ \sum_{i<j} (P_{ij}^{(0,0)}\lambda_i\lambda_j + P_{ij}^{(1,0)}(1-\lambda_i)\lambda_j + \\ P_{ij}^{(0,1)}\lambda_i(1-\lambda_j) + P_{ij}^{(1,1)}(1-\lambda_i)(1-\lambda_j))\end{aligned} \qquad (12)$$

Collecting all terms, we find a quadratic dependence of $\tilde{P}$ on the pairwise terms $P_{ij}$ with the parameters $\lambda_i$ (eq 12). Consequently, the derivatives are linear with respect to $\lambda_i$ (eq 13).

$$\frac{\partial \tilde{P}}{\partial \lambda_i} = P_i^{(0)} - P_i^{(1)} + \sum_{j \neq i} ([P_{ij}^{(0,0)} - P_{ij}^{(1,0)}]\lambda_j + [P_{ij}^{(0,1)} - P_{ij}^{(1,1)}](1-\lambda_j)) \qquad (13)$$

From eq 13, a pruning argument for minimization can be derived, which is equivalent to the first-order DEE pruning rule applied to the special case of only two options at each site. Whenever the gradient with respect to a parameter $\lambda_i$ is negative for all values of $\boldsymbol{\lambda}$ in the hypercube, then $\lambda_i = 1$ minimizes $\tilde{P}$. This condition is precisely met when inequality 14 is fulfilled.

$$\frac{\partial \tilde{P}}{\partial \lambda_i} < 0 \Leftrightarrow P_i^{(0)} - P_i^{(1)} < \sum_{j \neq i} \min\{P_{ij}^{(1,0)} - P_{ij}^{(0,0)}, P_{ij}^{(1,1)} - P_{ij}^{(0,1)}\} \qquad (14)$$

Conversely, a positive gradient (eq 15) implies that $\lambda_i = 0$ minimizes $\tilde{P}$. Thus, it has been demonstrated that $\tilde{P}$ naturally leads to DEE-like algorithms.

$$\frac{\partial \tilde{P}}{\partial \lambda_i} > 0 \Leftrightarrow P_i^{(0)} - P_i^{(1)} > \sum_{j \neq i} \max\{P_{ij}^{(1,0)} - P_{ij}^{(0,0)}, P_{ij}^{(1,1)} - P_{ij}^{(0,1)}\} \qquad (15)$$

**2.1. Library Construction and Ordering.** The choice of enumeration of the library $\mathscr{L}$ determines the assignment of specific molecules to $\boldsymbol{\lambda}$. Consequently, this choice greatly influences the characteristics of $\tilde{P}$, such as its smoothness. Considering the example of Figure 1, the energy rises in all directions only for $C_4H_{10}$. But if $CH_4$ and $C_3H_8$ exchange places in the order, then going from $C_3H_8$ to either of its two neighbors ($C_2H_6$ or $CH_4$) increases the energy, so that a "hurdle" has to be overcome to reach $C_4H_{10}$. Just exchanging the position of two neighboring molecules in the library changes the sign of the derivative at the corresponding $\boldsymbol{\lambda}$. If the Hessian of the pairwise-additive property function is positive-semidefinite, the corresponding $\tilde{P}$ is convex and optimization quickly reaches the global minimum. Using steepest gradient or Newton−Raphson algorithms locates property extrema (minima). It is beneficial to find an ordering of the library that produces a convex property surface. The linearity in each parameter $\lambda_i$ implies convexity of $\tilde{P}$ with respect to that parameter.

Assuming that molecules of similar structure have similar properties, a measure of similarity may be used to decrease the ruggedness/convexity of $\tilde{P}$. One choice to facilitate smooth property surfaces is the enumeration of molecules

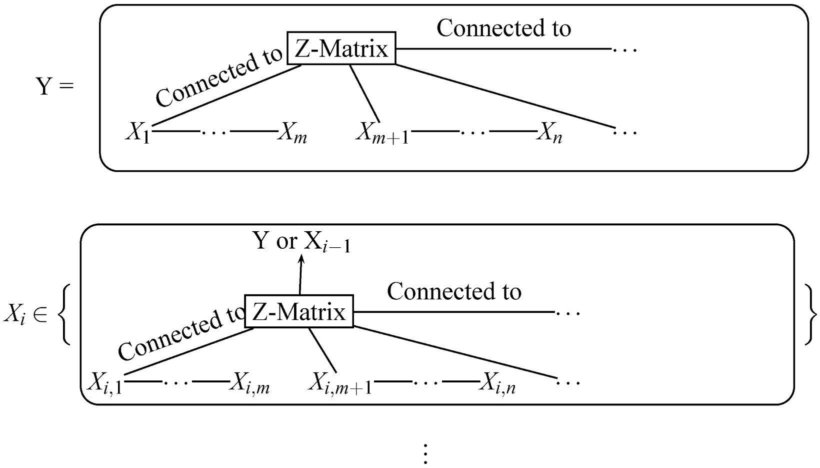

Y =
Z-Matrix
Connected to
Connected to
$X_1$ ... $X_m$ $X_{m+1}$ ... $X_n$ ...
$X_i \in$
Y or $X_{i-1}$
Z-Matrix
Connected to
Connected to
$X_{i,1}$ ... $X_{i,m}$ $X_{i,m+1}$ ... $X_{i,n}$ ...

***Figure 2.*** Substitution pattern hierarchy. Y contains a Z-matrix that has several open "valences". The first can be filled with substituents found in $X_1$, which are connected to substituents found in $X_2$, etc. The second is filled from $X_m$ in the same manner. The $X_i$ themselves are taken from a set of substitution patterns of the same kind as Y. Each instance is anchored to Y at the appropriate valence. The substitutions are terminated by Z-matrices that have no open valences.

by subsequent substitutions from a starting compound (see Figure 2). Returning to the example in Figure 1, Y contains the Z-matrix of $CH_3$ with the connectivity information for $X_1$, which consists of the Z-matrix of H and $CH_2X_{1,1}$ and connectivity information for $X_{1,1}$ (level 1), which contains H and $CH_2X_{2,1}$ (level 2), which finally contains H and $CH_3$ (level 3). Evidence provided by the LCAP approach supports the supposition of smoothness when using substitutions.[21] The substitutions may be defined recursively; therefore, each level of a hierarchy of substitutions consists of a molecular fragment or atom to be connected to the next higher level, a list of substitution sites, and a set of subsequent levels for each site (see Figure 2). Each element of the set of subsequent levels is identified with a coefficient between 0 and 1, and the sum of these coefficients for each set must equal 1 (see eq 16). For a case in which more than two possible substitutions are available at a site, the bit-string representation must be extended to allow mixed numeric bases $b_k$. The general properties discussed in the preceding sections remain unchanged in this alternative interpolation (eq 18). The advantage of this description is the increase of convexity throughout a single substitution site.

$$\sum_j \lambda_{ij} = 1, \qquad j \in \{0, ..., b_i - 1\} \tag{16}$$

$$s = \sum_i \left(\prod_{k=0}^{i-1} b_k\right) \sum_{j \in b_i} s(i,j) \times j, \qquad s(i,j) \in \{0,1\}, \qquad \sum_j s(i,j) = 1 \tag{17}$$

$$\tilde{P}(\{\lambda_{ij}\}_{j \in \{0,...,b_i\},i}) = \sum_{s=0}^{N-1} P_s\left(\prod_i \prod_{j=0}^{b_i-1} \lambda_{ij}^{s(i,j)}\right) \tag{18}$$

**Inclusion of Multiple Conformational States.** For each molecule, it is important to find low-energy conformers for the property optimization to be physically meaningful. For each molecule in the molecular library, another optimization can be started with the (second) library consisting of the corresponding conformers. Each dihedral degree of freedom can be treated as a substitution site at the lowest level with a number of rotations as possible substitutions, as is commonly done in conformational searches.[6,22] In this manner, the conformational search can be introduced as the lowest level in the previously described substitution hierarchy. Thus, the conformational search precedes property computation in property optimizations. More general constraints on the optimal molecule can be introduced *via* alternate methods, like Lagrange multipliers or stochastic algorithms. Lagrange multipliers can be implemented using (soft) penalty functions with weightings that increase throughout the optimization.

**Algorithm.** Here, a line search algorithm is used; in particular, each parameter $\lambda_i$ is followed to a minimum in that direction before varying the next parameter $\lambda_{i+1}$. Maximization *via* this algorithm can be achieved for instance by minimizing the negative objective function. This line search algorithm is an implicit branch-and-bound algorithm. A flowchart for the employed recursive algorithm appears in Figure 3, and application of the algorithm to a small example will be discussed in section 3 under the subsection Framework A (see also the accompanying Figure 6).

Since $\tilde{P}(\lambda)$ is locally convex, this algorithm converges locally. The line-search steps 4−7 in Figure 3 correspond to a linear tree search or branch-and-bound algorithm. The computational complexity is on the order $O(\log N)$ in the library size $N$ due to the linear dependence on the log $N$ variables. In contrast to conventional branch-and-bound methods, no structures are explicitly excluded from the search space. Since each molecule chosen in step 8 in Figure 3 is strictly better in the sense of property optimization than its predecessor, the algorithm quickly converges to a local property value minimum in the library.[20]

All property minima for this algorithm are minima for the steepest-descent derived method and *vice versa*. This algorithm traverses the library in a smoother fashion compared to the steepest-descent derived method, successfully employed by Keinan et al.,[15] because the molecules are

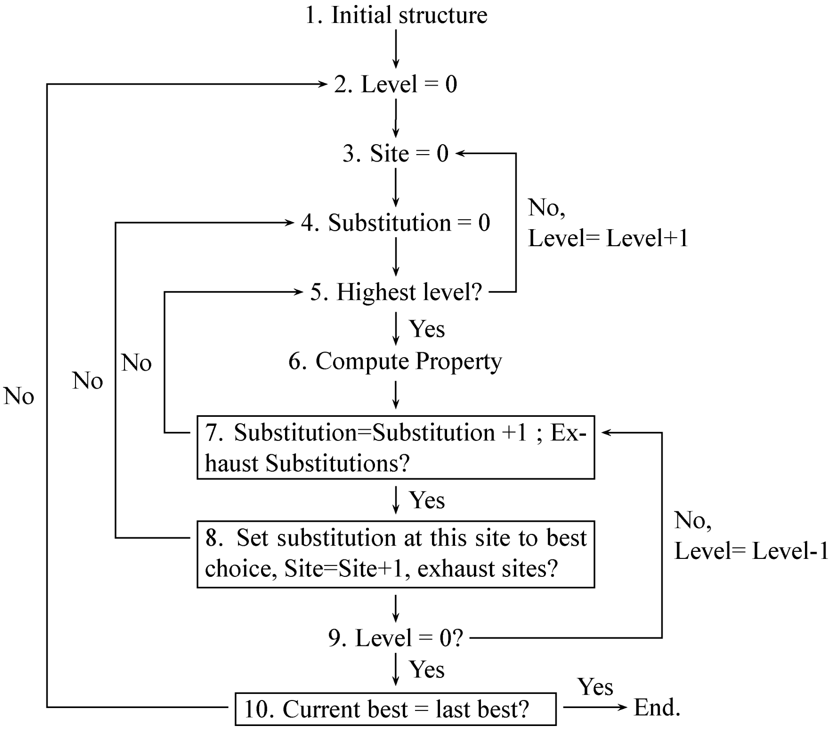


***Figure 3.*** Flowchart of the algorithm.

traversed variationally by single substitutions. While on one hand the steepest-descent based approach can sidestep barriers in the immediate vicinity efficiently, due to the simultaneous change of potentially several bits, the variational nature of this line search guarantees convergence, which is particularly useful on rugged property surfaces.

For the sake of computational accessibility, all geometries were optimized using the semiempirical Austin model 1 (AM1) method as implemented in *Gaussian03*.[23] The static electronic hyperpolarizability was computed using the INDO/S method as implemented in CNDO by Reimers et al.[24] using the sum-overstates expression in eq 19. The configuration interaction (CI) space was spanned by up to 100 unoccupied or occupied orbitals to accommodate for the large number of electrons in some of the investigated systems

$$\beta_{ijk} = \sum_{\nu\kappa} \frac{\langle 0|x_i|\nu\rangle\langle\nu|x_j - \mu_j|\kappa\rangle\langle\kappa|x_k|0\rangle}{E_{0\nu}E_{0\kappa}} \tag{19}$$

$$\beta_i = \frac{1}{3}\sum_j (\beta_{ijj} + \beta_{jij} + \beta_{jji}) \tag{20}$$

$$\beta_\mu = \frac{\vec{\mu}}{||\vec{\mu}||}\cdot\vec{\beta}, \beta_0 = ||\vec{\beta}|| \tag{21}$$

where $E_{0\nu}$ is the excitation energy from the ground state to the $\nu$th excited state, $\vec{\beta}$ is the static electronic hyperpolarizability with components $\beta_i$ and corresponding hyperpolarizability tensor elements $\beta_{ijk}$, $\beta_0$ is the isotropic hyperpolarizability, $\beta_\mu$ is the hyperpolarizability component in direction of the ground state dipole moment, $\vec{x}$ is the dipole operator with components $x_i$, and $\vec{\mu}$ is the ground state dipole moment with components $\mu_i$.

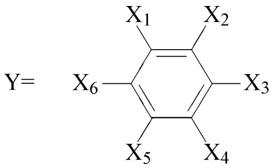


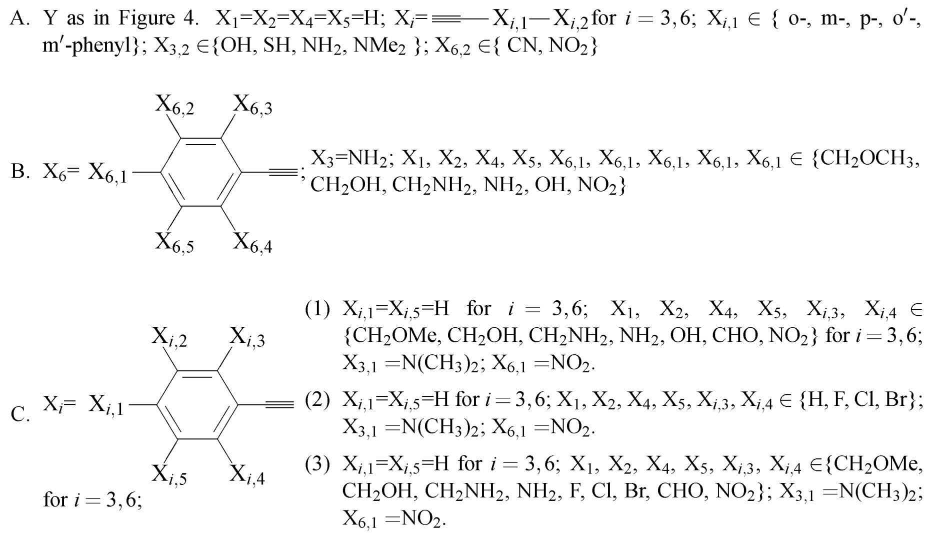
***Figure 4.*** Top level of the substitution scheme (see Figure 2).

Figures 4 and 5 summarize the tolane-based system studies. Tolane spectroscopic properties are favorable for applications, so their first and second hyperpolarizabilities have been studied extensively.[25,26] In addition, these structures are readily modified[27] and present a large number of possible derivatives. Tolanes therefore present a particularly rich testbed for these optimization studies.

## 3. Results and Discussion

Overall, five different tolane libraries were investigated (general structure in Figure 2). The first three sets of molecules are optimized with respect to their static isotropic hyperpolarizability $\beta_0$ (eq 21), while the remaining sets are optimized with respect to the component of the hyperpolarizability in direction of the dipole $\beta_\mu$ (eq 21).

**Framework A.** Validation of the algorithm was performed on the structure framework A in Figure 5. Figure 6 shows the progress of the algorithm. There are 200 molecules in this library, but hyperpolarizabilities of only 24 different molecules were computed during the optimization, the minimum number of molecules required for the algorithm to finish the optimization. Regardless of the starting structure, the algorithm consistently finishes with the global hyperpolarizability optimum (Figure 6), which has also been confirmed experimentally.[28] For comparison, if the library



A. Y as in Figure 4. $X_1$=$X_2$=$X_4$=$X_5$=H; $X_i$=≡—$X_{i,1}$—$X_{i,2}$ for $i = 3,6$; $X_{i,1} \in$ { o-, m-, p-, o′-, m′-phenyl}; $X_{3,2} \in${OH, SH, $NH_2$, $NMe_2$ }; $X_{6,2} \in${ CN, $NO_2$}

B. $X_6$= $X_{6,1}$ ; $X_3$=$NH_2$; $X_1$, $X_2$, $X_4$, $X_5$, $X_{6,1}$, $X_{6,1}$, $X_{6,1}$, $X_{6,1}$, $X_{6,1}$ ∈ {$CH_2OCH_3$, $CH_2OH$, $CH_2NH_2$, $NH_2$, OH, $NO_2$}

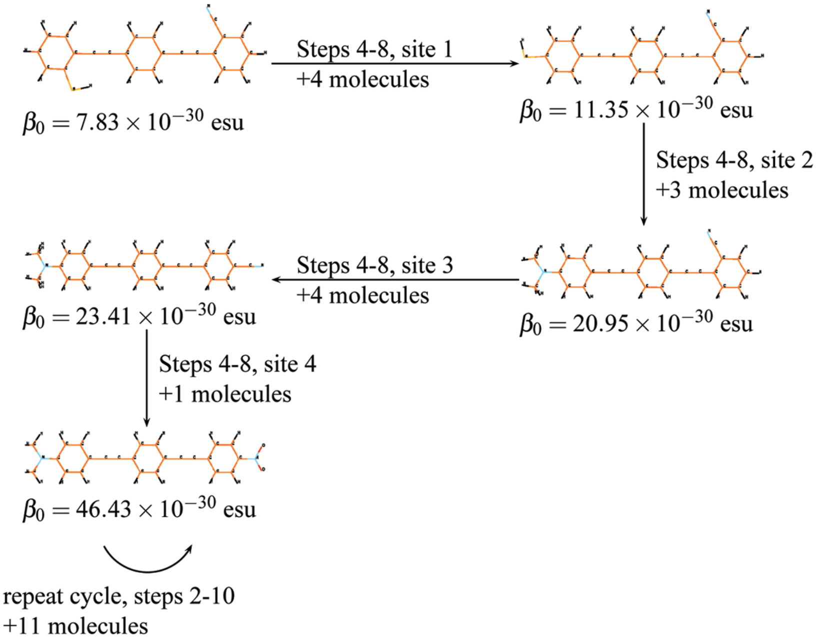
C. $X_i$= $X_{i,1}$ for $i = 3,6$;

(1) $X_{i,1}$=$X_{i,5}$=H for $i = 3,6$; $X_1$, $X_2$, $X_4$, $X_5$, $X_{i,3}$, $X_{i,4}$ ∈ {$CH_2OMe$, $CH_2OH$, $CH_2NH_2$, $NH_2$, OH, CHO, $NO_2$} for $i = 3,6$; $X_{3,1}$ =$N(CH_3)_2$; $X_{6,1}$ =$NO_2$.

(2) $X_{i,1}$=$X_{i,5}$=H for $i = 3,6$; $X_1$, $X_2$, $X_4$, $X_5$, $X_{i,3}$, $X_{i,4}$ ∈ {H, F, Cl, Br}; $X_{3,1}$ =$N(CH_3)_2$; $X_{6,1}$ =$NO_2$.

(3) $X_{i,1}$=$X_{i,5}$=H for $i = 3,6$; $X_1$, $X_2$, $X_4$, $X_5$, $X_{i,3}$, $X_{i,4}$ ∈{$CH_2OMe$, $CH_2OH$, $CH_2NH_2$, $NH_2$, F, Cl, Br, CHO, $NO_2$}; $X_{3,1}$ =$N(CH_3)_2$; $X_{6,1}$ =$NO_2$.

***Figure 5.*** Tolane libraries investigated. Terminology as in Figure 2 with the top level as in Figure 4.



Steps 4-8, site 1
+4 molecules
$\beta_0 = 7.83 \times 10^{-30}$ esu
$\beta_0 = 11.35 \times 10^{-30}$ esu
Steps 4-8, site 2
+3 molecules
Steps 4-8, site 3
+4 molecules
$\beta_0 = 23.41 \times 10^{-30}$ esu
$\beta_0 = 20.95 \times 10^{-30}$ esu
Steps 4-8, site 4
+1 molecules
$\beta_0 = 46.43 \times 10^{-30}$ esu
repeat cycle, steps 2-10
+11 molecules

***Figure 6.*** Progress of the optimization algorithm. The steps refer to the steps in Figure 3. The number of molecules indicated is the number of previously unvisited molecules for which the property is computed in performing the steps. Carbons are marked in orange, hydrogens in white, oxygens in red, and nitrogens in light blue.

is searched randomly, the expected number of computed molecules before finding the global minimum is 200 molecules. If repeats are avoided, then still 101 molecules would need to be computed on average in order to obtain the same result.

**Framework B.** The static hyperpolarizability $\beta_0$ of framework B in Figure 5 optimizes to an unstable, perhaps explosive structure with mostly nitro- and amino-substituents (Figure 7). The final computed $\beta_0$-value was 131.9 × $10^{-30}$ esu after 121 computed structures from $6^8 \approx 1.7 \times 10^6$ possible molecules. Additionally, conformational analysis was performed. CHO and OH were allowed two possible orientations in the plane of the tolane. For $CH_2OH$ and $CH_2NH_2$, 3-fold rotation around the C−O and C−N bonds, respectively, was included, while only 2-fold rotations around the bonds connecting to the tolane framework were allowed.

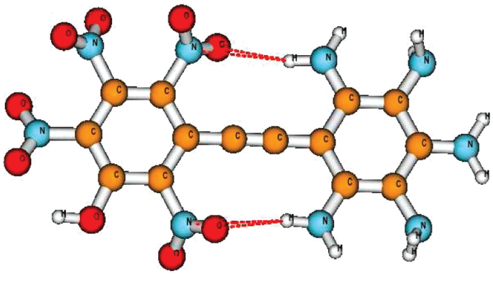
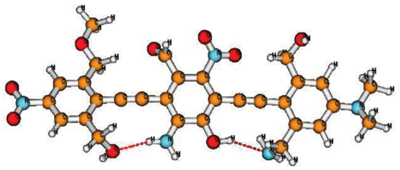
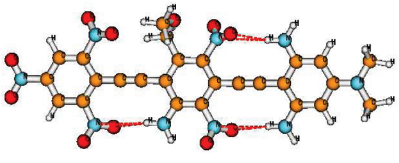
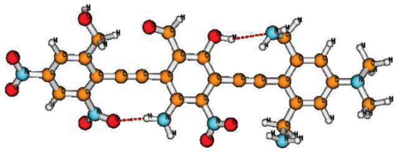
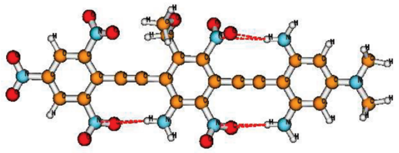
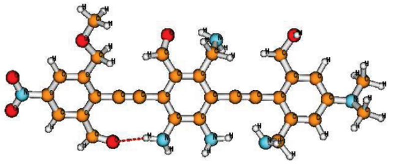
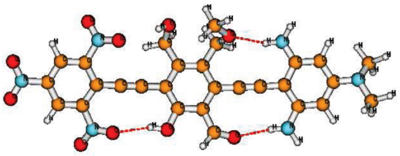


***Figure 7.*** Final structure of framework B. Carbons are marked in orange, hydrogens in white, oxygens in red, and nitrogens in light blue.

**Framework C-1.** The static hyperpolarizability for compounds in C-1 of Figure 5 was optimized starting from three different randomly chosen initial structures. A total of $7^8 \approx 5.8 \times 10^6$ possible molecules exist in this family. Confor-

***Table 1.*** Starting and Final Structures of Framework C-1 of Figure 5[a]

| Run | Initial structure | Final structure |
|---|---|---|
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |

[a] Carbons are marked in orange, hydrogens in white, oxygens in red, and nitrogens in light blue.

***Table 2.*** Starting and Final Hyperpolarizabilities and Number of Computed Molecules for Framework C-1 in Figure 5[a]

| run | initial $\beta_0/10^{-30}$ esu | final $\beta_0/10^{-30}$ esu | molecules computed |
|---|---|---|---|
| 1 | 55.1 | 214.6 | 157 |
| 2 | 71.0 | 214.6 | 109 |
| 3 | 49.9 | 216.9 | 169 |

[a] See also Table 1.

mational considerations were treated as in framework B. Two of the three runs converged to the same structure ($\beta_0 = 214.6 \times 10^{-30}$ esu), while the third converged to a second structure with comparable hyperpolarizability ($\beta_0 = 216.9 \times 10^{-30}$ esu, see Tables 1 and 2). All three runs finished after computing less than 0.1% of all possible molecules and achieved 3- to 4-fold improvements of the hyperpolarizability. Comparing the two structures, we see some common motifs emerge: the variable fragments $X_{3,3}$ and $X_{3,4}$ contain nitro-groups, while $X_{6,3}$ and $X_{6,4}$ are occupied by amino-groups. Furthermore, positions $X_2$ and $X_4$ are occupied by electron acceptors, and sites $X_1$ and $X_5$ are occupied by electron donors. It is notable that not all positions are occupied by the "strongest" donors or acceptors in the substitution set, i.e., $NH_2$ and $NO_2$, respectively.

**Framework C-2.** Halogen substituents do not necessitate extensive conformational analysis, so they allow the evaluation of the optimization method without added constraints. The structures C-2 in Figure 5 were optimized for the hyperpolarizability in the direction of the dipole moment ($\beta_\mu$, see eq 21). Entries a and c in Table 3 show the results of two optimizations of framework C-2 in Figure 5 starting from the same initial structure with all substitutions set to hydrogens. In this case, convergence to a hyperpolarizability maximum is confirmed to be logarithmic in the library size; i.e., squaring the library size from 256 to 65536 leads to roughly twice the number of computed molecules.

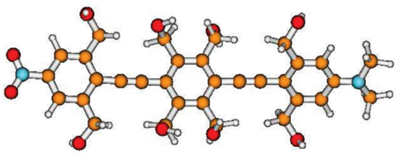
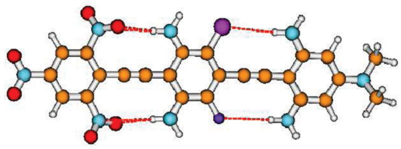
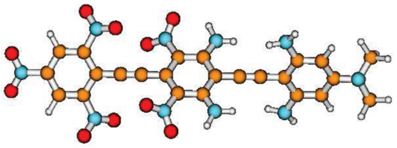
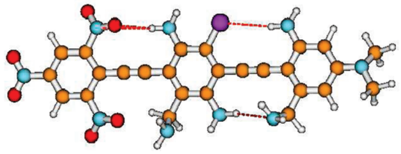
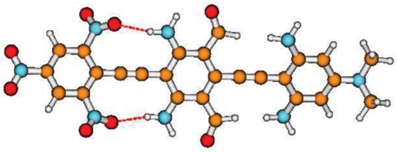
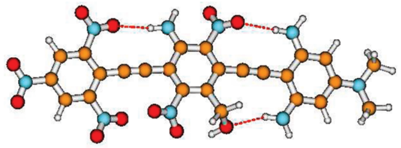
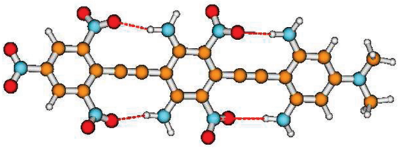
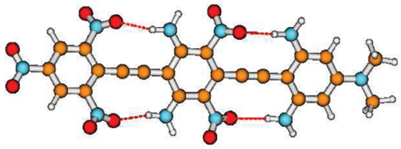
***Table 3.*** Optimized Structures for Frameworks C-2 in Figure 5

| | compound | $\beta_\mu/10^{-30}$ esu | molecules computed | library size |
|---|---|---|---|---|
| a | $X_1$, $X_5$, $X_{3,3}$, $X_{3,4}$ = H<br>$X_{6,3}$, $X_{6,4}$ = Br<br>$X_2$ = C1<br>$X_4$ = F | 84.1 | 67 | 65536 |
| b | $X_1$, $X_2$, $X_4$, $X_5$, $X_{3,3}$, $X_{3,4}$ = H<br>$X_{6,3}$, $X_{6,4}$ = Br | 77.4 | 69 | 65536 |
| c | $X_1$, $X_5$ = Br<br>$X_{3,3}$, $X_{3,4}$ = H<br>$X_{6,3}$, $X_{6,4}$ = Br<br>$X_2$, $X_4$ = F | 83.5 | 28 | 256 |
| d | $X_1$, $X_5$, $X_{3,3}$, $X_{3,4}$ = H<br>$X_{6,3}$, $X_{6,4}$ = Br<br>$X_2$, $X_4$ = Br | 83.2 | 28 | 256 |

The stability of the optimization procedure was tested by constraining substitutions to be symmetric with respect to the mirror plane perpendicular to the plane of the backbone (runs (c) and (d) in Table 3), as well as starting from different initial structures: runs (a) and (c) were started with all substituents set to hydrogen, while run (b) starts from $X_{6,3}$ = Br and $X_{3,4}$ = F, and run (d) starts from $X_{6,3} = X_{6,4}$ = Br and $X_{3,3} = X_{3,4}$ = F. The hyperpolarizabilities of the initial structures were within 4 units of $50 \times 10^{-30}$ esu. Since the procedure is not a global optimization algorithm, it is possible to end at different local maxima, here each run ended in a different structure with corresponding hyperpolarizabilities ($\beta_\mu/10^{-30}$ esu = 84.1, 77.4, 83.5, 83.2, respectively, see Table Table 3). Nonetheless, the optimizations lead to significant and comparable improvements between runs. The found maxima all place bromine in the $X_{6,3}$ and $X_{6,4}$ positions, implying that a large fraction of the gain in $\beta_\mu$ arises from bromine to amino charge transfer interactions.

**Framework C-3.** In combination with parts of libraries of C-1 and C-2 in Figure 5, structures C-3 in Figure 5 were subjected to optimization of the static hyperpolarizability in

***Table 4.*** Starting and Final Structures of Framework C-3 in Figure 5[a]

| Run | Initial structure | Final structure |
|---|---|---|
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |
| 4 |  |  |

[a] Carbons are marked in orange, hydrogens in white, oxygens in red, nitrogens in light blue, bromine in dark red, fluorine in dark blue, and chlorine in purple.

***Table 5.*** Starting and Final Hyperpolarizabilities and Number of Computed Molecules for Framework C-3 in Figure 5[a]

| run | initial $\beta_\mu/10^{-30}$ esu | final $\beta_\mu/10^{-30}$ esu | no. comp |
|---|---|---|---|
| 1 | 37.0 | 181.5 | 181 |
| 2 | 55.6 | 171.6 | 153 |
| 3 | 139.8 | 191.6 | 161 |
| 4 | 173.3 | 173.3 | 65 |

[a] See also Table 4.

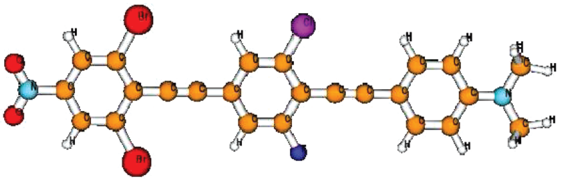


***Figure 8.*** Largest $\beta_\mu$ structure for framework C-2 in Figure 5. Compare to entry a in Table 3. Carbons are marked in orange, hydrogens in white, oxygens in red, nitrogens in light blue, bromine in dark red, fluorine in dark blue, and chlorine in purple.

the direction of the dipole moment ($\beta_\mu$). Four optimizations from different starting configurations were performed (see Tables 4 and 5 for results). The "unbiased" first optimization leads to a 5-fold increase in $\beta_\mu$ (37.0 → 181.5 × $10^{-30}$ esu). The final structure (see Table 4) indeed is a mixture of the results for C-1 and C-2 in Figure 5. The second optimization was started with a structure concentrating equal numbers of donors on one side and acceptors on the other, analogous to the final structure of framework B in Figure 5. This starting structure exhibited only a marginally larger hyperpolarizability (55.6 × $10^{-30}$ esu) than the "unbiased" starting structure, but optimized to an alternating donor−acceptor arrangement (171.6 × $10^{-30}$ esu) that failed to reach the optimum found in the first optimization. The low hyperpolarizability is presumably due to the benzene rings twisting out of plane and reducing conjugation.

A biased starting point, with alternating donor and acceptor groups, leads to a marginally increased final hyperpolarizability (191.6 × $10^{-30}$) over the first optimization. The attempt to exceed this value by substituting the "strongest" electron donors and acceptors, $NH_2$ and $NO_2$, fails despite the fact that this structure is indeed a local maximum (173.3 × $10^{-30}$ esu). All four optimization runs finish computing less than 0.001% out of the possible $9^8 \approx 4.3 \times 10^7$ molecules.

## 4. Summary and Conclusion

We have introduced an embedding of discrete molecular spaces in a continuous space, similar to the embedding of discrete Hamiltonians in LCAP.[21] From the embedding, an optimization based on differentiation in the continuous space was developed. The embedding is based on the chemically intuitive ordering of molecules by substitutions. Assuming that single substitutions are small perturbations, the ordering also increases smoothness in the resultant continuous space. Although the framework is very general, it is limited to properties that can be derived and computed by defining substitution patterns as well as computational accessibility, such as binding problems, linear spectra, or stress−strain curves of molecules.

The theoretical framework transforms a discrete optimization problem into a continuous optimization problem, which then gives rise to a discrete optimization strategy. The theoretical complexity of the used line-search algorithm is $O(\log N)$ in the library size $N$, and applications of the algorithm to a variety of conditions confirm the method's effectiveness. A design strategy for tolanes of alternating donors and acceptors along a conjugated framework is suggested by the optimization results. Choosing a set of initial structures increases the likelihood of finding the global optimum. Further applications and improvements are under study including an extension to second-order derivative methods, probabilistic methods[29,30] and dynamic ordering of the parameters to achieve overall convexity.

**Acknowledgment.** We would like to thank S. Keinan, B. Desinghu, G. Lindsay, A. Chafin, and M. Davis for helpful discussions. We are thankful to DARPA Predicting Real Optimized Materials through ARO (W911NF-04-1-0243) and the Army Research Laboratory for funding.

## References

(1) Andrekson, P. A.; Westlund, M. *Laser Photonics Rev.* **2007**, *1*, 231–248.

(2) Bergmann, G.; Ellis, C.; Hindmarsh, P.; Kelly, S. M.; O'Neill, M. *Mol. Cryst. Liq. Cryst.* **2001**, *368*, 4439–4446.

(3) Dalton, L. R.; Sullivan, P. A.; Bale, D. H.; Bricht, B. C. Theory-inspired nano-engineering of photonic and electronic materials: Noncentro symmetric charge-transfer electro-optic materials. In *3rd Nano and Giga Forum*; Pergamon-Elsevier Science Ltd.: Oxford, U.K., 2007; pp 1263−1277.

(4) van Deursen, R.; Reymond, J.-L. *Chem. Med. Chem.* **2007**, *2*, 636–640.

(5) Michalewicz, Z.; Fogel, D. B. *How to Solve It: Modern Heuristics*; Springer Verlag: Berlin, 2002.

(6) Gordon, D. B.; Mayo, S. L. *J. Comput. Chem.* **1998**, *19*, 1505–1514.

(7) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. *Science* **1983**, *220*, 671–680.

(8) Muhlenbein, H.; Gorgesschleuter, M.; Kramer, O. *Parallel Comput.* **1988**, *7*, 65–85.

(9) Franceschetti, A.; Dudiy, S. V.; Barabash, S. V.; Zunger, A.; Xu, J.; van Schilfgaarde, M. *Phys. Rev. Lett.* **2006**, *97*, 047202.

(10) Dudiy, S. V.; Zunger, A. *Phys. Rev. Lett.* **2006**, *97*, 046401.

(11) Franceschetti, A.; Zunger, A.; van Schilfgaarde, M. *J. Phys.: Condens. Matter* **2007**, *19*, 242203.

(12) Piquini, P.; Graf, P. A.; Zunger, A. *Phys. Rev. Lett.* **2008**, *100*, 186403.

(13) Wang, M. L.; Hu, X. Q.; Beratan, D. N.; Yang, W. T. *J. Am. Chem. Soc.* **2006**, *128*, 3228–3232.

(14) Keinan, S.; Hu, X. Q.; Beratan, D. N.; Yang, W. T. *J. Phys. Chem. A* **2007**, *111*, 176–181.

(15) Keinan, S.; Paquette, W. D.; Skoko, J. J.; Beratan, D. N.; Yang, W. T.; Shinde, S.; Johnston, P. A.; Lazo, J. S.; Wipf, P. *Org. Biomol. Chem.* **2008**, *6*, 3256–3263.

(16) von Lilienfeld, O. A.; Lins, R. D.; Rothlisberger, U. *Phys. Rev. Lett.* **2005**, *95*, 153002.

(17) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *J. Chem. Phys.* **2005**, *122*, 014113.

(18) von Lilienfeld, O. A.; Tuckerman, M. E. *J. Chem. Phys.* **2006**, *125*, 154104.

(19) Marcon, V.; von Lilienfeld, O. A.; Andrienko, D. *J. Chem. Phys.* **2007**, *127*, 064305.

(20) Desinghu, B.; Yang, W.; Beratan, D. N. *J. Chem. Phys.* **2008**, *129*, 174105.

(21) Xiao, D.; Yang, W.; Beratan, D. N. *J. Chem. Phys.* **2008**, *129*, 44106.

(22) Izgorodina, E. I.; Lin, C. Y.; Coote, M. L. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2507–2516.

(23) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 03 Technical Report*; Gaussian Inc.: Wallingford, CT, 2004.

(24) Tejerina, B.; Reimers, J. *CNDO/INDO*; 2007 (accessed 08/08/2009); DOI 10254/nanohub-r3352.5.

(25) Liu, C. P.; Liu, P.; Wu, K. C. *Acta Chim. Sinica* **2008**, *66*, 729–737.

(26) Oliva, M. M.; Casado, J.; Hennrich, G.; Navarrete, J. T. L. *J. Phys. Chem. B* **2006**, *110*, 19198–19206.

(27) Traber, B.; Oeser, T.; Gleiter, R. *Eur. J. Org. Chem.* **2005**, 1283–1292.

(28) Nguyen, P.; Lesley, G.; Marder, T. B.; Ledoux, I.; Zyss, J. *Chem. Mater.* **1997**, *9*, 406–408.

(29) Hu, X.; Beratan, D. N.; Yang, W. *J. Chem. Phys.* **2008**, *129*, 064102.

(30) Mueller, T.; Ceder, B. *Phys. Rev. B* **2009**, *80*, 024103.

CT900325P

JCTC Journal of Chemical Theory and Computation

# Erratum

**Efficient Diffuse Basis Sets: cc-pV*x*Z+ and maug-cc-pV*x*Z.** [*J. Chem. Theory Comput. 5,* 1197–1202 (2009)]. By Ewa Papajak, Hannah R. Leverentz, Jingjing Zheng, and Donald G. Truhlar*.

Pages 1199. Some data in Tables 3-6 are corrected. These corrections do not change any of our discussion or conclusions in the paper. In the second paragraph of Section 4, “cc-pVDZ+” should be “cc-pVTZ+”.

***Table 3.*** Mean Unsigned Errors (MUEs) (in kcal/mol) in Ionization Potentials

| | B3LYP | M06-2X | CCSD(T) |
|---|---|---|---|
| cc-pVDZ+ | 4.88 | 3.09 | 8.57 |
| aug-cc-pVTZ | | 2.70 | |

***Table 4.*** Mean Unsigned Errors (MUEs) (in kcal/mol) in Electron Affinities

| | B3LYP | M06-2X | CCSD(T) |
|---|---|---|---|
| cc-pVDZ | | 20.10 | |
| cc-pVDZ+ | 3.17 | 2.66 | 9.77 |
| aug-cc-pVDZ | | 2.37 | |
| cc-pVTZ | | 9.85 | |
| cc-pVTZ+ | | 1.92 | |
| aug-cc-pVTZ | | 1.55 | |

***Table 5.*** Mean Unsigned Errors Per Bond (MUEPBs) in (kcal/mol) in Atomization Energies

| | B3LYP | M06-2X | CCSD(T) |
|---|---|---|---|
| cc-pVDZ+ | 3.15 | 2.40 | 8.88 |

***Table 6.*** Mean Unsigned Errors (MUEs) (in kcal/mol) in the Barrier Heights of the DBH24/08 Database

| | HATBH6 | NSBH6 | UABH6 | HTBH6 | DBH24 |
|---|---|---|---|---|---|
| B3LYP/cc-pVDZ+ | 7.57 | 3.79 | | 5.81 | 4.80 |
| M06-2X/cc-pVDZ+ | 2.02 | 1.17 | | 1.37 | 1.45 |
| CCSD(T)/cc-pVDZ+ | 4.15 | 0.82 | | 1.79 | 2.05 |

**Addendum**. We also present here some further calculations that do not correct an error in the original article but that provide further relevant information. In particular, we note that the article tested the new plus basis sets for ionization potentials, electron affinities, atomization energies, barrier heights, and basis set superposition errors. We then presented tests of another set of basis sets, called maug basis sets, obtained by truncating the aug basis sets to the same size as the plus basis sets. The maug basis sets were tested only for barrier heights and basis set superposition errors, and we found very similar performance to the plus basis sets. As an example of the differences in the basis sets, diffuse functions on O in maug-cc-pVTZ have exponential parameters of 0.07376 for *s* functions and 0.05974 for *p*; these exponential parameters are smaller than those in the plus basis set, where both parameters are 0.0845. The most difficult tests of the adequacy of a scheme for diffuse basis functions are provided by electron affinities. We have now tested maug-cc-pV*x*Z against cc-pV*x*Z+ with both $x = \mathrm{D}$ and $x = \mathrm{T}$ for electron affinities, and we found better performance with the maug basis sets for M06-2X (better on average) and CCSD(T) (always better), especially for systems containing oxygen atoms (and to a lesser extent for $Si^-$ and $C^-$), but better performance (on average) with the plus basis set for B3LYP. However, in all 78 cases the anion energies are lower for the maug basis set than the corresponding plus one, so the improvement of the plus basis sets for B3LYP electron affinities is presumably due to cancellation of basis set error with a large error in the opposite direction from the functional itself. Table A1 gives two additional rows for the original Table 4 that show the mean unsigned errors in electron affinities with two maug basis sets. The conclusion is that anion energies and electron affinities are more sensitive than barrier heights and basis set superposition errors to the precise values of the diffuse exponents, and the maug basis sets are more accurate for such calculations, probably because the exponents were optimized for atomic anions.[1]

***Table A1.*** Mean Unsigned Errors (MUEs) (in kcal/mol) in Electron Affinities

| | B3LYP | M06-2X | CCSD(T) |
|---|---|---|---|
| maug-cc-pVDZ | 3.19 | 2.46 | 9.41 |
| maug-cc-pVTZ | 2.49 | 1.57 | 4.88 |

## Reference

(1) Kendall, R. A.; Dunning, Jr., T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.

CT9004905

10.1021/ct9004905
Published on Web 10/15/2009

Published on Web 10/15/2009