JCTC Journal of Chemical Theory and Computation

Using Molecular Dynamics Simulations To Provide New Insights into Protein Structure on the Nanosecond Timescale: Comparison with Experimental Data and Biological Inferences for the Hyaluronan-Binding Link Module of TSG-6

Andrew Almond,*[,†] Charles D. Blundell,[†,§] Victoria A. Higman,[§] Alexander D. MacKerell, Jr.,[||] and Anthony J. Day[‡,§]

Manchester Interdisciplinary Biocentre, Faculty of Life Sciences, University of Manchester, Princess Street, Manchester M1 7DN, U.K., Michael Smith Building, Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester M13 9PT, U.K., MRC Immunochemistry Unit, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, U.K., and Department of Pharmaceutical Chemistry, School of Pharmacy, University of Maryland, 20 Penn Street, Baltimore, Maryland 21201

Received July 20, 2006

Abstract: Link module domains play an essential role in extracellular matrix assembly and remodeling by binding to the flexible glycosaminoglycan hyaluronan. A high-resolution NMR-structure of the Link module from the protein product of tumor necrosis factor-stimulated gene-6 (Link_TSG6) has been determined, but a fuller appreciation of protein dynamics may be necessary to understand its hyaluronan-binding. Therefore, we have performed a 0.25 μs MD simulation, starting from the lowest-energy NMR-derived solution structure of Link_TSG6, with explicit water and ions, using the CHARMM22 protein force field. The simulation was as good a fit to the NMR data as the ensemble from simulated annealing, except in the β5-β6 loop. Furthermore, analysis revealed that secondary structure elements extended further than previously reported and underwent fast picosecond time scale dynamics, whereas nanosecond dynamics was found in certain loops. In particular, surface side chains proposed to interact with glycosaminoglycans were predicted to be highly mobile and be directed away from the protein surface. Furthermore, the hyaluronan-binding β4-β5 loop remained in a closed conformation, favoring an allosteric interaction mechanism. This enhanced view of the Link module provides general insight into protein dynamics and may be helpful for understanding the dynamic molecular basis of tissue assembly, remodeling, and disease processes.

1. Introduction

Submicrosecond conformational rearrangements in proteins provide the specificity and plasticity required to associate with other molecules.[1−4] Unfortunately, the two principal techniques used for defining protein structure at atomic resolution, X-ray crystallography (XRC), and nuclear magnetic resonance (NMR) spectroscopy have a limited capacity to obtain such information on conformational rearrangements.[5,6] While, XRC B-factors can, in principle, provide information about the distribution of conformers present in the crystal,[7] the nonphysiological conditions (i.e., the protein is in the solid state and often at very low temperatures) and crystal packing artifacts can perturb protein structure, particularly at surface-exposed side chains and loops.[8,9] In

* Corresponding author phone: +44 161 30 64 199; fax: +44 161 30 68918; e-mail: andrew.almond@manchester.ac.uk.
† Manchester Interdisciplinary Biocentre, Faculty of Life Sciences, University of Manchester.
‡ Michael Smith Building, Faculty of Life Sciences, University of Manchester.
§ University of Oxford.
|| University of Maryland.

10.1021/ct600236q CCC: $37.00
Published on Web 12/02/2006

the case of NMR, while performed in the solution state at physiological temperatures, it typically relies on nuclear Overhauser enhancements (NOEs) to provide local geometric information between neighboring protons. These NOEs are resultant from dynamic molecular processes but are generally interpreted as a set of static structures through simulated annealing, which can lead to unrealistic conformational bias.[10]

A more detailed analysis of protein dynamics in solution can be performed using experimental techniques such as fluorescence anisotropy,[11] measurements of NMR relaxation parameters, and residual dipolar couplings.[12−14] While NMR relaxation is in principle capable of investigating ns-time scale protein dynamics at residue-specific positions, a serious problem is encountered because both internal motion and overall tumbling contribute to dipolar relaxation. Since overall tumbling in proteins occurs on the nanosecond time scale, separation of the two processes cannot be achieved effectively even in the ideal case of isotropic motion. Therefore, ^{15}N-relaxation (and other NMR-relaxation based experiments) is more effective at studying fast picosecond-time scale dynamics rather than those on the nano- to microsecond timescales, on which local protein conformational rearrangements occur.[15]

Due to these experimental limitations, protein nanosecond-time scale dynamics are relatively unexplored in proteins, which is of significant concern considering its crucial role in ligand-binding events. In the absence of detailed experimental data about molecular conformational rearrangements, computer simulations can be employed. Molecular dynamics (MD) simulation using a molecular-mechanics force field is one of the best all-atom theoretical techniques presently available for investigating ns-time scale dynamic atomic motions in proteins and their attendant solvent molecules and ions. With increases in computing power, simulations of proteins are now being performed more routinely on the 0.1 μs time scale.[15] Such simulations are sufficient to investigate nanosecond-time scale dynamics effectively. Furthermore, MD is able to predict with relatively high accuracy the rotamer conformation of the side chains important for molecular recognition, and even in simulations of unbound protein, these frequently visit the rotamer conformations seen in the complex.[4,16] However, MD simulations are based on assumptions about molecular potential energies and at present need to be compared carefully against experimental data to ensure that realistic results are being obtained. In this regard, two recent studies that compared MD simulations of proteins against a variety of experimental data are illustrative.[15,17] The first concerns two simulations of ubiquitin, which were compared against NMR data,[15] and the second is a 3 ns simulation of bacterial cytochrome *c*, which was compared against amide exchange rates and XRC B-factors.[17] These studies show that MD simulations are accurate enough to predict a wide array of experimental data and therefore provide crucial insight into nanosecond dynamics.

Such promising results now give the impetus to understand dynamics in a wider range of experimentally characterized proteins and particularly in proteins that bind to flexible ligands. The Link module, for example, is a structural domain of approximately 100 amino acids that plays a fundamental role in extracellular matrix assembly via its association with the high molecular weight polysaccharide hyaluronan.[18−21] This domain is found in extracellular matrix proteins such as cartilage link protein, aggrecan, cell surface receptors (including CD44), and the inflammation-associated protein tumor necrosis factor-stimulated gene-6 (TSG-6). TSG-6 is a secreted protein comprising a Link and a CUB module that is not constitutively expressed in normal adult tissues. It has roles in inflammation, matrix remodeling, and ovulation via interactions with hyaluronan[21] and also interacts with a wide range of other extracellular matrix molecules including sulfated glycosaminoglycans and various proteins.[22,23] The Link module from human TSG-6, which has been expressed in bacteria (as an 11-kDa, 98-residue construct termed Link_TSG6) and shown to support hyaluronan binding,[24] comprises two α-helices and two antiparallel triple-stranded β-sheets arranged around a large hydrophobic core,[24,25] as illustrated in Figure 1(a). Link_TSG6 interacts with hyaluronan via a binding groove that is opened by rearrangements within the β1-α1 and β4-β5 loops that allow the key residues (K11, Y12, Y59, F70, Y78, and R81) to bind to hyaluronan.[25,26] Bikunin, a serine protease inhibitor that is potentiated through its association with TSG-6, also binds to a similar region (i.e., K11, Y12, Y59, and Y78) via a protein−protein interaction.[27] In contrast, heparin interacts with surface-exposed basic residues (including K20, K34, K41, and K54) on a different face of Link_TSG6.[27] Since hyaluronan and heparin are locally dynamic molecules,[28−31] it is reasonable to suggest that the entropic penalty associated with their binding to the Link module may be lessened by binding them using a dynamic platform of loops and surface residues in the interaction sites. Nanosecond time scale dynamics is therefore likely to be important to the biological roles of Link modules in the extracellular matrix.

NMR structures of Link_TSG6 have been determined in both the free and hyaluronan-bound state.[24,26,32] Since virtually every atom was assigned in free Link_TSG6, the solution structure was determined to high resolution (over 13 NOE constraints per residue) and is therefore a good starting conformation for MD simulation. These assignments also enabled a considerable amount of other NMR data to be collected, making a large experimental data set that can be used to test simulations. Here, we present a 0.25 μs MD simulation of Link_TSG6 that was started from the lowest energy NMR-derived structure[26] with additional side-chain refinements including salt-bridges. The results are compared against NOE intensities, ^{15}N-NMR-relaxation measurements, vicinal scalar couplings, amide hydrogen exchange times, and temperature coefficients. By suitable analysis, the simulation is used to investigate nanosecond time scale dynamics and provide an enhanced view of loops and surface side chains that could not be obtained by NMR alone. The resultant dynamic picture provides a more realistic molecular view of the TSG-6 Link module and serves as a basis for understanding how this might affect its interactions with glycosaminoglycans, compared to the static and conforma-

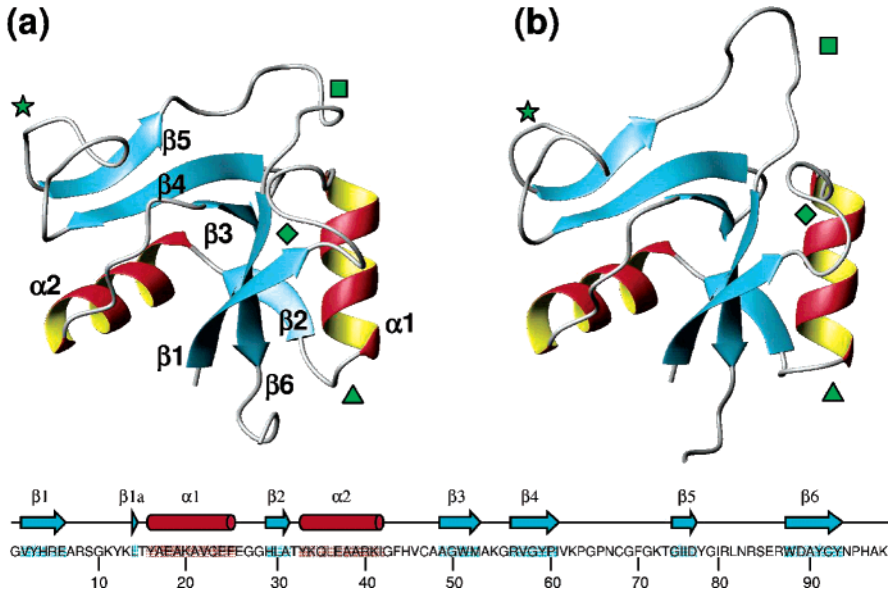


Figure 1. (a) The average solution backbone structure for Link_TSG6 of the lowest-energy 125 (of 250) minimized conformations derived from simulated annealing[26] and (b) the average backbone structure derived by averaging a 0.25 μs aqueous simulation of Link_TSG6 (125,000 frames). The following loops are marked: β1-α1 (green diamond), α1-β2 (green triangle), β4-β5 (green star), and β5-β6 (green square). In the lower portion of the figure the relationship between primary sequence and secondary structure is shown.

tionally biased views produced by XRC and NMR structure determination.

2. Materials and Methods

2.1. Molecular Dynamics Simulation. In all calculations, the CHARMM program[33] was used together with the CHARMM22 parameters for proteins.[34,35] Nonbonded interactions were truncated using the switching function between 0.8 nm and 1.0 nm, and nonbonded list generation was stopped at 1.2 nm, with the dielectric constant set to unity. Periodic boundary conditions were applied to the whole system by application of a face-centered cubic lattice transformation (the unit cell corresponds to a truncated octahedron) and using the Ewald summation convention with order set to 6, $\kappa = 0.44$ and 64 points over the box. Potential energy minimizations were performed using an implementation of the adopted-basis Newton−Raphson algorithm. The molecular dynamics simulation produced an NPT ensemble (variable volume) by application of the Nosé−Hoover algorithm with suitable parameters to maintain the pressure at 1 atm and the temperature at 298 K. The SHAKE methodology was used to keep bonds to hydrogen atoms rigid and hence permit an integration time-step of 2 fs. Numerical integration of the equations of motion applied the Leapfrog variant of the original Verlet algorithm.

Simulations were started from a low-energy conformer derived from NMR data as described previously.[26] However, slight structural refinements were made (including hydrogen bonds to side chains and intramolecular salt-bridges inferred from experiments conducted over a wide-range of pH values), and the majority charge state of each histidine residue was inferred at pH 6.0 by chemical-shift titrations (C. D. Blundell, I. D. Campbell, and A. J. Day, manuscript in preparation); H29 and H96 were made positively charged, while H4 and H45 were left uncharged. Basic (R, K) and acidic (D, E) amino acids and the N- and C-termini were given the relevant charge at pH 6.0; all other amino acids were neutral giving an overall charge of $+10e$. The starting structure was centered inside a pre-equilibrated truncated octahedral water box, which initially contained 6000 TIP3P water molecules,[36] and water molecules that overlapped with protein were deleted. Ions (19 Na^+ and 29 Cl^-) were added in a random fashion clear of the protein surface, to explicitly neutralize the charge of the protein and simulate a 0.2 M background solution of NaCl (i.e., comparable to physiological ionic strength). The protein was then fixed, and the surrounding water and ions were minimized (100 steps) and subjected to 500 ps of equilibrating dynamics at 300 K (strong coupling to heat bath and pressure). Following this, the whole system was minimized (100 steps) and equilibrated with free dynamics for a further 500 ps. The simulation was continued for a further 0.25 μs at constant pressure and temperature (by weak coupling to heat bath and pressure). Coordinates were saved at 2 ps intervals for later analysis.

2.2. Analysis of Global Protein Conformation. The 0.25 μs simulation resulted in 125 000 separate frames, containing protein, water, and ions. For the majority of analyses, these data sets were not used directly but aligned prior to analysis. In order to achieve this, first water and ions were removed, and then each frame from the simulation was translated and rotated such that the root-mean-square-deviation (RMSD) of backbone atoms (N, C′, C^α) within secondary structural elements from the NMR-derived starting structure was minimized; this set of structures will be referred to as being "in the molecular frame". A similar molecular frame calculation was performed for the 125 lowest-energy minimized structures derived from NMR data.[26] Although this is more than normally reported, none of these structures had serious steric clashes, which was not true for many of the remaining 125 higher-energy structures calculated. Average structures

for each ensemble were generated by calculating the average atomic positions in the molecular frame. Surface racer 3.0[37] was used to calculate the surface area of each coordinate set in the simulation, with a probe radius of 0.1 nm to reveal solvent accessible areas of the protein. Calculations were performed on the whole protein and repeated with the β4-β5 loop (residues V62-T73) omitted. The radius of gyration was calculated using standard equations (without mass weighting).

2.3. Prediction of Hydrogen Bonds and Comparison with Experimental Data. Hydrogen bonds were defined by distance and angular criteria, namely that the A-H...B angle deviated no more than 60° from direct and the distance between A and B did not exceed 0.35 nm (A and B are the electronegative donor and acceptor, respectively). Backbone amide hydrogen and oxygen atoms were examined for possible hydrogen bonds in each frame of the simulation, and average occupancies were calculated from this. Hydrogen bonds in side chains and salt-bridges were calculated in an analogous way.

Amide hydrogen exchange rates were calculated from the simulation using an empirical relationship,[38] which has recently been used to provide structural restraints.[39] In the EX2-limit of exchange,[40] it is assumed that both the presence of intramolecular hydrogen bonds (dominant factor) and the burial of amide hydrogen atoms in the interior of the protein (subsidiary factor) leads to protection, which slows the rate of exchange. Therefore, if N_h is the average number of hydrogen bonds that each amide hydrogen partakes in and N_c is the number of contacts made by each residue, then the protection factor, P, can be calculated by $\ln P = \beta_c N_c + \beta_h N_h$. The original values ($\beta_c = 1$ and $\beta_h = 5$) have been used unmodified here,[38] but it should be noted that other authors have tried to optimize them on the basis of molecular dynamics simulations.[17] The value of N_c was calculated by finding all residues that had a closest-approach distance less than 0.25 nm (any atom), and the value of N_h was calculated from the average number of hydrogen bonds made to and from amides during the simulation (using the definition described above).

Measurement of the amide hydrogen residence times using a H_2O/D_2O exchange NMR experiment has been described previously.[26] Temperature coefficients were obtained from ^{1}H-^{15}N HSQC spectra collected between 5 °C and 30 °C at 5 °C intervals at 750 MHz ^{1}H-frequency on a 0.3 mM sample of ^{15}N-labeled Link_TSG6 containing 0.33 mM DSS, as described previously.[41] Chemical shift deviations (CSDs) were calculated and used to improve the prediction of hydrogen bonds using a cutoff line of $-2.97 - 2.19 \times$ CSD.[42] Structure calculations were carried out (as described previously) to test hydrogen bonds for compatibility with the NMR structure.[26] These data were used in combination to determine hydrogen bond permanency.[43]

2.4. Prediction of NOESY Cross-Peak Intensities and Comparison with Experimental Data. Several NMR experiments, described and performed previously,[26] were used to obtain NOE distance constraints: a 2D ^{1}H-NOESY (unlabeled protein) in D_2O, a 3D ^{13}C-NOESY-HSQC (^{13}C,^{15}N-labeled protein) in H_2O, a 3D ^{15}N-NOESY-HSQC, and a 3D ^{15}N-HSQC-NOESY-HSQC (^{15}N-labeled protein) both in H_2O. In order to make a comparison between the simulation and this experimental data, the raw intensities of assigned peaks were measured in each spectrum.[26] In total, there were 188, 799, 859, and 32 assigned peaks from the spectra listed above, respectively, i.e., 1878 nonduplicated intensities. The theoretical prediction of the intensity used the independent spin-pair approximation in the slow tumbling regime.[44] In this case, the theoretical intensity is $k\langle r^{-3}\rangle^2$ where k is an experimentally derived constant (dependent on each particular experimental setup), r is the distance between protons, and the angular brackets represent an average over the simulation. The value of k was calculated for each of the NOESY-based experiments, and $\log I$ (I is the experimentally measured intensity) was plotted against $\log(k\langle r^{-3}\rangle^2)$ for each set of assigned peaks. In the case of degenerate methyl or methylene groups, $\langle r^{-3}\rangle^2$ was calculated between all combinations of proton pairs and summed.

2.5. Prediction of ^{15}N-Relaxation and Comparison with Experimental Data. Angular autocorrelation functions[45] of N−H-vectors were calculated at each amide directly from the molecular dynamics simulation. Although their decays were exponential, the overall correlation time estimated from the simulation was ∼2.4 ns, i.e., lower than that obtained from experimental observations (∼4.5 ns), see below. It was therefore concluded that the simulations were not capable of providing a realistic description of the overall molecular tumbling. Therefore, the angular autocorrelation functions were calculated in the molecular frame (see above), which provides an approximation to their internal correlation functions, $C^i(t)$, as described in the Lipari-Szabo model-free method for describing NMR relaxation.[45] These functions have a value of 1.0 at time $t = 0$ and are theorized to decay to an asymptotic value $S^2 \leq 1.0$, where S^2 is the order parameter for internal motions. The correlation functions were calculated out to 1 ns (500 points at 2 ps intervals), and the value of S^2 was taken at this endpoint. In order to calculate the ^{15}N-relaxation parameters at each residue, their internal motion autocorrelation functions were extended to 2^{22} points by setting $C^i(t) = S^2$ for $t > 1$ ns. Each correlation function was multiplied by $0.2\exp(-t/\tau_m)$ at every point, where τ_m is the overall correlation time. The resultant data sets were then converted to spectral density functions, $J(\omega)$, by performing a real fast Fourier transform. These were used to calculate $J(\omega)$ for specific values of angular frequency, ω, which were then input into the standard equations used to calculate ^{15}N-relaxation parameters, i.e., T_1-relaxation rates and steady-state NOE enhancement, η.[12,46] In these calculations, the N−H-bond length was assumed to be 0.102 nm, the chemical shift anisotropy was set to −160 ppm, and τ_m was set to 4.5 ns in accordance with the experimental data. In this manner, the values were predicted from the simulation at each nitrogen in the protein backbone. The calculations were repeated for the N−C$^\alpha$-bond with the same parameters, except for the overall correlation time, which needed to be increased to 5.5 ns to agree with the experimental data (this change may seem unusual at first sight, but it should be remembered that τ_m is part of the Lipari-Szabo model and is not a well-defined physical parameter in a dynamic

molecular system because internal motion and overall tumbling cannot be separated). The ^{15}N-T_1, ^{15}N-T_2, and ^{1}H-^{15}N steady-state NOE (η) data were acquired in a similar manner to that described previously[29] at a ^{1}H frequency of 500 MHz on 0.3 mM and 2 mM samples, respectively, of ^{15}N-labeled Link_TSG6 at pH 6.0, 25 °C, in 10% (v/v) D_2O. Lipari-Szabo generalized order parameters were calculated from the relaxation data using standard relaxation equations that assume a single overall correlation time and an exchange contribution.[12,45]

2.6. Prediction of Scalar Couplings and Comparison with Experimental Data. The backbone ϕ-dihedral angles were extracted from the simulation at all residues, except for proline and glycine. From these values the vicinal couplings were calculated at each point using the 'zero motion' Karplus equation: $^{3}J_{\mathrm{HNH}\alpha}(\phi) = 9.5\cos^2\phi - 1.4\cos\phi + 0.3$, from which the predicted value was computed by averaging across all time frames.[47] A ^{15}N-HMQC-*J* spectrum was recorded on a sample of ^{15}N-Link_TSG6 (2 mM, pH 6.0, 10% v/v D_2O) at 500 MHz for the direct measurements of $^{3}J_{\mathrm{HNH}\alpha}$ coupling constants.

3. Results

3.1. The Global Secondary Structure. The 125 lowest-energy minimized structures of Link_TSG6, derived from previous NMR measurements,[26] were averaged in the molecular frame (see Methods). Figure 1(a) shows the resulting conformation using traditional secondary structural visualization of the backbone, which is 0.1 nm RMSD from the lowest energy NMR structure. A similar average conformation, calculated from the 0.25 μs molecular dynamics simulation, is shown beside it in Figure 1(b). The Link module of TSG-6 contains two α-helices (α1, residues Y16-F25 and α2, residues Y33-I42) and two antiparallel β-sheets.[24,26] Sheet I comprises strands β1 (V2-E6), β1a (L14), β2 (H29-A31), and β6 (D89-Y93), and sheet II comprises strands β3 (A49-M52), β4 (R56-I61), and β5 (G74-D77). Comparison of the average structures, Figure 1(a,b), immediately shows that the simulation has maintained all the secondary structural elements of the molecule. Furthermore, the average backbone conformations of most of the loops are similar in the two average structures; the major exception is the β5-β6 loop (green star in Figure 1), which deviates significantly from the average NMR backbone conformation. Also, a slight deviation was noticed in the β1-α1 loop (green square in Figure 1), which is in steric contact with the β5-β6 loop in the NMR structure. Therefore, it should be expected that deviations may occur in experimental predictions from the simulation at these positions.

These findings can be explored in more detail by examining the range of conformations present during the simulation and in the NMR-derived ensemble. Figure 2(a) shows an overlay in the molecular frame of the 125 lowest-energy minimized structures derived from NMR data (of 250), in which the C-terminal tail, the β1-α1 loop, the β4-β5 loop, and the β5-β6 loop are not as well-defined as the other parts of the molecule. This can be compared with the output from the simulation, Figure 2(b), in which 125 structures (taken at 2 ns intervals) have been overlaid. It is apparent that the simulation ensemble is a more disordered set of structures than the NMR ensemble, Figure 2(a), even in the secondary structure elements. Of particular note are the β1-α1, α2-β3, and β5-β6 loops, which are substantially more disordered in the simulation ensemble. However, the β4-β5 loop (that forms one side of the hyaluronan-binding groove) has a similar range of conformations in both the simulation and the NMR-derived ensemble.

The overall change in conformation of the protein during the simulation can be visualized by plotting the minimum RMSD between each simulated backbone conformation and the starting structure, Figure 3(a). It can be seen that the simulation started at around 0.2 nm RMSD from the starting structure (after equilibration) and relaxes over the first 50 ns of the simulation. Comparison to a similar RMSD calculation with the β5-β6 loop omitted, Figure 3(a), shows that the structural relaxation is predominantly due to movement and partial unfolding of this loop. Following this, the simulation reaches an equilibrium, being on average 0.3 nm RMSD from the NMR structure.

3.2. Intramolecular Hydrogen Bonds. The decrease in the exchange rate incurred by folding from the free polypeptide chain and intramolecular hydrogen bonding can be described by an amide-hydrogen protection factor, *P*. Therefore, the NMR and simulation ensembles were used to estimate the value of ln(*P*) at each amino acid, as plotted on Figure 3(b). These values can be compared against the experimentally observed amide-hydrogen persistence for a protonated protein placed into D_2O.[26] However, it should be pointed out that the protection factor does not take into account differences in intrinsic exchange rate between amino acids, and the comparison will remain qualitative. At this level of analysis, a key difference between the predictions made by the NMR and simulation ensembles is in the region of Y16. The lower protections factors predicted by the simulation in this region (as compared with the alpha-helical regions) are in better qualitative agreement with the experimental data than the NMR ensemble, which predicts relatively large protection factors. Furthermore, within the majority of the secondary structural elements (i.e., 30 amides) the agreement between experiment and theoretical prediction from the simulation is good. Notable exceptions are Y3H^{N}, T32H^{N}, A53H^{N}, Y59H^{N}, I61H^{N}, I76H^{N}, Y78H^{N}, and Y91H^{N}, which are denoted with arrows on Figure 3(b). Several of those discrepancies can be accounted for by the NMR experimental parameters, in which data sets could only be collected every 105 min, and with limited resolution in the indirect dimension. For example, although Y3, T32, I76, Y59, and Y78 HN-resonances could not been seen in the first spectrum of the free protein (>105 min after reconstitution with D_2O), they were seen in the same experiment on the hyaluronan-bound protein, where global exchange is ~10 times slower.[26] However, the involvement of these amides in hydrogen bonds is also consistent with the observed NOE pattern in the free protein. Furthermore, temperature coefficient data are consistent with all of these discrepancies (see the Supporting Information). In particular, analysis suggests that several hydrogen bonds are present in the region A53-E86 that were not determined by amide exchange experi-

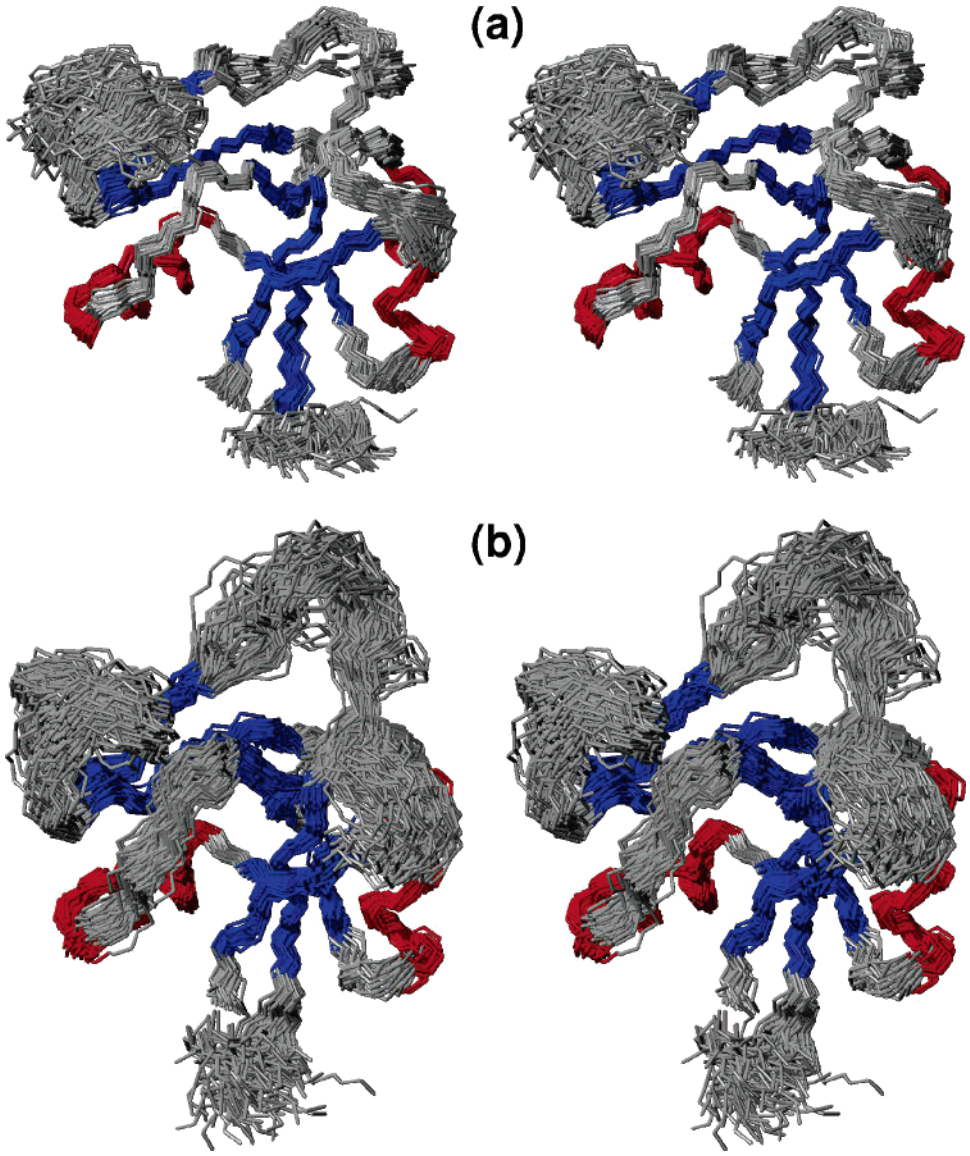


Figure 2. (a) Stereo overlays of 125 (of 250) minimized structures derived from simulated annealing and (b) from equally spaced (2 ns) conformations extracted from a 0.25 μs simulation of Link_TSG6. Both structures are colored according to secondary structure (red α-helices and blue β-strands).

ments, which includes the β4, β5 strands and the β4-β5 loop. The other discrepancies (I61HN, A53HN, and Y91HN) can be accounted for by assignment problems or overlap with other resonances. A single slowly exchanging hydroxyl hydrogen, on T32, was also observed in NMR spectra of the free protein,[26] and the NMR structures do not provide any evidence for a hydrogen bond in spite of the hydroxyl group being very well defined. Such slowly exchanging hydrogen atoms have been seen in less than 5% of reported threonine assignments (according to BioMagResBank). The simulations suggested that this hydroxyl group was in close contact with solvent and not involved in any substantial intramolecular hydrogen bonding interactions, so also does not provide an explanation.

Hydrogen bonds were characterized according to their temporal occupancy during the 0.25 μs simulation using criteria based on angle and distance (see Methods). Those hydrogen bonds in the backbone with a fractional occupancy greater than 50% (considered to be strong hydrogen bonds) are detailed in Table 1. As expected, high occupancy hydrogen bonds were restricted to the secondary structure elements, Figure 4. The hydrogen bonds found in the simulations are marked with dotted lines and labeled with their percentage occupancy in Figure 4.

Within both α-helices (Figure 4) the hydrogen bond pattern predicted by the simulation was as expected, with the characteristic $i + 4 \rightarrow i$ H^N...O interactions (from hereafter hydrogen bonds will be represented by "...") and fractional occupancies in excess of 70%. In both helices, it was found that hydrogen-bonding interactions from $i+3 \rightarrow i$ H^N...O (as found in 3_{10}-helices) could coexist with those normally associated with α-helices but with a lower occupancy (~25%). At the N-terminal end of both α-helices the backbone hydrogen bonds were consistent with those inferred from experimental data,[26] see Figure 4. The two hydrogen bonds forming the α1-helix cap (not shown in Figure 4) were observed in the simulation: T15H^N...Q18$O^{\epsilon 1}$ (26%) and Q18H^N...T15$O^{\gamma 1}$ (25%). However, they were found to have

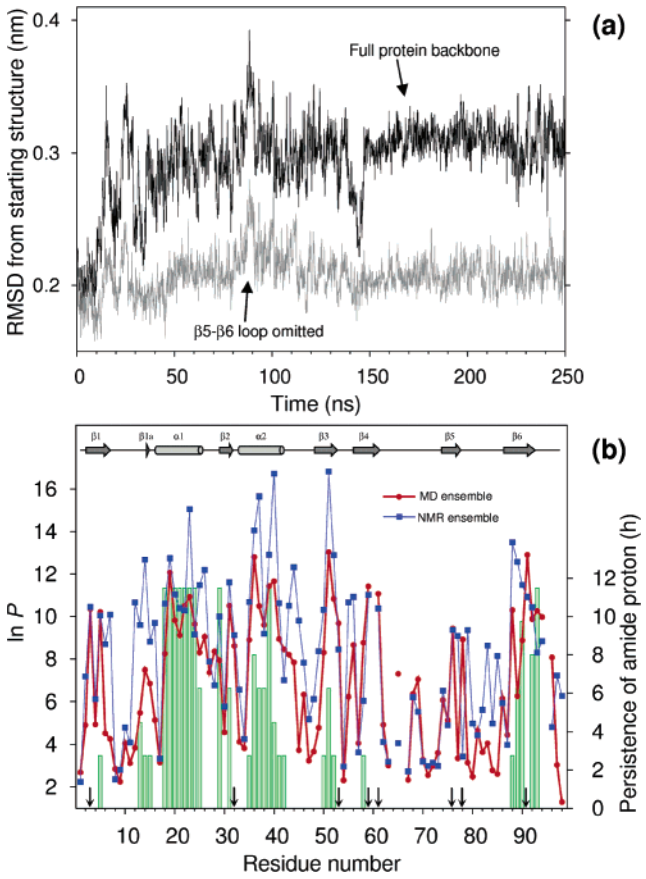


Figure 3. (a) RMSD of the backbone between the simulation and the NMR-derived starting structure (as a function of time) with and without consideration of the β5-β6 loop in the sums. (b) Calculated values for the amide protection factors in Link_TSG6 derived from a 0.25 μs simulation (red line and circles) and from the NMR ensemble (blue line and squares). The green bars indicate the latest time at which each amide could be observed in a H_2O/D_2O exchange experiment. Arrows denote obvious inconsistencies between the theoretically predicted protection factors and experimental measurement.

a relatively low persistence in the simulation and another hydrogen bond was identified, T15H$^{\gamma 1}$...Q18O$^{\epsilon 1}$ (26%), which has not been confirmed experimentally. At the N-terminus of the α2-helix T32H^{N}...Q35O$^{\epsilon}$ (81%) and Q35H^{N}...T32O$^{\gamma}$ (62%) were found to cap it, with a rather large persistence. At their C-termini the α-helices were predicted to have more hydrogen bonds than determined previously by NMR.[26] First, G27 and G43 were inferred to be involved in hydrogen bonds (to C23 and A39, respectively) in the simulation, and while a slowly exchanging amide has not been identified experimentally here, amide temperature coefficients support their existence (see supplementary Table 1). Also, at the C-terminus of both helices a capping hydrogen bond was found from $i + 5 \rightarrow i$ (as found in π-helices), again not seen experimentally. This was of high frequency in α1 (G28H^{N}...C23O) at 92% and lower in α2 (F44H^{N}...A39O) at 50% (see Figure 4). Again, temperature coefficient data indicate that they are likely to be present in the free protein.

The simulations reproduced the majority of the hydrogen bonds in β-sheets, as shown in Figure 4, in the sense that hydrogen bonds had a high-percentage occupancy where slowly exchanging amides were observed experimentally.[26] The two obvious exceptions to this agreement were D89H^{N}...A49O and L14H^{N}...W88O, which were identified as slowly exchanging amides, but have persistencies of only 11% and 28%, respectively, in the simulations. Furthermore, the simulation indicated the presence of more persistent hydrogen bonds across the strands than had been reported previously.[26] In particular, the β1-β6, β3-β6, β3-β4, β4-β3, and β4-β5 strand pairs were all suggested to have one or two persistent hydrogen bonds at positions that had not been deduced previously by amide exchange data (Figure 4 and Table 1), e.g., Y78H^{N}...V57O and G65H^{N}...G69O. Such extensions are consistent with recent temperature coefficient measurements and with structure building exercises (see the Supporting Information, Table 1). Also, of particular note is a tertiary structural hydrogen bond between A53H^{N} and A31O (99%), which has not been proposed previously (not shown in Figure 4). Temperature coefficient measurements were supportive of this observation.

The strongest salt-bridges (present for greater than 80% of the time) were found between R56...D77 (end of β4 to β5), R81...E86 (within the β5-β6 loop), and R5...E26 (end of β1 to end of α1). These were present in the starting structure and have been maintained during simulation. Weaker salt-bridges (15−50%) were predicted between K13...E18, K20...E24, K34...E37, R8...E26, R87...E18, and H29...E24. The residue K11 formed a bifurcated salt-bridge to both E6 and D89.

3.3. Comparison with Experimental NOESY Data. As noted above, the simulations compared favorably with previous characterizations of secondary structure outside of the β5-β6 loop. Comparison with experimental NMR intensity data was therefore used to further test the simulations. Figure 5(a) shows the log$I$ values plotted against the predictions, log(kr^{-6}), for the NMR-derived starting structure. Figure 5(b) shows (using the same axes ranges) the same data averaged over the simulation; in this case the predictions are log($k<r^{-3}>^{2}$). It can be seen that, overall, the correlation between the theoretical predictions and experimental measurements is good in both cases. However, the simulation produced some outliers, Figure 5(c), which were not present in the original NMR-derived structure. These data are shown again, relative to the protein sequence, Figure 5(d), where the absolute difference between the experimental and theoretical log values have been plotted. The regions where disagreement is present are shown using bars and localize to residues 10−18, 50−58, and 80−86. These NOEs are between residues in the β1-α1, β3-β4, and β5-β6 loops. It is inferred that this is principally due to large motions of the β5-β6 loop in the simulation (which has experimental NOE cross-peaks to protons in the other two loops). The RMS deviation between log$I$ and log(kr^{-6}) is 0.584 for the lowest-energy NMR-derived structure. For the simulation if all data are taken into consideration, the deviation is 0.831, and if the 42 outliers (out of 1878) are taken out, then the deviation becomes 0.588, which is not significantly different from the starting structure. In other words, the dynamic representation of the protein given by the simulation is as

Table 1. Hydrogen Bonds between Backbone Amide Hydrogen (H^N) and Carbonyl (O) Present for More than 50% of Frames in the Simulation

H^N donor	O acceptor	% occupancy	secondary structural element	H^N donor	O acceptor	% occupancy	secondary structural element
G1	N94	82.9	β1-β6 strand	I42[a]	A38	72.2	α2 helix
Y3[a]	C92	99.7	β1-β6 strand	G43	A39	98.2	α2 helix
R5[a]	A90	98.9	β1-β6 strand	F44	A39	51.0	α2 helix
A19[a]	T15	99.8	α1 helix	G50[a]	G58	74.8	β3-β4 strand
K20[a]	Y16	94.5	α1 helix	W51[a]	D89	99.9	β3-β6 strand
A21[a]	A17	78.3	α1 helix	M52[a]	R56	98.9	β3-β4 strand
V22[a]	E18	96.1	α1 helix	A53	A31	99.2	tertiary
C23[a]	A19	82.5	α1 helix	R56	M52	83.0	β3-β4 strand
E24[a]	K20	92.0	α1 helix	G58[a]	G50	83.6	β3-β4 strand
F25[a]	A21	90.1	α1 helix	Y59[a]	I76	97.5	β4-β5 strand
E26	V22	82.7	α1 helix	I61	G74	88.0	β4-β5 strand
G27	C23	77.1	α1 helix	G65	G69	80.0	β4-β5 loop
G28	C23	92.3	α1 helix	G69	G65	62.3	β4-β5 loop
H29[a]	Y93	75.8	β2-b6 strand	I76[a]	Y59	98.6	β4-β5 strand
A31[a]	Y91	99.2	β2-b6 strand	Y78	V57	83.2	β4-β5 strand
L36[a]	T32	92.9	α2 helix	W88[a]	L14	80.2	β1a-β6 strand
E37[a]	Y33	98.7	α2 helix	A90[a]	R5	80.4	β1-β6 strand
A38[a]	K34	93.3	α2 helix	Y91	Y51	83.9	β3-β6 strand
A39[a]	Q35	96.0	α2 helix	C92[a]	Y3	94.6	β1-β6 strand
R40[a]	L36	97.9	α2 helix	Y93[a]	H29	99.9	β2-β6 strand
K41[a]	E37	76.9	α2 helix	N94	G1	98.8	β1-β6 strand

[a] Experimentally determined slowly exchanging amide hydrogen.

good a fit to the experimental data as the static representation from simulated annealing (except in the β5-β6 loop). Furthermore, recent work has shown that a physically correct ensemble also correctly predicts the nonobserved NOEs.[48] Such analysis is possible using the aromatic region (F, H, Y, and W residues) of the 2D-NOESY spectrum recorded in D_2O, which is the least ambiguous part of any of the NOESY spectra collected. In this region, 36 NOEs were predicted by both the lowest-energy NMR structure and the simulation ensemble (<5 Å) that were not assigned experimentally (it should be stressed that some of these 'missing assignments' may be due to artefacts or overlap). However, a further 68 NOEs were predicted by the NMR structure alone and 56 by the simulation alone that were not assigned experimentally. Therefore, in fact, the simulation predicts fewer nonassigned/nonobserved NOEs than the static lowest-energy NMR structure.

For direct comparison with the NMR spectra, the average distance ($<r^3>^{-1/3}$) was calculated from the simulation for all assigned NOEs in the 2D ^{1}H-NOESY in D_2O. It was assumed that if an NOE could be observed between protons in an NMR spectra, their average distance must be less than 0.5 nm (as is typically done). Using chemical shift tables, cross-peaks were labeled directly onto the raw NMR spectrum. A section of the 2D NOESY in the aromatic-fingerprint region of the spectra is shown in Figure 6. Predictions that are within 0.5 nm are labeled in blue, while those in which the prediction was greater than 0.5 nm are labeled in red; this part of the spectrum was chosen to show both the agreement with the experimental data and also some disagreement. Out of 216 peaks assigned in the 2D NOESY, the simulation predicted 185 to be within 0.5 nm (86%). The static starting structure placed 195 within the same distance (90%).

3.4. Comparison with ^{15}N-Relaxation Data. While both a static structure and a dynamic simulation can predict the average NOEs reasonably well, only the simulation is capable of predicting parameters that are highly dependent on dynamics. One such set of observables can be obtained from ^{15}N-relaxation NMR experiments. Theoretically, NMR dipole−dipole relaxation (as discussed here) is related to angular correlation functions calculated from vectors joining the relaxing nuclei.[45,49] First, autocorrelation of the N−H-vectors on long timescales directly from the aqueous simulation resulted in an overall rotational correlation time of 2.4 ns that was too rapid for a molecule of this size.[50] The Link_TSG6 molecule diffused with axial symmetry, but the diffusion tensor was found to have a slight anisotropy ($D_{\parallel/\perp}$) of 1.13.

Therefore, rather than computing N−H-bond-vector autocorrection functions for the freely rotating protein, they were calculated in the molecular frame (removing overall tumbling). In this frame the correlation functions did not decay to zero but to a constant value that describes the Lipari-Szabo order parameter (S^2), see Figure 7(a). Comparison with the experimentally derived values gave a correlation coefficient of 0.38 and an RMSD of 0.14 (see the Supporting Information). The internal correlation functions were multiplied by an exponential representing overall tumbling, and the resultant data set was Fourier-transformed to produce a spectral density function. From this, the ^{1}H-^{15}N heteronuclear steady-state NOE enhancement (η) and T_1-values were calculated at residue specific positions. These were compared

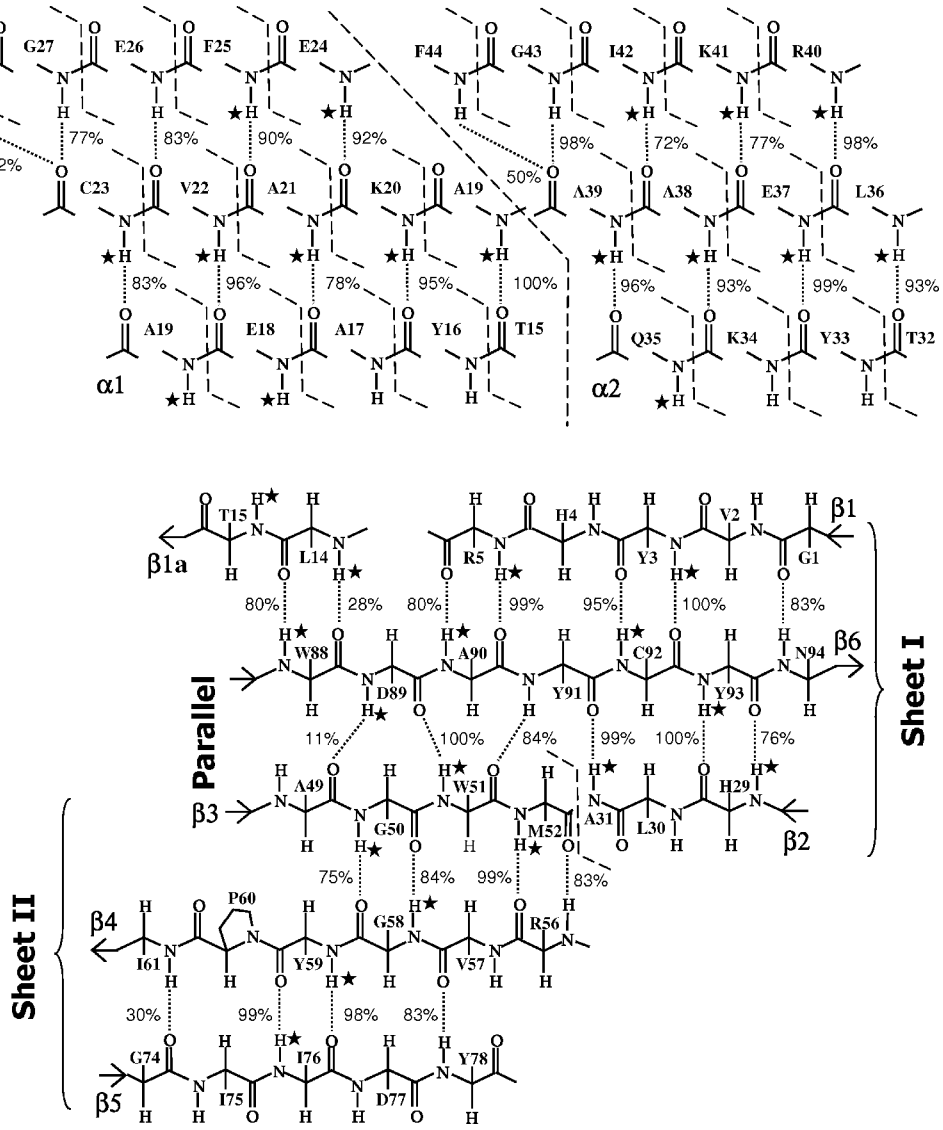


Figure 4. Hydrogen bonds within the secondary structure elements of Link_TSG6. Experimentally determined slowly exchanging amides[26] are marked with stars. Hydrogen bonds found in the simulation (with a fractional occupancy greater than 50%) are marked by a dotted line and labeled with their percentage occupancy.

against actual NMR measurements made on ^{15}N-labeled protein. The best fit to the overall correlation time was 4.5 ns, leading to the comparison shown in Figure 7(b,c). For the η- and T_1-values the respective correlation coefficients to experimental data were 0.74 and 0.15, and the respective RMSDs were 0.16 and 0.44 s^{-1} (see the Supporting Information).

As found previously in a simulation of ubiquitin,[15] some of the residues had T_1-values that were significantly larger and η-values that were significantly lower than the experimental data. This was primarily due to rapid local motion (on the picosecond time scale), rather than long time scale flexibility, perhaps due to the libration of the peptide plane. To investigate this, the calculation was repeated but this time using the N−C$^\alpha$-bond rather than the N−H-bond, i.e., a bond located along the axis of motion of the peptide plane. Due to the increased order parameters resultant from such a calculation, Figure 7(a), the overall correlation time that fitted the experimental data had to be increased to 5.5 ns. The correlation coefficient to experimental data increased to 0.57, and the RMSD decreased to 0.11 (see the Supporting Information). Calculations using these values are shown using thick dotted lines in Figure 7(b,c). It can be seen that the fit is much closer to the experimental data than calculations using the N−H-bond. For the η- and T_1-values the respective correlation coefficients to experimental data were 0.74 and 0.28, and the respective RMSDs were 0.09 and 0.26 s^{-1} (see the Supporting Information). A disagreement between theory and experiment is still present around residue 80, corresponding to the β5-β6 loop where most of the NOE violations were found.

3.5. Comparison with Scalar Coupling Data. The average conformation of the backbone can be assessed using vicinal scalar-couplings measured from NMR spectra. One coupling that can be measured routinely at most residues is the vicinal coupling ($^3J_{HNH\alpha}$) between H^N and H^α, which has a high value in β-sheets (∼10 Hz) and a low value in α-helices (∼4 Hz). Figure 8(a) shows that the couplings predicted from the simulation ensemble are in good agreement with experimental measurements (from an HMQC-*J*

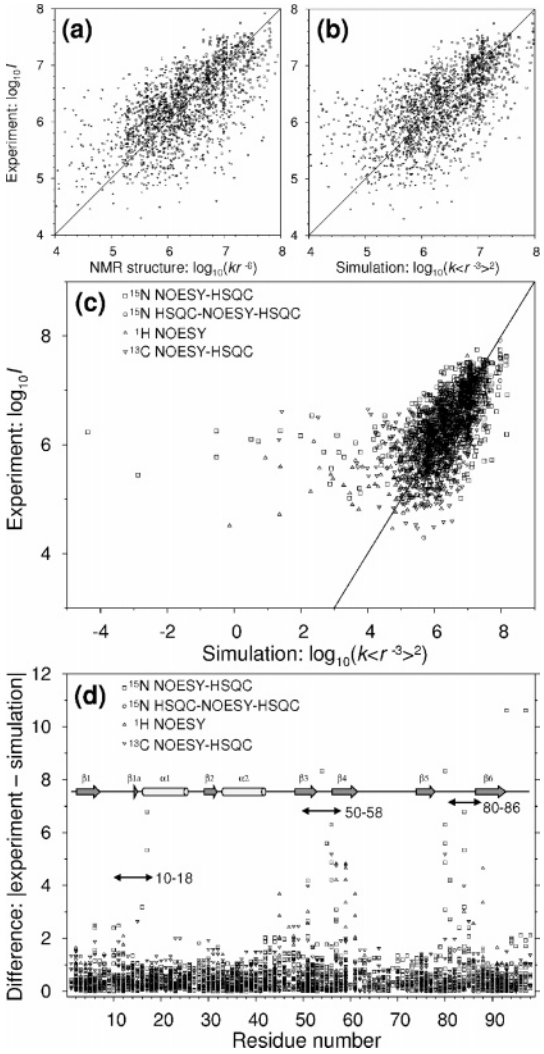


Figure 5. Correlation between the calibrated experimental NOESY measurements (log intensity, *I*) from four NMR data sets[26] and (a) the original lowest energy NMR structure and (b) the 0.25 μs simulation performed here. (c) Shows this correlation to the simulation for all of the predicted NOEs and (d) shows the difference between the log values along the backbone of the protein. Regions of disagreement are marked with double arrows.

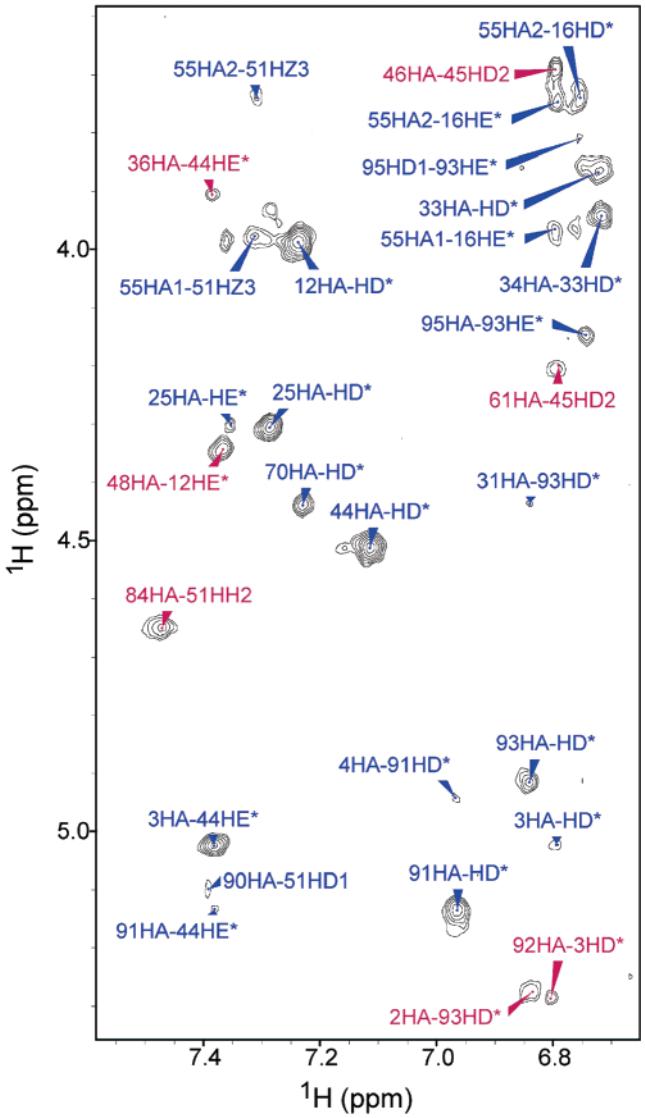


Figure 6. Part of the aromatic fingerprint region extracted from the ^{1}H-^{1}H NOESY experiment performed in D_2O on Link_TSG6, showing the NOEs that are, on average, predicted to be within 0.5 nm in the simulation (blue) and those that are not (red). Each peak is labeled with its NMR assignment. This region was purposely selected as one with more disagreements than average.

experiment). In fact, the agreement with the simulation is better than the lowest energy NMR structure. The simulation ensemble and NMR structure have correlation coefficients to experimental data of 0.70 and 0.60, respectively, while the RMSDs to experimental data are 1.48 and 1.91, respectively (see the Supporting Information).

Pairs of (ϕ, ψ) values were extracted from the simulation, and scatter plots were constructed for every peptide linkage in the protein (see the Supporting Information). It was found that many of the linkages have multiple regions of exploration and, in particular, those N-terminal to glycine residues. For those linkages that had a single region of exploration (monomodal), an average and standard deviation can be meaningfully expressed. Most of the averages are restricted to the right-handed α-helix and β-region of the Ramachandran plot as expected (see the Supporting Information).

3.6. The Dynamic Structure of the Backbone and Side Chains. Having shown that the simulation predicts the diverse range of experimental data as well as the static ensemble of NMR-structures (except at the β5-β6 loop) it can be used to inform us about nanosecond time scale dynamics in the protein backbone and side chains. Previously, it was shown that upon hyaluronan-binding, the β4-β5 loop opens under conformational rearrangements involving the C47−C68 disulfide bridge. It is therefore pertinent to examine further whether the free protein momentarily adopts such a conformation (which would substantially increase the protein surface accessible area). Figure 8(b) shows that the initial protein surface area (60 nm^2) decreases momentarily (to 52 nm^2) before rising again. Following this, the protein undergoes an oscillatory motion with the surface area varying between 53 nm^2 and 62 nm^2, over a period of 150 ns. Recalculation of the accessible surface area with the β4-β5 loop omitted indicated that this oscillatory motion did

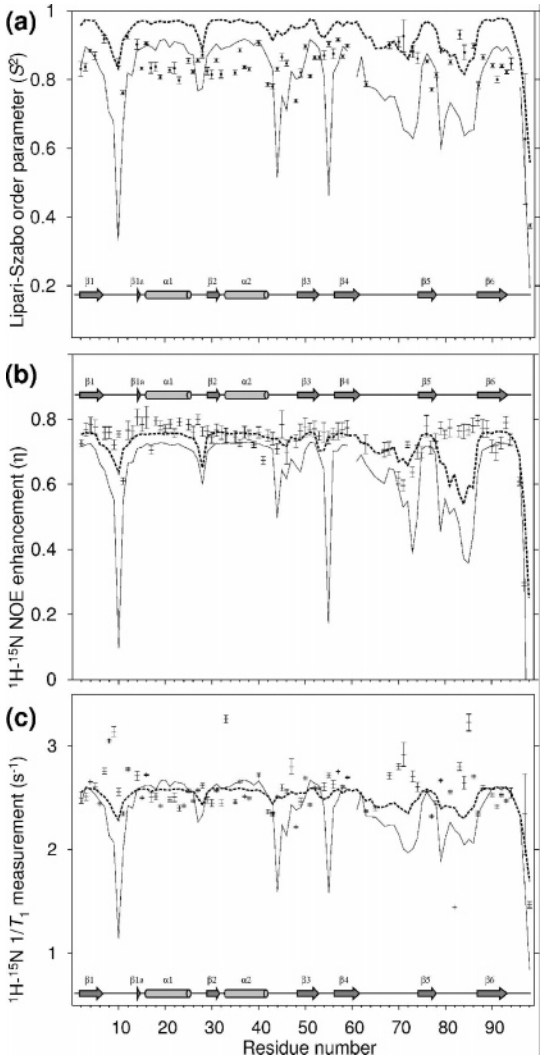


Figure 7. (a) Calculated Lipari-Szabo order parameters from the simulation of Link_TSG6. The thin, continuous line represents order parameters calculated using the NH-vector and the thick, dotted line from the N−C$^\alpha$-vector. (b,c) ^{15}N-relaxation parameters calculated from correlation functions calculated in the molecular frame and an overall tumbling time of 4.5 ns and 5.5 ns for the NH− and N−C$^\alpha$-bonds, respectively. The experimental data are also plotted, using error bars.

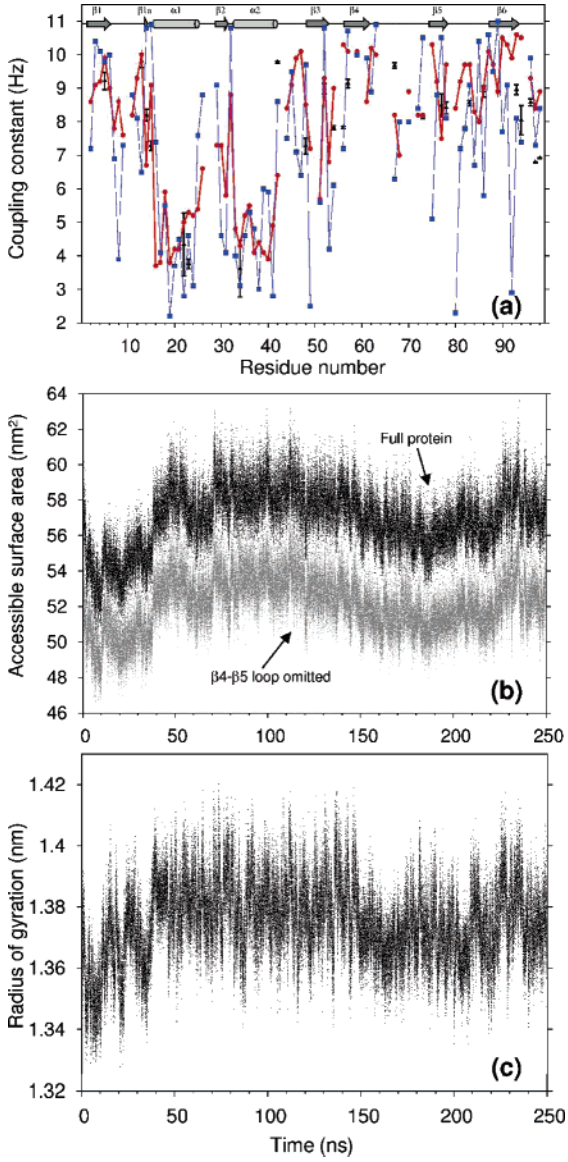


Figure 8. (a) Comparison between experimentally determined $^3J_{H\alpha HN}$ (using error bars to one standard deviation) and the values calculated from the simulation (blue squares and thin lines) and from the lowest energy NMR-derived structure (red circles and thick lines). (b) Calculated surface area for the whole protein as a function of simulation time (black) and for the protein with the β4-β5 loop omitted (gray). (c) Radius of gyration as a function of time for the whole protein (without mass weighting).

not come from mobility of this loop and the momentary exposure of a binding groove. Further, the overall trend in the radius of gyration (calculated without mass weighting), shown in Figure 8(c), was similar to that seen in accessible surface area. Therefore, it is suggested (since the radius of gyration provides a view of global protein reorganization) that this trend in surface area simply comes from a slow overall breathing motion of the protein on the nanosecond time scale.

The flexibility of various parts of the protein simulation was quantified by calculating the RMSF (root-mean-square fluctuation) for N−H-groups in the molecular frame, Figure 9(a). In the majority of cases, the amide hydrogen atoms have more flexibility than their neighboring nitrogen atoms during the simulation because the peptide planes undergo local librational motion; this may account for the discrepancies between the calculated ^{15}N-relaxation measurements.

Amide groups not conforming to this general relationship occur exclusively around R81 and localize to regions of the protein simulation that disagreed with the experimental NOE measurements. The simulation predicts that the major points of flexibility of the amide-hydrogen atoms are K98, G10, N67, and F25 (RMSFs ∼0.4 nm). Residues G10 and K98 have low ^{1}H-^{15}N NOE values, Figure 7(a), indicative of flexibility, and the side chain of N67 has previously been shown to be flexible.[51] The simulations also suggest that this flexibility is present in the adjacent nitrogen atoms, except at F25, which is one of the least flexible points of the backbone. Therefore, as expected, there is a good correlation between amide hydrogen and nitrogen RMSFs in the

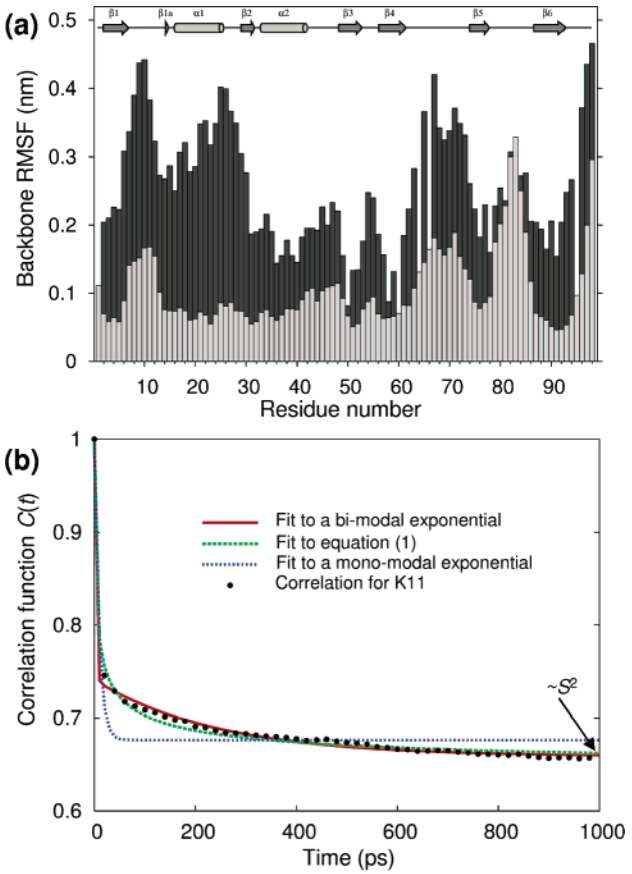


Figure 9. (a) Histogram of the root-mean-square fluctuation (RMSF) of the backbone H^N (black bars) and N (gray bars) atoms at specific residue positions. (b) Autocorrelation function for the N−H vector of K11 fitted to mono- and bimodal exponential and the equation described in eq 1 of the text.

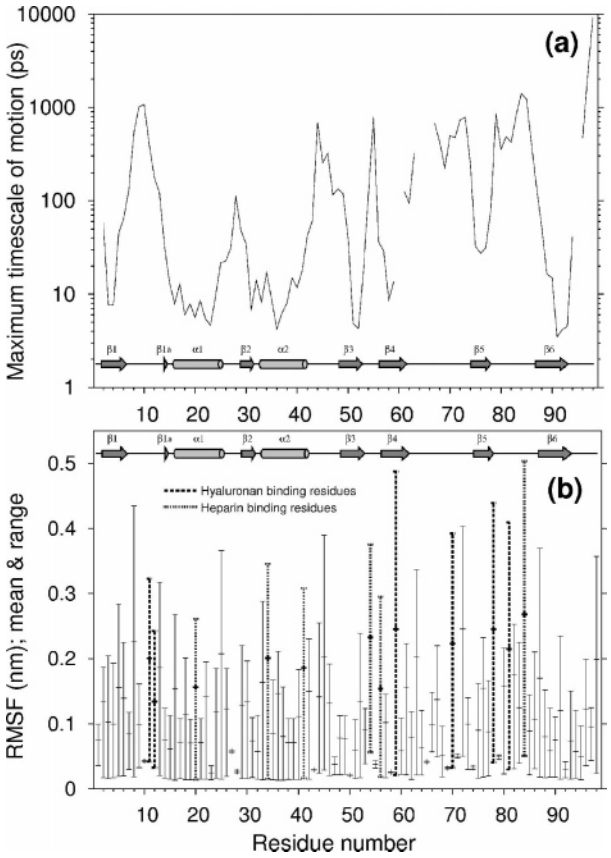


Figure 10. (a) The maximum time scale of internal motions at residue-specific positions calculated by fitting an empirical equation (see text) to the molecular frame correlation functions. (b) The minimum, maximum, and average RMSF of side chains are shown with key hyaluronan and heparin binding residues indicated. Note: K20, K34, K41, and K54 have been shown experimentally to bind heparin; R56 and R84 were predicted to be involved, but folded mutants could not be obtained for these residues.

simulation. From Figure 9(a), G50 and G58 have the most ordered amide hydrogen atoms. These residues are located in the core of the protein and form hydrogen bonds to each others' carbonyl oxygen atoms across the β3- and β4-strands, Figure 4. However, in the calculations of molecular frame order parameters (S^2) described above, Figure 7(a), these residues had S^2-values that were indistinguishable from residues in the other secondary structural elements. Furthermore, K98, G10, G55, and F44 had low calculated S^2 values, but the RMSFs of K98 and G10 are high (∼0.4 nm), whereas, for G55 and F44, the RMSFs are lower (∼0.2 nm). Therefore, there does not appear to be a relationship between the molecular frame order parameter (S^2) and amide-hydrogen RMSF.

The angular autocorrelation functions for N−H-vectors contain information on the time scale of motion of the backbone, i.e., the faster the decay, the faster the motion. It was found that calculated correlation functions ($C^{(i)}$) in the molecular frame did not fit very well to an exponential form as postulated originally by Lipari and Szabo.[45] Rather, interpretation of the correlation functions as an exponential resulted in ambiguities in establishing the decay constant; this may be due to the fact that it is theoretically impossible to define a consistent molecular frame for a dynamic molecule. However, the empirical eq 1 was found to fit all of the decay curves well, and a meaningful decay constant (k) could be extracted to estimate the time scale of motion, Figure 9(b).

$$C^{(i)}(t) = S^2 + \frac{1 - S^2}{1 + (t/k)^{1/2}} \qquad (1)$$

For time $t = 361k$ the correlation function has decayed almost fully to $0.05(1 + 19S^2) \approx S^2$. Hence, this is a time parameter that represents the longest time scale correlations found at each amide hydrogen, and these are plotted for each residue in Figure 10(a). It is interesting to note that the backbone amide groups within the secondary structure elements are predicted to undergo rather rapid motion (picosecond time scale), whereas those in loops undergo motion on the nanosecond time scale. The tail is predicted to undergo motion on an even longer time scale.

3.7. Side-Chain Dynamics of Key Binding Residues. Figure 10(b) shows the average, maximum, and minimum RMSF found within each side chain in Link_TSG6; key residues that bind to hyaluronan and heparin are indicated. Residues involved in binding to glycosaminoglycans are particularly flexible in the free protein and particularly so at their ends, where they would make contact with ligand. One

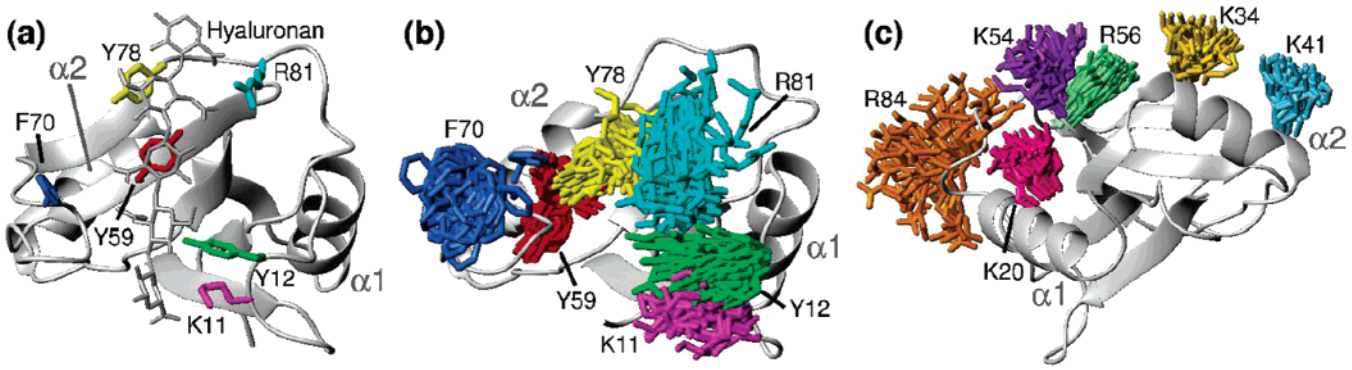


Figure 11. (a) The hyaluronan-bound NMR-structure of Link_TSG6 with a hyaluronan oligosaccharide modeled into the binding groove, reproduced from Blundell et al., with key residue indicated. (b,c) Overlay of 50 structures from the simulation of Link_TSG6 showing (b) the hyaluronan-binding residues, aligned with the model in (a) and (c) the heparin-binding residues with the protein viewed from the side.

of the most flexible side chains in the whole molecule is Y59, which is found at the center of the hyaluronan-binding groove in the structure of the complexed protein. However, this residue appears well defined in the family of NMR structures.[26] A possible explanation for this is that a large number of NOE cross-peaks are often observed to tyrosine residues, which may result in simulated annealing overconstraining them on the surface. Furthermore, the region K54−R84, while the simulation suggests that it contains the highest density of dynamic side chains in the molecule and many hyaluronan and potential heparin binding residues, has relatively slow backbone dynamics (nanosecond time scale), which may have an important part to play in ligand specificity and binding.

Figure 11(a) shows the lowest-energy NMR-structure of hyaluronan-bound Link_TSG6,[26,32] while in Figure 11(b), the 50 conformations of the same residues from the simulation of the free protein are given for comparison. All of the residues can be seen to be particularly dynamic, and their orientation is largely projecting into solution away from the protein surface. It can also be seen clearly that residue Y78 moves closer to Y59, due to unfolding of the β5-β6 loop. In Figure 11(c), the heparin binding side chains (which are on another face of the protein) are shown. All these basic side chains are again projecting away from the protein surface, but in this case are relatively more ordered (except for R84, which is probably partly due to movement of the unfolded β5-β6 loop).

4. Discussion

The 0.25 μs simulation of Link_TSG6 performed here, while as good a fit to the experimental NOE data as static NMR-structures outside of the β5-β6 loop, has the ability to provide additional new insight into protein dynamics. Deviations from the experimental data were almost exclusively localized to the β5-β6 loop, which lost its original structure and became disorganized within the first 50 ns of simulation time, Figure 3(a). In the rest of the protein, the RMSF of backbone nitrogen atoms was less than 0.1 nm for the majority of the backbone, Figure 9(a); notable exceptions were found in the backbone around G10, R81, and N67 and at the protein C-terminus (K98). Similar observations were made in two simulations of ubiquitin, started from either NMR and XRC structures,[15] where both trajectories were found to be stable with an average RMSF for the backbone atoms of around 0.1 nm. Comparison with NMR NOESY cross-peaks from ubiquitin showed that only four restraints were violated by more than 0.05 nm in the NMR-trajectory, while the XRC trajectory had slightly more, with 13 violations (out of 985 unambiguous restraints); this is a total of 1.3% violations in ubiquitin compared to 2.2% in the simulation presented here. As we have found for the β5-β6 loop of Link_TSG6, all the ubiquitin violations were restricted to one section of the protein.

The simulation of Link_TSG6 correctly predicted high-occupancy intramolecular hydrogen bonds in secondary structure where they had been experimentally determined on the basis of amide exchange rates and NOE patterns but were predicted to extend further out from the ends of β-strands than previously reported.[26] The most occupied hydrogen bonds were localized to α-helices and toward the middle of β-sheets. A highly occupied tertiary structure hydrogen bond was predicted linking A53H^{N} (in the β3-β4 loop) to A31O (in strand β2) that has not been documented previously; this could account for the high-resolution obtained for the A53 methyl side chain during the NMR-structure determination, which was surprising given that A53 is the central residue in a β-turn. The simulations also predicted that the single slowly exchanging hydroxyl group of T32, found by NMR measurements,[26] was solvent exposed and not involved in extensive hydrogen bonds. The reason for this slow exchange, therefore, still remains a mystery, although it may indicate an ion-binding site. In a previous 3 ns simulation of bacterial cytochrome *c*,[17] intramolecular hydrogen bonds were used as a basis for calculating amide protection factors. While amide hydrogen bonds were found throughout the protein, their fractional occupancy was highest in the helices, giving them the highest amide protection factors. However, bacterial cytochrome *c* does not possess β-sheets, so they could not be considered. In the present simulation, it was found that high amide hydrogen protection can be found in both helices and sheets. Within β-sheets, the most central strands (β3, β6) have the highest protection factors, as might be expected.

The aqueous simulation of Link_TSG6 predicted a rotational tumbling time of 2.4 ns, which is lower than that calculated (4.5−5.5 ns) from ^{15}N-relaxation measurements. This compares favorably with a previous aqueous simulation of ubiquitin,[15] which had an even greater disagreement (0.74 ns as opposed to an experimental value of 4.03 ns), suggesting that the current simulation is making reasonably accurate predictions on the nanosecond time scale. Both SPC (used in the ubiquitin simulations) and TIP3P (used here) display good thermodynamic properties, but their kinetic properties differ from the experimental values, and, in particular, they severely overestimate the diffusion coefficient of water while underestimating bulk water viscosity.[52−54] These disagreements underline the fact that molecular mechanics simulations of this type, i.e., using simple models of water, cannot predict the overall tumbling rates for proteins. In spite of these disagreements, the diffusional anisotropy ($D_{\parallel\perp}$) calculated for ubiquitin was close to the experimental value of 1.17.[15] The diffusional anisotropy calculated for Link_TSG6 is similar to ubiquitin and hence may be a good predictor of the experimental value (not yet experimentally determined).

The simulation of Link_TSG6 appears to contain too much fast time scale (subnanosecond) fluctuation at the amide hydrogens, which results in a rapid initial decay of the angular correlation function and order parameters that are lower than those required to predict the ^{15}N-relaxation parameters. The flexibility of the amide hydrogen around F25, in particular, is not reflected in the RMSD of the neighboring nitrogen backbone, Figure 9(a), nor is reflected in the ^{15}N-heteronuclear NOE measurements, Figure 7(b). The apparent disagreement between the heteronuclear NMR data and back-calculations from simulations is therefore likely to be due to rapid local librations of the peptide planes. Simulations of ubiquitin exhibited similar phenomena, and the calculated order parameters deviated from the experimentally derived values. Interestingly, however, it was found that internal motions in the 2−6 ns range are not well represented by heteronuclear relaxation data because they are obscured by relaxation due to the overall tumbling time of the molecule, which is of the same order. Therefore, ^{15}N-relaxation experiments that are traditionally used to infer dynamic information in proteins may be largely insensitive to important nanosecond time scale dynamics and vastly more sensitive to picosecond time scale dynamics. Simulations therefore currently play an essential role in investigating motions on these longer nanosecond timescales that are not accessible to NMR, but it is crucial that each simulation is tested against experimental data to assess its general validity before conclusions are drawn.

Residues Y78 to W88 in the β5-β6 loop are seen to have a high flexibility in the simulation, and the mean structure deviates from the NMR-derived conformation. Comparison with experimental data suggest that this motion is not realistic. This β5-β6 loop of Link_TSG6 is surrounded by several acidic and basic residues (D77, R81, R84, E86, R87, and D89), which suggests that there are a complex set of electrostatic interactions. Since the original structure calculations may not have placed the side chains in ideal positions, this could have led to some unfavorable electrostatic interactions that destabilized the loop. The unfolding of the loop leads to Y78 moving to be in the proximity of R81, as shown in Figure 11(b). However, it should be noted that this loop is the longest in the protein (twice as long as any other) and has no secondary structure, disulfide bridges, hydrogen bonds, or salt-bridges to stabilize it. Furthermore, it is also the location of a sizable insertion in the second Link module of the hyaluronan-binding domain of the Type C members of the Link module superfamily.[32]

Furthermore, it is known that CHARMM22 does not model the α-L region of (ϕ,ψ) potential energy surface (including glycines) by significantly overestimating the relative energy, which may explain some of these observed anomalies within the loops of TSG-6.[55,56] Simulations performed using the CMAP correction to the CHARMM22 force-field may be able to more accurately model the protein and alleviate some of these problems.[35]

The hyaluronan-binding β4-β5 loop was difficult to define by NMR and simulated annealing because fewer NOEs per residue than average could be obtained[26] since it contains two proline residues (P64 and P66) and three glycine residues (G65, G69, and G71). Due to this, the β4-β5 loop had a similar flexibility in both the calculated NMR ensemble and the simulation, which is largely fortuitous. The simulation also indicated that other loops in the protein are more flexible than the NMR ensemble suggests, probably because simulated annealing results in overconstraining when there is a large amount of NOE data, i.e., it purposefully uses experimental data to generate the impression of a static structure when no such structure exists in solution. This overconstraining may be due to the fact that there are often more NOE restraints than rotatable degrees of freedom in the molecule.

The simulations, while long (0.25 μs), provided no evidence for the opening/closing motion of the hyaluronan-binding β4-β5 loop on the nanosecond time scale. This suggests that the bound conformation of this loop is not sampled in unbound protein, further indicating an induced-fit mechanism for binding of hyaluronan.[26] A second set of NMR-assignments was found for the β4-β5 loop at I61, V62, and K63, which was hypothesized to be due to cis−trans isomerization of P64 or P66.[26] It should be noted that the current simulation is not capable of investigating such long-time scale phenomena.

It was also found that the backbone of loops and the C-terminal tail were involved in slower time scale motions (nanosecond) than secondary structural elements, which were involved in faster time scale motions (picosecond). In particular, the backbone within elements α1, α2, β3, and β6 have the fastest time scale motions (less than 10 ps), which interestingly coincide with the regions in the protein that have the highest amide hydrogen exchange protection factors. In fact, it could be hypothesized that there is an inverse relationship between time scale of local dynamic motion and amide hydrogen protection factors (EX2-regime) according to this calculation.

Surface side chains are generally not well characterized in NMR structure ensembles since there are relatively few

NOE constraints to side chain nuclei. A consequence of this is that NMR ensembles generally show them in a variety of conformations, some flat against the protein surface, simply due to lack of experimental data. The most favorable configuration for many of these side chains is to be pointing away from the protein surface and highly solvent exposed, as shown in Figure 11(c). This is particularly important for understanding how heparin binds to this Link-module, which is likely to occur via the positive charge at the end of the long and flexible side chains of lysine and arginine residues. However, it should be pointed out that side chains are the most difficult part of the protein to characterize experimentally, and the predictions made for side chains by the simulation are largely untested at this stage.

In the NMR-determined structure of Link_TSG6 complexed with hyaluronan,[26] Figure 11(a), Y59 and Y78 lie flat against the protein surface and are highly ordered, even unusually displaying distinct H^{ϵ}/H^{δ} chemical shifts on each side of the ring. In contrast, in the free protein they project outward and are highly mobile, having only one chemical shift for the H^{δ} and H^{ϵ} ring protons. This is reflected in the simulation, which predicts that amino acids known to be involved in hyaluronan, bikunin, and heparin binding are among the most mobile in the protein. Therefore, the specific, average side-chain orientations determined from the few observed NOEs is a gross simplification arising from simulated annealing, which does not account for side-chain mobility. However, it may be possible to gain further insight into the specific dynamics of side chains by other methods and simultaneous use of NOEs and residual dipolar couplings[57] or by improving the experimental data set by using specific isotopic labeling.[58]

This new representation of the Link-module of TSG-6 provides information about dynamics on the nanosecond time scale while being as good a fit to experimental data over the majority of the structure as a static ensemble derived by simulated annealing. It will be invaluable for understanding the dynamic molecular basis of its interactions with ligands and the role and conservation of individual amino acids in the Link-module superfamily. Ultimately, these advances will underpin a dynamic understanding of extracellular matrix assembly, remodeling, and the changes that occur in physiological and disease states.

Abbreviations: XRC − X-ray crystallography; NMR − nuclear magnetic resonance; MD − molecular dynamics; NOE − nuclear Overhauser effect; NOESY − NOE spectroscopy; HSQC − heteronuclear single quantum coherence; HMQC − heteronuclear multiple quantum coherence; CSD − chemical shift deviation; TSG-6 − tumor necrosis factor stimulated gene-6; Link_TSG6 − recombinant link-module domain from human TSG-6; RMS − root-mean-square; RMSD − root-mean-square deviation; RMSF − root-mean-square fluctuation.

Acknowledgment. We thank the Oxford Centre for Molecular Sciences (University of Oxford, U.K.), and in particular Professor Iain D. Campbell, for the use of their NMR spectrometers. We acknowledge Julian Heuberger for helping to collect NMR temperature coefficient data. A.A. was funded by a Biotechnology and Biological Sciences Research Council Sir David Phillips Research Fellowship. C.D.B. was funded by the ARC (Arthritis Research Council) and Yamouchi Research Institute. V.A.H. and A.J.D. were funded by the ARC (16539). A.D.M. was funded by the NIH (GM51501).

Supporting Information Available: Temperature coefficients and chemical shift deviations for Link_TSG6, experimental order parameters, scatter plots for each of the (φ,ψ)-angles from the molecular dynamics simulation of Link_TSG6, average and standard deviation for these (φ,ψ)-angles, and graphs of the simulation predictions (order parameters, relaxation measurements. and scalar couplings) plotted against experimental measurements with correlation coefficients and RMSDs. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) D'Aquino, J. A.; Freire, E.; Amzel, L. M. *Proteins* **2000**, *S4*, 93−107.

(2) Monecke, P. *Biophys. J.* **2006**, *90*, 841−850.

(3) Rajamani, D.; Thiel, S.; Vajda, S.; Camacho, C. J. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 11287−11292.

(4) Kimura, S. R.; Brower, R. C.; Vajda, S.; Camacho, C. J. *Biophys. J.* **2001**, *80*, 635−642.

(5) Wider, G. *Biotechniques* **2000**, *29*, 1278−1290.

(6) Dauter, Z. *Acta Crystallogr. D* **2006**, *62*, 1−11.

(7) Furnham, N.; Blundell, T. L.; DePristo, M. A.; Terwilliger, T. C. *Nat. Struct. Mol. Biol.* **2006**, *13*, 184−185.

(8) Derewenda, Z. S.; Vekilov, P. G. *Acta Crystallogr. D* **2006**, 62, 116−124.

(9) Eyal, E.; Gerzon, S.; Potapov, V.; Edelman, M.; Sobolev, V. *J. Mol. Biol.* **2005**, *351*, 431−442.

(10) Schwieters, C. D.; Clore, G. M. *J. Biomol. NMR* **2002**, *23*, 221−225.

(11) Schröder, G. F.; Alexiev, U.; Grubmüller, H. *Biophys. J.* **2005**, *89*, 3757−3770.

(12) Farrow, N. A.; Muhandiram, D. R.; Singer, A. U.; Pascal, S. M.; Kay, C. M.; Gish, G.; Shoelson, S. E.; Pawson, T.; Forman-Kay, J. D.; Kay, L. E. *Biochemistry* **1994**, *33*, 5984−6003.

(13) Malliavin, T. E. *Curr. Org. Chem.* **2006**, *10*, 555−568.

(14) Clore, G. M.; Schwieters, C. D. *J. Mol. Biol.* **2006**, *355*, 879−886.

(15) Nederveen, A. J.; Bonvin, A. *J. Chem. Theor. Comp.* **2005**, *1*, 363−374.

(16) Camacho, C. J. *Proteins: Struct., Funct., Bioinform.* **2005**, *60*, 245−251.

(17) Kieseritzky, G.; Morra, G.; Knapp, E. W. *J. Biol. Inorg. Chem.* **2006**, *11*, 26−40.

(18) Tammi, M. I.; Day, A. J.; Turley, E. A. *J. Biol. Chem.* **2002**, *277*, 4581−4584.

(19) Day, A. J.; Prestwich, G. D. *J. Biol. Chem.* **2002**, *277*, 4585−4588.

(20) Day, A. J.; Sheehan, J. K. *Curr. Opin. Struct. Biol.* **2001**, *11*, 617−622.

(21) Day, A. J.; de la Motte, C. A. *Trends Immunol.* **2005**, *26*, 637−643.

(22) Milner, C. M.; Higman, V. A.; Day, A. J. *Biochem. Soc. Trans.* **2006**, *34*, 446−50.

(23) Milner, C. M.; Day, A. J. *J. Cell. Sci.* **2003**, *116*, 1863−73.

(24) Kohda, D.; Morton, C. J.; Parkar, A. A.; Hatanaka, H.; Inagaki, F. M.; Campbell, I. D.; Day, A. J. *Cell* **1996**, *86*, 767−75.

(25) Mahoney, D. J.; Blundell, C. D.; Day, A. J. *J. Biol. Chem.* **2001**, *276*, 22764−71.

(26) Blundell, C. D.; Mahoney, D. J.; Almond, A.; DeAngelis, P. L.; Kahmann, J. D.; Teriete, P.; Pickford, A. R.; Campbell, I. D.; Day, A. J. *J. Biol. Chem.* **2003**, *278*, 49261−70.

(27) Mahoney, D. J.; Mulloy, B.; Forster, M. J.; Blundell, C. D.; Fries, E.; Milner, C. M.; Day, A. J. *J. Biol. Chem.* **2005**, *280*, 27044−27055.

(28) Almond, A.; Brass, A.; Sheehan, J. K. *J. Mol. Biol.* **1998**, *284*, 1425−1437.

(29) Almond, A.; DeAngelis, P. L.; Blundell, C. D. *J. Am. Chem. Soc.* **2005**, *127*, 1086−1087.

(30) Mulloy, B.; Forster, M. J. *Glycobiology* **2000**, *10*, 1147−56.

(31) Verli, H.; Guimaraes, J. A. *Carbohydr. Res.* **2004**, *339*, 281−90.

(32) Blundell, C. D.; Almond, A.; Mahoney, D. J.; DeAngelis, P. L.; Campbell, I. D.; Day, A. J. *J. Biol. Chem.* **2005**, *280*, 18189−201.

(33) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, 4, 187−217.

(34) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(35) Buck, M.; Bonnet, S.; Pastor, R. W.; MacKerell, A. D., Jr. *Biophys. J.* **2006**, *90*, L36−L38.

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

(37) Tsodikov, O. V.; Record, M. T., Jr.; Sergeev, Y. V. *J. Comput. Chem.* **2002**, *23*, 600−609.

(38) Vendruscolo, M.; Paci, E.; Dobson, C. M.; Karplus, M. *J. Am. Chem. Soc.* **2003**, *125*, 15686−7.

(39) Best, R. B.; Vendruscolo, M. *Structure* **2006**, *14*, 97−106.

(40) Englander, S. W.; Sosnick, T. R.; Englander, J. J.; Mayne, L. *Curr. Opin. Struct. Biol.* **1996**, *6*, 18−23.

(41) Blundell, C. D.; Deangelis, P. L.; Almond, A. *Biochem. J.* **2006**, *396*, 487−98.

(42) Andersen, N. H.; Neidigh, J. W.; Harris, S. M.; Lee, G. M.; Liu, Z. H.; Tong, H. *J. Am. Chem. Soc.* **1997**, *119*, 8547−8561.

(43) Cierpicki, T.; Otlewski, J. *J. Mol. Biol.* **2000**, *302*, 1179−1192.

(44) Tropp, J. *J. Chem. Phys.* **1980**, *72*, 6035−6043.

(45) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4546−4559.

(46) Brüschweiler, R. *Curr. Opin. Struct. Biol.* **2003**, *13*, 175−83.

(47) Bruschweiler, R.; Case, D. A. *J. Am. Chem. Soc.* **1994**, *116*, 11199−11200.

(48) Zagrovic, B.; van Gunsteren, W. F. *Proteins: Struct., Funct., Bioinform.* **2006**, *63*, 210−218.

(49) Farrow, N. A.; Zhang, O.; Forman-Kay, J. D.; Kay, L. E. *Biochemistry* **1997**, *36*, 2390−402.

(50) Krishnan, V. V.; Cosman, M. *J. Biomol. NMR* **1998**, *12*, 177−182.

(51) Kahmann, J. D.; O'Brien, R.; Werner, J. M.; Heinegard, D.; Ladbury, J. E.; Campbell, I. D.; Day, A. J. *Struct. Fold. Des.* **2000**, *8*, 763−774.

(52) Guo, G. J.; Zhang, Y. G. *Mol. Phys.* **2001**, *99*, 283−289.

(53) Smith, P. E.; Vangunsteren, W. F. *Chem. Phys. Lett.* **1993**, *215*, 315−318.

(54) Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D., Jr. *Chem. Phys. Lett.* **2006**, *418*, 245−249.

(55) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., III *J. Am. Chem. Soc.* **2004**, *126*, 698−9.

(56) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., III *J. Comput. Chem.* **2004**, *25*, 1400−1415.

(57) Peti, W.; Meiler, J.; Bruschweiler, R.; Griesinger, C. *J. Am. Chem. Soc.* **2002**, *124*, 5822−33.

(58) Rosen, M.; Gardner, K.; Willis, R.; Parris, W.; Pawson, T.; Kay, L. E. *J. Mol. Biol.* **1996**, *263*, 627−636.

CT600236Q

JCTC Journal of Chemical Theory and Computation

Temperature-Dependent Probabilistic Roadmap Algorithm for Calculating Variationally Optimized Conformational Transition Pathways

Haijun Yang,[†] Hao Wu,[†] Dawei Li,[†] Li Han,*[,‡] and Shuanghong Huo*[,†]

Gustaf H. Carlson School of Chemistry and Biochemistry and Department of Mathematics and Computer Science, Clark University, 950 Main Street, Worcester, Massachusetts 01610

Received August 18, 2005

Abstract: In this paper we present a method to calculate a temperature-dependent optimized conformational transition pathways. This method is based on the maximization of the flux derived from the Smoluchowski equation and is implemented with a probabilistic roadmap algorithm. We have tested the algorithm on four systems—the Müller potential, the three-hole potential, alanine dipeptide, and the folding of β-hairpin. Comparison is made with existing algorithms designed for the calculation of protein conformational transition and folding pathways. The applications demonstrate the ability of the algorithm to isolate a temperature-dependent optimal reaction path with improved sampling and efficiency.

I. Introduction

Protein/peptide conformational transitions are closely related to their biological function. Here, the conformational transition is broadly defined, including the transition from an unfolded state to the folded state, from a partially unfolded intermediate state to the native state, or the disorder-to-order transition for intrinsically disordered proteins. It is of great challenge and significance to describe this long-time process accurately and effectively using computational approaches.[1] This process can be investigated using the "chain of states" methods to identify all intermediates and transition states between the two-end conformations concomitantly.[2−22] A number of these methods search for the steepest descent path or minimum-energy path. While the steepest descent path may make significant contributions to the reaction rate, it has been noticed that, in general, the observed average dynamics is not determined solely by the steepest descent path.[13,23] Therefore, including temperature effect in the calculation of transition pathway is desired. Methods have been developed to generate an ensemble of transition pathways.[24−27] In principle, the transition-path sampling (TPS) method can be applied to any system to provide detailed information about the transition at a given temperature. However, in practice, they are too computationally demanding for larger systems, e.g. a collection of a few hundred unfolding pathways of a 16-residue peptide took ∼3 months on an 8-node, 1-GHz Athlon PC cluster.[28] Furthermore, the quality of the path sampling relies on a proper definition of the stable states by order parameters. It may need an even longer time to fulfill the requirements of order parameters by trial and error.

Elber and Shalloway have developed a cost-effective method to include temperature effect based on a classical expansion of path integral for a stochastic process. However, in their method, the time for transition is a variable.[29] Huo and Straub have developed a time-independent algorithm to search for protein/peptide conformational transition pathways at a well-defined temperature.[30] This algorithm is based on the work of Berkowitz and co-workers who derived variational formulas for the optimal transition pathway connecting the reactant and the product.[3] Suppose that the conformational transition is a stochastic process, the probability distribution of the system is well described by the Smoluchowski equation.[31] Assuming that the friction (γ) of the system is isotropic and spatially independent, the optimal

* Corresponding author fax: (508) 793-8861; e-mail: shuo@clarku.edu (S.H.) and fax: (508) 421-3715; e-mail: lhan@clarku.edu (L.H.).

[†] Gustaf H. Carlson School of Chemistry and Biochemistry.

[‡] Department of Mathematics and Computer Science.

10.1021/ct0502054 CCC: $37.00
Published on Web 11/15/2006

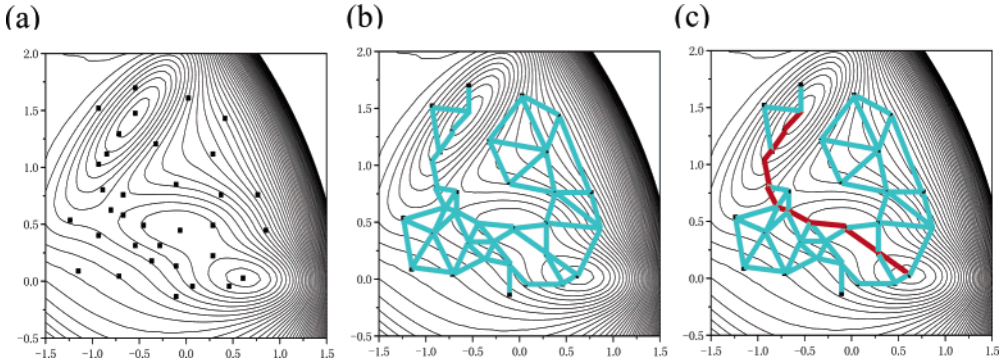


Figure 1. Schematic representation of the PRM approach applied to the Müller potential.[40,41] (a) Node generation. Each black dot represents a node. (b) Roadmap connection. All of the blue lines are edges connecting the nodes. (c) Roadmap query. The shortest path is denoted by the red line.

reaction pathway is defined as the one that maximizes the reactive flux (j) between the reactant ($\{\mathbf{r}_R\}$) and the product ($\{\mathbf{r}_P\}$), $Max\{j \propto [1/\gamma \int e^{\beta U(\mathbf{r})} dl(\mathbf{r})]\}$, or minimizes the mean first passage time (τ),[3,30] $Min\{\tau \propto \gamma \int e^{\beta U(\mathbf{r})} dl(\mathbf{r})\}$. The objective is to minimize the line integral

$$P = \int_{\Gamma_R}^{\mathbf{r}_P} e^{\beta U(\mathbf{r})} dl(\mathbf{r}) \tag{1}$$

In eq 1, $\beta = 1/kT$, where k is the Boltzmann constant, and T is the absolute temperature. $U(\mathbf{r})$ is the effective energy of the system for a given conformation $\{\mathbf{r}\}$, including the intramolecular interaction energy of the protein/peptide and the solvation free energy. Also in eq 1, $dl(\mathbf{r})$ is the line segment along the path. By minimizing the line integral of eq 1 using the self-penalty walk method,[7] one can obtain an optimal pathway corresponding to the fastest reaction rate.[30] This algorithm is called MaxFlux and has been successfully applied to study the conformational change of peptides.[32,33]

Based on the same idea of maximizing the flux, but using the differential equation derived by Berkowitz et al. instead of the integral equation,[3] Crehuet and Field[34] described an alternative MaxFlux by implementing the nudged-elastic-band method,[11] called MaxFlux-NEB. It has been demonstrated that MaxFlux-NEB works well for small molecules and model peptides.[34] However, both the original MaxFlux and MaxFlux-NEB rely on global minimization methods. They both start from an initial guess which is usually a linear interpolation between the reactant and the product. If the initial guess happens to be close to the optimal pathway, a local minimization method is good enough to find the optimal pathway.[30] However, when the initial guess is far away from the final path, computationally demanding global minimization methods are required, for example, simulated annealing or a molecular dynamics parallel tempering scheme.[35] Note that we are talking about searching for the global minimum in the path space, instead of the conformational space. The former is more time-consuming than the latter. One way to deal with the global minimization problem is to employ a stochastic sampling method proposed by Woolf and co-workers.[26] However, applications in biomolecular systems are still the major challenge for this method.

In this work, we present an alternative way to tackle the sampling issue. We propose to combine MaxFlux with the probabilistic roadmap (PRM) motion planning method that is originally developed for robotics motion planning and has been applied to study various problems.[36−38] We do not intend to give a complete review of PRM, rather, we mainly introduce Amato and co-workers's PRM application on protein folding landscape and kinetics[39] because their work is closely related to our present article. The details of the combined approach of MaxFlux and PRM are presented in the MaxFlux-PRM method section, and the applications are in Results and Discussion section.

II. The PRM Method

The basic idea of PRM is to extract the pathways from a roadmap. There are three stages of PRM as shown in Figure 1: (1) node generation; (2) roadmap connection; and (3) roadmap query. The objective of the node generation is to sufficiently sample the region of conformational space surrounding the final conformation or the product, e.g. the native state of protein. A conformation is called a node (q) in the roadmap. Biased sampling strategies in the (ϕ,ψ) space have been successfully applied for medium-size proteins (more than 100 amino acids).[38,39,42] A threshold can be set to remove the high energy nodes. The details of the techniques of node generation to ensure adequate coverage of the conformational space will be presented in the MaxFlux-PRM method section.

The second stage is to connect the generated nodes to obtain a roadmap or a graph. The objective of this stage is to construct a roadmap encoding the representative paths. For each generated node or conformation, its k nearest neighbors will be found. The neighbor can be defined using the root-mean-square (rms) distance or Euclidean distance. The value of k is adjustable. Previously, $k = 20$ was set.[38,39,42] The connection between a node (e.g. q_1) and its neighbor (e.g. q_2) is called an edge. A weight is assigned to each edge defined as

$$\omega(q_1, q_2) = -\ln(P_{\text{connection}}) \tag{2}$$

where

$$P_{\text{connection}} = \begin{cases} e^{-\Delta E/k_B T} & \text{if } \Delta E > 0,\ \Delta E = E(q_2) - E(q_1) \\ 1 & \text{if } \Delta E \le 0 \end{cases} \tag{3}$$

By connecting all the nodes, the roadmap is constructed and

appears like a net lying on the energy surface (Figure 1(b)). However, note that in eq 3, there is no difference between downhill movement and moving on a flat surface because when either $\Delta E < 0$ or $\Delta E = 0$, $P_{\text{connection}} = 1$. Moreover, the magnitude of downhill movement is not taken into account, for example, $\Delta E = -0.0001$ kcal/mol has the same value of $P_{\text{connection}}$ as $\Delta E = -1000$ kcal/mol.

The final stage is to query the roadmap to extract the optimal pathway from a given initial conformation to the final destination, such as the native state. Dijkstra's algorithm[43] is employed to find the smallest weight path between the initial and final conformation (Figure 1(c)). The optimal path from q_1 to q_n through $n-2$ nodes satisfies

$$w(q_1,q_2,.....q_n) = \min \{\sum_{i=1}^{n-1} - \ln(p_{i\text{connection}})\} \qquad (4)$$

where n is not fixed and may vary from path to path. If we plug eq 3 into eq 4, we can get an equivalent criterion of the optimal path as

$$w(q_1,q_2,.....q_n) = \min \left\{\frac{1}{kT}\sum_{\text{uphill}} \Delta E_i\right\} \quad \text{or}$$
$$w(q_1,q_2,.....q_n) = \min \{\sum_{\text{uphill}} \Delta E_i\} \qquad (5)$$

The querying process searches for the **least uphill** movement. As a result, this PRM method misses the intermediate states if any. In other words, it is more suitable to two-state transitions than multistate transitions.This shortcoming can be illustrated on model potentials as shown in the section of results. Furthermore, since $1/kT$ can be moved outside the sum in eq 5, the pathway does not change with temperature. As a result, the pathway found by PRM is not temperature dependent. However, the strength of PRM is its enhanced sampling and efficiency. It can map the potential energy landscape and generate sets of representative transition pathways from a single roadmap. The system can avoid being trapped in a local minimum, and no detailed simulations are needed plus it is initial-guess free. The PRM method has been applied to study protein folding pathways of 14 proteins up to the size of 110 amino acids.[39] The order of secondary structure formation along the PRM pathway has been found to be in good agreement with the experimental results.[39,42] Note that all of the proteins that have been studied by PRM have two-state folding behavior. PRM and MaxFlux naturally complement each other. By replacing the edge weight function of PRM in eqs 2−5 with the MaxFlux equation (eq 1), we can search for temperature-dependent transition pathways with enhanced sampling and the capability to study the multistate transition pathways. This idea falls in the same category as a recent effort in network models for kinetics,[44−46] while the differences are in node generation and the definition of edge weight. In MaxFlux-PRM, nodes are not generated by molecular dynamics or replica exchange as other network models.

III. The MaxFlux-PRM Methods

We have tested the MaxFlux-PRM method on two model potentials, Müller and 3-hole, and two peptides, alanine dipeptide (AD) and β-hairpin. For the Müller potential and 3-hole potential, we uniformly searched the potential energy surface to generate 5000 and 20 000 nodes. All of the generated nodes were accepted without setting the energy threshold. Five intermediate conformations were added between two nodes by linear interpolation. The edge weight between nodes, e.g. q_0 and q_1, can be defined as

$$w_i(c_0 = q_0, c_1, ..., c_m, c_{m+1} = q_1) = \sum_{i=0}^{m} e^{\beta U(c_i)} |r(c_{i+1}) - r(c_i)| \qquad (6)$$

where $m = 5$ and $|r(c_{i+1}) - r(c_i)|$ is the rms distance between conformations c_{i+1} and c_i. $\beta = 1/kT$, where k is the Boltzmann constant, and T is the absolute temperature. U is the energy of the system. For these model potentials, 5000 nodes can be considered a complete search; therefore, interpolation between nodes did not significantly improve the accuracy of the path but it made the path smoother.

For AD and β-hairpin, we describe the protein intramolecular interaction energy using CHARMM 19 polar hydrogen energy function[47] and an effective energy function (EEF1) for solvation.[48] We search for the path in the full $3n$-dimension Cartesian coordinate. For AD, a node was generated by assigning each backbone dihedral angle a random value in its allowable range ($[-\pi, \pi]$ in this simulation). Once the dihedral angles were generated, the conformation was minimized 100 steps using the adopted basis Newton−Raphson method with restraints on the backbone dihedral angles. Suppose that the thermal fluctuation of each dihedral angle is 10°, a search with more than 36×36 nodes can be considered complete for AD. We randomly generated 5000 nodes. Twenty nearest neighbors were found for each node. No linear interpolation between neighbors was applied.

Due to the high dimensionality of the protein conformation space, we adopt a focused sampling strategy[42] for β-hairpin based on the fact that the reactant and product conformations are known. The MaxFlux-PRM procedure was illustrated in the flowchart (Figure 2). In particular, we generate a set of Gaussian distributions around the backbone dihedral angles of the reactant and product with a set of standard deviations (STDs) of {3°, 5°, 10°, 15°, 20°, 25°, 30°, 40°, 50°, 60°, 80°, 100°, 130°, 160°}. The small STDs capture the detail around the reactant and product, while the larger STDs ensure adequate roadmap coverage of the conformation space. By random sampling from these distributions, we initially generated 200 000 nodes. The side chains were built using CHARMM and minimized (3000 steps or $\Delta E < 0.005$ kcal/mol whatever comes first) with the backbone dihedral angles restrained. This procedure helps to remove some bad contacts. To sample the side-chain conformations, 10^4 steps of MC were performed for each node by moving the side chain only using the Monte Carlo (MC) module in CHARMM.[49] By eliminating the nodes with a threshold, $E_{\max} = -400$ kcal/mol, we obtained 93 886 nodes, where $E_{\max}$ is adjustable and varies with the size of protein. The number of nodes is almost four times of that reported recently for the same system and the same force-field.[50] We used rms

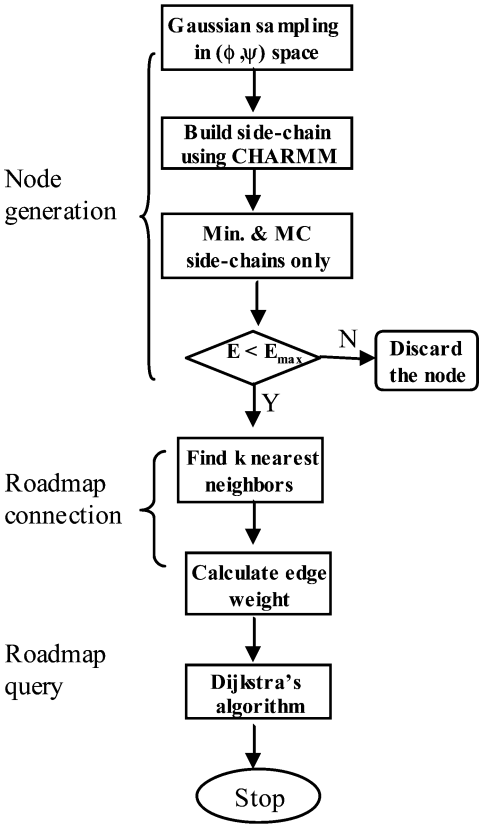


Figure 2. Flowchart for MaxFlux-PRM (β-hairpin).

deviation as the criterion to find 40 nearest neighbors for each node with the prerequisite that the node has no more than one backbone hydrogen bond difference with its neighbor. The nodes that do not satisfy the neighboring requirements were removed. In fact, this kind of node is rare. No interpolation between nodes was carried out.

To introduce the temperature effect, we define the edge weight ($\omega(q_0,q_1)$) as

$$w(q_0,q_1) = e^{\beta U(q_0)}|r(q_1) - r(q_0)| \tag{7}$$

where $|r(q_1) - r(q_0)|$ is the all-atom rms distance between conformation q_0 and its neighbor q_1 in the Cartesian coordinate. $\beta = 1/k_BT$, where k_B is the Boltzmann constant and T is equal to 300 K for AD and β hairpin. U is the effective energy of the system (protein plus solvent) including the intramolecular interaction energy of the protein and the solvation free energy. Once all the neighbors were identified and all the edge weights were calculated, the roadmap was constructed. To find the minimum weight path between the extended state and the native fold, that corresponds to the maximum flux path, Dijkstra's algorithm was used to query the roadmap. The minimum weight path between the initial and final states is defined in the discretized form of eq 1 as

$$\omega_{\min} = \min\{\sum_{i=0}^{n-1} e^{\beta U(q_i)}|\mathrm{r}(q_{i+1}) - \mathrm{r}(q_i)|\} \tag{8}$$

where the path connecting the initial and final states goes through $n-1$ nodes. When the samping is enough, either $U(q_i)$ or $U(q_{i+1})$ can be used in eq 8. Otherwise, the midpoint energy should be used. The weight of the path is the sum of the weights of its constituent edges. This is a well-defined single-pair shortest-path problem when all edges have nonnegative weight. As mentioned in the Introduction, the minimum weight path defined in eq 8 corresponds to the shortest mean first passage time or the fastest reaction rate.

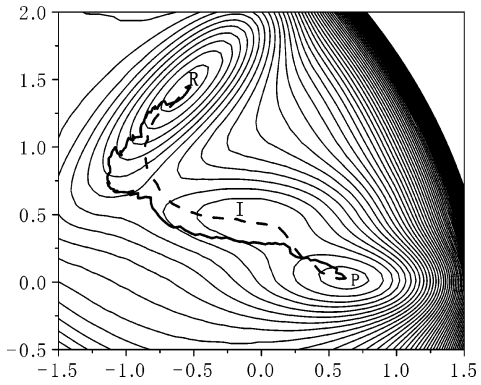


Figure 3. The transition pathway found for the Müller potential. The solid line is the PRM pathway, while the dashed line is the MaxFlux-PRM pathway. The pathways with 20 000 nodes are presented.

IV. Results and Discussion

Müller Potential. The MaxFlux-PRM method was applied on the two-dimensional model potential of Fukui, Kato, and Fujimoto[40] studied by Müller.[41] There are three minima in the Müller potential: the reactant, the intermediate (a shallow metastable minimum), and the product. As shown in Figure 3, the PRM method (solid line) misses the metastable intermediate state. If a pathway goes through an intermediate state, the system has to move downhill to visit the minimum and then move uphill to escape from the minimum. According to eq 5, in the PRM method, only the uphill transition contributes to the edge weight, while the downhill transition is not taken into account. As a result, the downhill-to-uphill edge weight to visit the intermediate state is larger than that obtained by avoiding the intermediate state and directly going down to the product well. When we replaced the PRM edge weight function (eq 5) with the MaxFlux criterion (eq 1) to search for the pathway with maximum reactive flux or minimum mean first passage time, we identified a new path that visited the intermediate state (dashed line). The MaxFlux-PRM path is an estimate of the best path for the reaction at nonzero temperature $1/\beta$ (β=0.06). This path is comparable to what would be defined by a minimum-energy path.[51]

3-Hole Potential. A more challenging test is the 3-hole potential developed by Huo and Straub[30] which has been used by other groups.[29,34] It is designed to isolate the globally optimal transition pathway. As shown in Figure 4, besides the reactant (lower left well) and product (lower right well) minima there is a third energy minimum (above center). There are two possible reaction pathways. The lower path moves roughly left-to-right between the reactant and product wells crossing a single high energy barrier. The upper path overcomes a low energy barrier, through a basin, and then overcomes another low energy barrier before reaching the product state. The optimal minimum-energy path is the lower path that has a high barrier but is shorter.[30] However, the

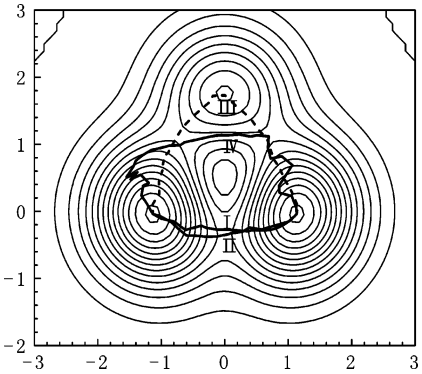


Figure 4. The transition pathway found for the 3-hole potential. MaxFlux-PRM has identified the lower reaction pathways (I and II) at high temperature ($\beta = 1$ or 2), while the upper pathway(III) is dominant at low temperature ($\beta = 3.3$) using MaxFlux-PRM. In contrast, PRM found either pathway IV or pathway II at whatever temperature. Pathways IV and II are identical based on the PRM criterion in eq 5. We found that 5000 nodes can yield a smooth path.

lower path is not always the path of maximum transition rate at nonzero temperature. According to the definition of the optimal path with maximum flux in eq 1, the lower reaction path will be the pathway of maximum flux only at high temperatures. At low enough temperatures the upper pathway, which crosses two lower energy saddle points, is dominant.[30]

Independent runs of PRM with different seeds of random number generator have identified either the pathway IV or the lower path II (Figure 4) at any of the tested temperatures because these paths are identical according to eq 5 and the method is temperature-independent. Interestingly, pathway IV is similar to the "last" stable path in the upper region identified by MaxFlux-NEB at high temperature.[34] Once again, this test has illustrated that the PRM method is more suitable to two-state transitions than multistate transitions. It fails to identify the intermediate state due to the flaw of the edge weight function (eqs 2−5). In contrast, the MaxFlux-PRM method has identified different pathways at different temperatures (Figure 4), consistent with the original MaxFlux results[30] and MaxFlux-NEB.[34]

Alanine Dipeptide. We have applied the MaxFlux-PRM method to alanine dipeptide (AD). The potential energy surface and the pathways of conformational transition of this system have been studied extensively.[7,30,34,52] Although the molecule is small, it has emerged as a standard test for reaction path algorithms. The reactant conformation is $C7_{eq}$ at ($\phi = -86°$, $\psi = 79°$) with effective energy equal to −28.74 kcal/mol, and the product conformation is $C7_{ax}$ ($\phi = 76°$, $\psi = -55°$) whose effective energy is −30.17 kcal/mol. For the pathway obtained at 300 K, $E_{barrier} - E_{reactant}$ is equal to 4.98 kcal/mol and $E_{barrier} - E_{product}$ is 6.40 kcal/mol, where $E_{barrier}$ is the maximum effective energy of the conformation along the optimal path. The effective energy barrier is ca. 10 times of RT at room temperature; therefore, the conformation transition pathway at 300 K is similar to the minimum-energy path.

β-Hairpin Formation. Investigating the folding of key secondary structural elements is proving to be useful for understanding the thermodynamics and kinetics of large protein folding. Therefore, the folding mechanism of β-hairpin has been studied extensively by experimental[53−64] and computational approaches.[45,46,50,65−84] Herein, we demonstrate the MaxFlux-PRM method can be applied to β-hairpin to locate optimal folding pathways from the extended state to the folded state (pdb entry: 3GB1,[85] residues 41−56). The folding path was analyzed using the all-atom rms deviation from the native state, radius of gyration (R_g) for the hydrophobic core (W43, Y45, F52, and V54 in Figure 6), and the number of native contacts (N_{NC}) as shown in Figure

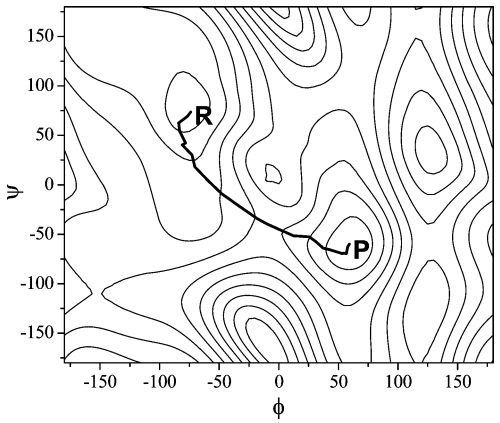


Figure 5. The reaction path of alanine dipeptide on the (ϕ,φ) potential energy map. The solid line connecting $C7_{eq}$($\phi = -86°$, $\varphi = 79°$) and $C7_{ax}$($\phi = 76°$, $\varphi = -55°$) is the transition pathway identified by the MaxFlux-PRM method.

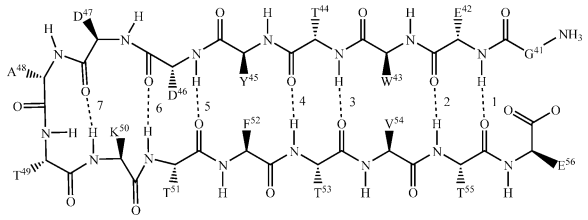


Figure 6. Schematic representation of the second β-hairpin of the B1 domain of streptococcal protein G (GB1). The main-chain hydrogen bonds are numbered from tail to the turn region.

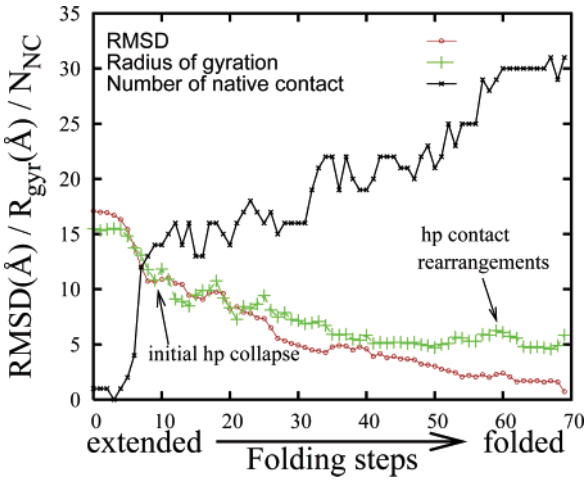


Figure 7. RMSD from the native state, radius of gyration for the hydrophobic core, and number of native contacts are computed along the path.

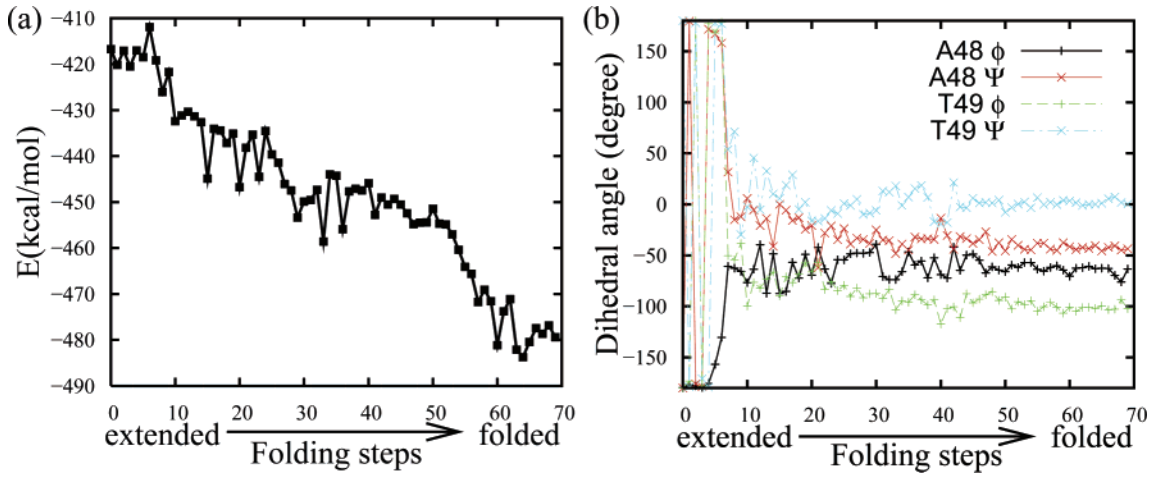


Figure 8. (a) Effective energy (molecular mechanics potential plus solvation energy) along the folding path. (b) Backbone dihedral angles of A48 and T49 in the turn region calculated along the folding path.

7. Two residues are considered to contact with each other if any heavy atom (C, N, O, S) of one residue is within 4.5 Å from any heavy atom of the other residue (the nearest neighbor in sequence is not included).[86] As shown in Figure 7, the very first step in the folding of β-hairpin from the extended state is hydrophobic (hp) collapse, as reflected in the initial drops in R_{gyr} for the hydrophobic core from 15.5 to 10.8 Å (at folding step #10). This initial hydrophobic collapse has been observed by a number of simulations[68,71,77,79] and experiments[63] on β-hairpin. Concomitantly, N_{NC} jumps to 14 among which almost all are the contacts between residues i and $i+2$. The side chains including hydrophobic contacts rearrange in the later stage to form the native packing (Figure 7, a local maximum in R_{gyr} at folding step #60). The effective energy spans from ca. −410 kcal/mol to ca. −480 kcal/mol along the folding path (Figure 8a). It is obvious that the native state (at folding step 70) is not the global minimum because there are two minima along the path which have even lower effective energy (Figure 8a). It is possible that multiple conformations coexist because experimentally estimated population of the β hairpin is only ca. 30%,[54,60,87] and the presence of multiple conformers with effective energy lower than native state energy was also reported by other computational work.[22]

To describe the turn formation, we used the dihedral angles of the central residues (($\phi_{A48} = -60°$, $\psi_{A48} = -30°$) and ($\phi_{T49} = -90°$, $\psi_{T49} = 0°$)).[88] If the four dihedral angles fall within ±30° of these values, we consider that the turn is formed. According to this criterion, the turn is first formed at the folding step #13 (Figure 8b), that is after the initial hydrophobic collapse (at folding step #10 (Figure 7)). The highest energy (at folding step #7, Figure 8a) along the folding pathway corresponds to the turn initiation when the dihedral angles (($\phi_{A48} = -61°$, $\psi_{A48} = 32°$) and ($\phi_{T49} = -50$, and $\psi_{T49} = 54$)) of the central residues started the transition from the extended state region to the β turn region, consistent with the experimental results which have suggested that the rate-limiting step is the turn formation.[61,63,64,89]

To investigate the role of the native hydrogen bonds and the hydrophobic cluster in the folding process, we present the native hydrogen bond formation and the accessible

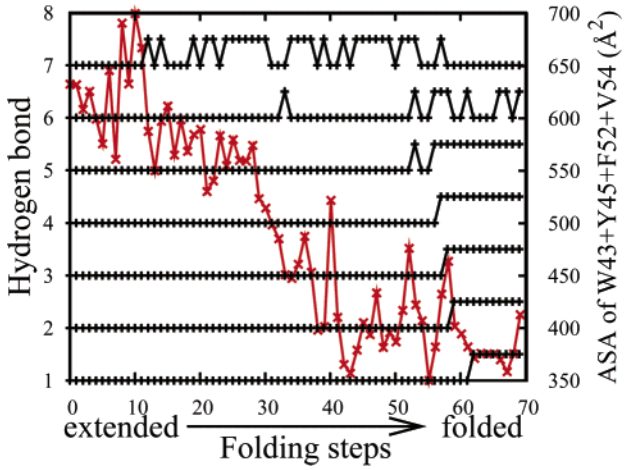


Figure 9. Interstrand main-chain hydrogen bond formation along the folding path. The hydrogen bonds are numbered in the same way as Figure 6. Hydrogen bond #7 was first formed at folding step 13. The accessible surface area (ASA) is denoted by the red line. A 1.4 Å probe was used.

surface area (ASA) of the hydrophobic side chains (W43, Y45, F52, and V54) along the folding path in Figure 9. We used the criteria of hydrogen bonds defined by Dinner et al.,[68] the distance between the corresponding heavy atoms is 3.4 Å and the out-of-line angle is less than 70°, who used the same force field on the same system. Even though the hydrogen bond in the turn region formed first, this hydrogen bond is not stable and even not present in the minimized structure because the distance between K50(N) and D47(O) is 3.49 Å. The hydrogen bond next to K50(N)−D47(O) is not stable either (Figure 9). The remaining hydrogen bonds are formed in the late stage of folding (after folding step #55). This kind stability is consistent with the analysis on the temperature-dependent hydrogen bond stability.[68,74,83] However, as demonstrated by Zhang et al.[84] in their simulations of trpzip2 the hydrogen bond with the highest stability is not necessarily important in the folding process. The hydrogen bond in the turn region with the lowest stability forms first in the transition state in their simulations,[84] which is in line with our results. After the turn formation, yet before the hydrogen bonds to continue to propagate, the ASA drops

Table 1. Computational Cost (β-Hairpin)

process	node generation minimization/MC[a] (h)	roadmap connection[b] (h)	roadmap query (min)
running time	18/82	74	10

[a] There are 200 000 nodes were generated. The minimization and MC were for the side chain only. [b] Ca. 3.6×10^6 edges were connected, and their edge weights were calculated.

to 456 Å^2 (close to the native state value 422 Å^2) at folding step #35 and remains below this value for the majority of the conformers along the rest of the path, strongly suggesting that the hydrophobic assembly is formed before the propagation of the hydrogen bonds from the turn to the tail. Figure 9 is in generally good agreement with Figure 3c of ref 68.

In general, our mechanism is a mixture of zipper model[54,65−67] and hydrophobic core-centric model.[68,69,71,79] The sequence of events along the folding pathway is as follows: (1) initial hydrophobic collapse; (2) formation of the β turn; (3) occurrence of hydrophobic cluster; and (4) repacking of the hydrophobic core accompanied by the propagation of the hydrogen bonds beyond the turn region (in the late stage of folding). Our mechanism is supported by the fact that the mutation in the turn region affects the folding rate, while the stronger hydrophobic cluster only decreases the unfolding rate,[64] suggesting that the turn forms in the transition state while the hydrophobic cluster formed after the transition state.

Advantages and Limitations of MaxFlux-PRM. To better assess the computational cost, we list the running time of each step of path searching in Table 1. All of the calculations are on a single-CPU of AMD Opteron (1.4 GHz). The two most time-consuming tasks are the MC in node generation and the pairwise rms deviation calculation during roadmap connection, as seen in Table 1. This computational cost of pairwise rms deviation, which scales up as N^2 (N = number of nodes), can be reduced by the two-step divide-and-conquer approach employed by Brooks and co-workers.[22] To save running time, for node generation, minimization, MC, and roadmap connection, the job can be readily partitioned into multiple CPUs.

As demonstrated on the Müller potential and 3-hole potential, the edge weight of MaxFlux-PRM is more physical than the original PRM which must miss the intermediate state due to the flaw in the edge weight definition. The advantage of MaxFlux-PRM over the original MaxFlux[30] is the enhanced sampling and efficiency. Huo and Straub started from the initial guess which is a straight line connecting the reactant and the product on the 3-hole potential. To identify the upper path at low temperature, they applied simulated annealing to minimize the path, which took hours on an SGI supercomputer. For the MaxFlux-PRM method, it only took a few minutes on a desktop PC. Furthermore, the original MaxFlux method requires several restraints on the path that are implemented by setting several parameters, for example, (1) a pseudobond restraint between nearest neighbor intermediate structures to encourage the mean-square distances between adjacent conformations along the path to be approximately equal and (2) a repulsion interaction between any non-nearest neighbor intermediate structures along the path. These restraints are implemented through three adjustable parameters. Additionally, one more parameter is needed to control the cooling rate for the simulated annealing. Even though the authors provided some rules of thumb to set the parameters,[30] it takes a huge amount of human time to adjust these parameters for each system. On the contrary, there are only a few adjustable parameters used in MaxFlux-PRM: n (the number of nodes), k (the number of nearest neighbors of each node), and E_{max} which vary with the size of the system but must be larger than the effective energy of the extended state.

For other available temperature-dependent reaction path algorithms such as the transition path sampling (TPS) method[24,25] and its analogue, discrete path sampling method,[50] the quality of the paths relies on a proper definition of the stable states by order parameters. To choose proper order parameters, one needs to carry out numerous trail simulations before production runs. In addition, TPS can tackle only one free-energy barrier at a time. Consequently, one needs to define a number of (meta)stable states. One limitation of the MaxFlux-PRM method is that it employs implicit solvation. The reason we used an implicit solvation model is that in Berkowitz's description of reactive flux,[3] which is based on Smoluchowski equation of stochastic processes, $U(\mathbf{r})$ is the potential of mean force of the system. The potential of mean force for a given conformation of a solvated macromolecule is the free energy of the system consisting of the macromolecule and the solvent with an average over all solvent degrees of freedom at a given temperature. The solvation free energy ($\Delta G_{solvation}$) can be calculated with the implicit solvation model, while the explicit water model is not practical to give $\Delta G_{solvation}$; nevertheless, the water expulsion in protein folding cannot be addressed with such implicit solvation. To compensate for this, one can use MaxFlux-PRM to get the initial path to define the (meta)stable states and employ TPS with explicit water to search for the ensemble of pathways. As a result, the definition of the (meta)stable states will be more reasonable than that obtained from an unfolding trajectory[28] because the (meta)stable states sampled by the unfolding simulation are not necessarily on the folding pathway.

In summary, as demonstrated in the applications of these four systems, we have successfully identified the temperature-dependent transition pathways at atomic level with improved sampling and efficiency by a new method, MaxFlux-PRM. The system can avoid being trapped in a local minimum, and no detailed simulations are needed plus initial-guess free**.** This method can be employed to study conformational transition of biomolecular systems not involving bond breaking or formation. The node-generation and roadmap connection are naturally parallel computing processes. The computational efficiency is expected to be dramatically increased by employing parallel processing techniques.[90]

Acknowledgment. This work was supported by NIH (1R15 AG025023-01 to S.H.). We are also thankful for support from the National Science Foundation Major Research Instrumentation Fund (DBI-0320875). S. Huo is very

grateful to Prof. Aaron Dinner for his help with the MC module in CHARMM.

References

(1) Elber, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 151−6.

(2) Berkov, D. V. *J. Magn. Magn. Mater.* **1998**, *186*, 199−213.

(3) Berkowitz, M.; Morgan, J. D.; McCammon, J. A.; Northrup, S. H. *J. Chem. Phys.* **1983**, *79*, 5563−5565.

(4) Cho, A. E.; Doll, J. D.; Freeman, D. L. *Chem. Phys. Lett.* **1994**, *229*, 218−224.

(5) Choi, C.; Elber, R. *J. Chem. Phys.* **1991**, *94*, 751−760.

(6) Chu, J.-W.; Trout, B. L.; Brooks, B. R. *J. Chem. Phys.* **2003**, *119*, 12708−12717.

(7) Czerminski, R.; Elber, R. *Int. J. Quantum Chem.* **1990**, *Suppl. 24*, 167−186.

(8) Elber, R.; Karplus, M. *Chem. Phys. Lett.* **1987**, *139*, 375−380.

(9) Elber, R.; Meller, J.; Olender, R. *J. Phys. Chem. B.* **1999**, *103*, 899−911.

(10) Gillilan, R. E.; Wilson, K. R. *J. Chem. Phys.* **1992**, *97*, 1757−1772.

(11) Jónsson, H.; Mills, G.; Jacobsen, K. W. Nudged Elastic Band Method for Finding Minimum Energy Paths of Transitions. In *Classical and Quantum Dynamics in Condensed Phase Simulations*, Berne, B. J., Ciccotti, G., Coker, D. F., Eds.; World Scientific: Singapore, 1998.

(12) Olender, R.; Elber, R. *J. Chem. Phys.* **1996**, *105*, 9299−9315.

(13) Olender, R.; Elber, R. *J. Mol. Struct.: THEOCHEM* **1997**, *398*−*399*, 63−71.

(14) Passerone, D.; Ceccarelli, M.; Parrinello, M. *J. Chem. Phys.* **2003**, *118*, 2025−2032.

(15) Passerone, D.; Parrinello, M. *Phys. Rev. Lett.* **2001**, *87*, 108302/1−108302/4.

(16) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. *J. Chem. Phys.* **2004**, *120*, 7877−7886.

(17) Smart, O. S. *Chem. Phys. Lett.* **1994**, *222*, 503−512.

(18) Trygubenko, S. A.; Wales, D. J. *J. Chem. Phys.* **2004**, *120*, 2082−2094.

(19) Ulitsky, A.; Elber, R. *J. Chem. Phys.* **1990**, *92*, 1510−1511.

(20) Woodcock, H. L.; Hodoscek, M.; Sherwood, P.; Lee, Y. S.; Schaefer, H. F. I.; Brooks, B. R. *Theor. Chem. Acc.* **2003**, *109*, 140−148.

(21) Zaloj, V.; Elber, R. *Comput. Phys. Commun.* **2000**, *128*, 118−127.

(22) Khavrutskii, I. V.; Byrd, R. H.; Brooks, C. L. *J. Chem. Phys.* **2006**, *124*, 194903/1−194903/14.

(23) Northrup, S. H.; McCammon, J. A. *J. Chem. Phys.* **1983**, *78*, 987−989.

(24) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291−318.

(25) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 1964−1977.

(26) Zuckerman, D. M.; Woolf, T. B. *J. Chem. Phys.* **1999**, *111*, 9475−9484.

(27) Zuckerman, D. M.; Woolf, T. B. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2000**, *111*, 016702−016711.

(28) Bolhuis, P. G. *Biophys. J.* **2005**, *88*, 50−61.

(29) Elber, R.; Shalloway, D. *J. Chem. Phys.* **2000**, *112*, 5539−5545.

(30) Huo, S.; Straub, J. E. *J. Chem. Phys.* **1997**, *107*, 5000−5006.

(31) Gardiner, C. W. *Handbook of stochastic methods for physics, chemistry, and the natural sciences*; Springer-Verlag: Berlin, New York, 2001.

(32) Huo, S.; Straub, J. E. *Proteins* **1999**, *36*, 249−61.

(33) Straub, J. E.; Guevara, J.; Huo, S.; Lee, J. P. *Acc. Chem. Res.* **2002**, *35*, 473−81.

(34) Crehuet, R.; Field, M. J. *J. Chem. Phys.* **2003**, *118*, 9563−9571.

(35) Brumer, Y.; Golosov, A. A.; Chen, Z. D.; Reichman, D. R. *J. Chem. Phys.* **2002**, *116*, 8376−8383.

(36) Kavraki, L.; Svestka, P.; Latombe, J. C.; Overmars, M. *IEEE Trans. Robot. Automat.* **1996**, *12*, 566−580.

(37) Apaydin, M. S.; Brutlag, D. L.; Guestrin, C.; Hsu, D.; Latombe, J. C.; Varma, C. *J. Comput. Biol.* **2003**, *10*, 257−81.

(38) Song, G.; Amato, N. M. *IEEE Trans. Robot. Automat.* **2004**, *20*, 60−71.

(39) Amato, N. M.; Dill, K. A.; Song, G. *J. Comput. Biol.* **2003**, *10*, 239−55.

(40) Fukui, K.; Kato, S.; Fujimoto, H. *J. Am. Chem. Soc.* **1975**, *97*, 1−7.

(41) Muller, K.; Brown, L. D. *Theor. Chim. Acta* **1979**, *53*, 75−93.

(42) Amato, N. M.; Song, G. *J. Comput. Biol.* **2002**, *9*, 149−68.

(43) Cormen, T. H.; Leiserson, C. E.; Rivest, R. L. *Introduction to Algorithms*; MIT Press: Cambridge, MA, 1992.

(44) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299−306.

(45) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766−70.

(46) Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6801−6.

(47) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.

(48) Lazaridis, T.; Karplus, M. *Proteins* **1999**, *35*, 133−52.

(49) Hu, J.; Ma, A.; Dinner, A. R. *J. Comput. Chem.* **2006**, *27*, 203−16.

(50) Evans, D. A.; Wales, D. J. *J. Chem. Phys.* **2004**, *121*, 1080−90.

(51) Fischer, S.; Karplus, M. *Chem. Phys. Lett.* **1992**, *194*, 252−261.

(52) Bolhuis, P. G.; Dellago, C.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5877−82.

(53) Kobayashi, N.; Endo, S.; Munekata, E. In *Conformational study on the IgG binding domain of protein G*, *Proc. Jpn. Symp, 1992*; Yanaihara, N., Ed.; ESCOM: Leiden, The Netherlands, 1992; pp 278−80.

(54) Munoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196−9.

(55) Blanco, F.; Ramirez-Alvarado, M.; Serrano, L. *Curr. Opin. Struct. Biol.* **1998**, *8*, 107−11.

(56) Kobayashi, N.; Honda, S.; Yoshii, H.; Munekata, E. *Biochemistry* **2000**, *39*, 6564−71.

(57) Honda, S.; Kobayashi, N.; Munekata, E. *J. Mol. Biol.* **2000**, *295*, 269−78.

(58) Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5578−83.

(59) Espinosa, J. F.; Munoz, V.; Gellman, S. H. *J. Mol. Biol.* **2001**, *306*, 397−402.

(60) Fesinmeyer, R. M.; Hudson, F. M.; Andersen, N. H. *J. Am. Chem. Soc.* **2004**, *126*, 7238−43.

(61) Dyer, R. B.; Maness, S. J.; Peterson, E. S.; Franzen, S.; Fesinmeyer, R. M.; Andersen, N. H. *Biochemistry* **2004**, *43*, 11560−6.

(62) Yang, W. Y.; Gruebele, M. *J. Am. Chem. Soc.* **2004**, *126*, 7758−9.

(63) Du, D.; Tucker, M. J.; Gai, F. *Biochemistry* **2006**, *45*, 2668−78.

(64) Du, D.; Zhu, Y.; Huang, C. Y.; Gai, F. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15915−20.

(65) Munoz, V.; Henry, E. R.; Hofrichter, J.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5872−9.

(66) Kolinski, A.; Ilkowski, B.; Skolnick, J. *Biophys. J.* **1999**, *77*, 2942−52.

(67) Klimov, D. K.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2544−9.

(68) Dinner, A. R.; Lazaridis, T.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9068−73.

(69) Pande, V. S.; Rokhsar, D. S. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9062−7.

(70) Ma, B.; Nussinov, R. *J. Mol. Biol.* **2000**, *296*, 1091−104.

(71) Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins* **2001**, *42*, 345−54.

(72) Paschek, D.; Garcia, A. E. *Phys. Rev. Lett.* **2004**, *93*, 238105.

(73) Zagrovic, B.; Sorin, E. J.; Pande, V. *J. Mol. Biol.* **2001**, *313*, 151−69.

(74) Zhou, R.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931−6.

(75) Wu, X.; Wang, S.; Brooks, B. R. *J. Am. Chem. Soc.* **2002**, *124*, 5282−3.

(76) Tsai, J.; Levitt, M. *Biophys. Chem.* **2002**, *101−102*, 187−201.

(77) Zhou, Y.; Linhananta, A. *Proteins* **2002**, *47*, 154−62.

(78) Zhou, Y.; Zhang, C.; Stell, G.; Wang, J. *J. Am. Chem. Soc.* **2003**, *125*, 6300−5.

(79) Bolhuis, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12129−34.

(80) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. *J. Phys. Chem. B* **2004**, *108*, 6582−6594.

(81) Wei, G.; Mousseau, N.; Derreumaux, P. *Proteins* **2004**, *56*, 464−74.

(82) Snow, C. D.; Qiu, L.; Du, D.; Gai, F.; Hagen, S. J.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4077−82.

(83) Nguyen, P. H.; Stock, G.; Mittag, E.; Hu, C. K.; Li, M. S. *Proteins* **2005**, *61*, 795−808.

(84) Zhang, J.; Qin, M.; Wang, W. *Proteins* **2006**, *62*, 672−85.

(85) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Am. Chem. Soc.* **1999**, *121*, 2337−2338.

(86) Daggett, V.; Levitt, M. *J. Mol. Biol.* **1993**, *232*, 600−619.

(87) Blanco, F. J.; Rivas, G.; Serrano, L. *Nat. Struct. Biol.* **1994**, *1*, 584−90.

(88) Creighton, T. E. *Proteins Structures and Molecular Properties*, 2nd ed.; W. H. Freeman and Company: New York, 1993; p 507.

(89) McCallister, E. L.; Alm, E.; Baker, D. *Nat. Struct. Biol.* **2000**, *7*, 669−673.

(90) Amato, N. M.; Dale, L. K. *Proceedings of the 1999 IEEE International Conference on Robotics and Automation (ICRA '99)*; 1999; pp 688−694.

CT0502054

JCTC Journal of Chemical Theory and Computation

Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations

John D. Chodera,*,† William C. Swope,‡ Jed W. Pitera,‡ Chaok Seok,§ and Ken A. Dill||

Graduate Group in Biophysics and Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, California 94143, IBM Almaden Research Center, 650 Harry Road, San Jose, California 95120, and Department of Chemistry, College of Natural Sciences, Seoul National University, Gwanak-gu, Shillim-dong, san 56-1 Seoul 151-747, Republic of Korea

Received November 21, 2005

Abstract: The growing adoption of generalized-ensemble algorithms for biomolecular simulation has resulted in a resurgence in the use of the weighted histogram analysis method (WHAM) to make use of all data generated by these simulations. Unfortunately, the original presentation of WHAM by Kumar et al. is not directly applicable to data generated by these methods. WHAM was originally formulated to combine data from independent samplings of the *canonical* ensemble, whereas many generalized-ensemble algorithms sample from mixtures of canonical ensembles at different temperatures. Sorting configurations generated from a parallel tempering simulation by temperature obscures the temporal correlation in the data and results in an improper treatment of the statistical uncertainties used in constructing the estimate of the density of states. Here we present variants of WHAM, STWHAM and PTWHAM, derived with the same set of assumptions, that can be directly applied to several generalized ensemble algorithms, including simulated tempering, parallel tempering (better known as replica-exchange among temperatures), and replica-exchange simulated tempering. We present methods that explicitly capture the considerable temporal correlation in sequentially generated configurations using autocorrelation analysis. This allows estimation of the statistical uncertainty in WHAM estimates of expectations for the canonical ensemble. We test the method with a one-dimensional model system and then apply it to the estimation of potentials of mean force from parallel tempering simulations of the alanine dipeptide in both implicit and explicit solvent.

1. Introduction

The difficulty of computing equilibrium averages for complex systems such as solvated biopolymers by Monte Carlo or molecular dynamics simulation is well-known. Numerous minima and large free-energy barriers tend to slow exploration in phase space and trap the simulation in metastable regions of configuration space. This hampers the ability of the system both to equilibrate (reach the thermodynamically relevant region of phase space) and to sample sufficiently for estimates of ensemble averages to converge (reduce the statistical uncertainty in the estimate to an acceptable level) in finite computer time.

The emergence of a new class of simulation algorithms, termed *generalized-ensemble* algorithms,[1] has helped to

* Corresponding author e-mail: jchodera@gmail.com.
† Graduate Group in Biophysics, University of California at San Francisco.
‡ IBM Almaden Research Center.
§ Seoul National University.
|| Department of Pharmaceutical Chemistry, University of California at San Francisco.

10.1021/ct0502864 CCC: $37.00
Published on Web 11/09/2006

mitigate these problems. In a generalized-ensemble simulation, the probability distribution from which conformations are sampled is altered from a canonical distribution to one that will induce a broader sampling of the potential energy. Proper application should in principle allow the system to overcome energetic barriers and sample configuration space more thoroughly, at the expense of spending more time in high-energy regions that may be irrelevant at the temperature of interest. The particular method by which sampling is enhanced depends on the algorithm. In the *multicanonical algorithm* (MUCA),[5−9] conformations are sampled with a probability proportional to an approximation of the inverse potential energy density of states in an attempt to produce a random walk in the potential energy. In *simulated tempering* (ST),[10,11,3] a random walk between canonical ensembles at different temperatures is used to produce a random walk in energy, but an estimate of the free energy as a function of temperature is needed as input to ensure equal visitation of all temperatures. *Parallel tempering* (PT), a special case of the replica-exchange method (REM),[12,4] eliminates the need to know these free energies a priori by coupling temperature changes between pairs of a pool of simulated tempering simulations conducted in parallel. Several other algorithms and combinations thereof have also been proposed.[3,13−15]

In several of these algorithms, such as simulated tempering and parallel tempering, each replica generates configurations from a *mixed-canonical* distribution (a term coined in ref 16)—that is, a number of configurations are generated from the canonical distribution at each of several temperatures. To compute expectations over the canonical ensemble at a single temperature, either the configurations from all replicas that visit the temperature of interest must be collected and the remainder discarded (as in ref 17) or else a reweighting scheme must be used to properly weight the data generated at other temperatures. Fortunately, the weighted histogram analysis method (WHAM),[2] an extension of the single- and multiple-histogram methods introduced by Ferrenberg and Swendsen,[18,19] allows configurations generated from independent canonical simulations at different temperatures to be reweighted to compute expectations from the canonical ensemble at any temperature of interest. Okamoto and coworkers have applied this method to both replica-exchange simulated tempering (REST)[3] and parallel tempering[4] methods by reordering sampled configurations into pseudotrajectories, grouping configurations generated at a particular temperature together regardless of which replica they came from. Unfortunately, this permutation obscures the correlation among the stored configurations, causing the apparent correlation times for each pseudotrajectory to appear artificially shorter than the true correlation times within the independent replica trajectories. The permutation also introduces correlation *between* the pseudotrajectories, which is problematic because WHAM as presented in ref 2 is constructed to operate on *independent* canonical trajectories. Additionally, it is difficult to estimate the statistical uncertainty in the resulting estimate of the expectation from these pseudotrajectories, since standard autocorrelation analysis techniques[20−23] can no longer be applied. Recently, Gallicchio et al.[24] have described a new method for computing expectations and uncertainties from canonical simulations at different temperatures based on Bayesian inference. While Bayesian approaches are usually superior to those based on first-order Taylor expansion methods for the propagation of uncertainties (of the sort we describe in this work), they are less suitable for treating highly correlated measurements where the functional form of the correlation is essentially unknown.

Here, we derive variants of WHAM that operate on replica trajectories that are not reordered or collected by temperature. It should be noted that even if simulation data have been stored to disk sorted by temperature, they can be permuted back to the original replica trajectories to perform the proposed analyses if information about the replica-to-temperature mapping or swapping was stored. Our presentation takes a careful approach to the correlation times involved, and we show under which conditions the almost universally omitted statistical inefficiency term that appears in all formulations of WHAM-like methods can be properly neglected. Finally, we show how the statistical uncertainty in the estimator for the configuration space average for some observable can be estimated by considering the effect of temporal correlation. The method is simple and inexpensive enough to employ in all cases where WHAM is used, and we hope all researchers using WHAM will report these statistical uncertainties in the future to assess both the significance and the degree of reproducibility of results from simulations.

This paper is organized as follows: In section 2, we present a derivation of the Kumar et al. WHAM for independent simulations sampling from the canonical ensemble. Careful attention is paid to the proper treatment of time correlation in estimating the statistical uncertainty in the histograms and the resulting estimator for the expectation, and a novel way of obtaining estimates for multiple observables is presented. In section 3, we derive analogues of the method for treating simulated and parallel tempering simulations, STWHAM and PTWHAM, while properly capturing the correlations among sequential configurations. In section 4, we validate our uncertainty estimates in a one-dimensional model system and demonstrate an application for biomolecular systems by estimating the potential of mean force and corresponding uncertainties from parallel tempering simulations of alanine dipeptide in implicit and explicit solvent. An illustrative efficient implementation of the method in Fortran 95 for use in the analysis of simulated and parallel tempering simulations can be found in the Supporting Information.

2. Independent Canonical Simulations

In this section, we review the derivation of WHAM for computing expectations from multiple independent simulations in the canonical ensemble. Conducting independent simulations at the same or different temperatures can reduce statistical uncertainty while obtaining perfect parallelism (after the initial time to reach equilibrium has been discarded). Some of these simulations might be conducted at a higher temperature than the temperature of interest to

promote greater sampling across barriers, for example. Sometimes, the expectation value of one or more observables is desired over a range of temperatures. Additionally, simulations started from different initial conditions can be used as a check of equilibration and convergence.[25] Below, we follow roughly the same approach as Kumar et al.[2] in deriving the WHAM equations, though our notation differs substantially, and we include a more detailed treatment of statistical uncertainty. Additionally, we arrive at a novel way of computing expectations of multiple observables and avoid the use of many-dimensional histograms. While the method presented in ref 2 has the full generality of treating simulations conducted with arbitrary biasing potentials, we focus on the case of independent canonical simulations at different temperatures, since variations on this approach will allow us to consider simulated and parallel tempering simulations in section 3. (For an informative treatment of the case of a multiple biasing potentials at a single temperature, as in the case of umbrella sampling, see ref 26.)

2.1. Motivation and Definitions. Suppose we have an observable A that is only a function of the Cartesian coordinates of the system $\mathbf{q}$, and we wish to estimate the expectation of A over the canonical ensemble at some temperature of interest T. Instead of this temperature T, we will generally refer to its corresponding inverse temperature $\beta = (k_B T)^{-1}$, where k_B is the Boltzmann constant. We denote the expectation of A over the canonical ensemble at inverse temperature β by $\langle A \rangle_\beta$, which can be written as

$$\langle A \rangle_\beta = \frac{\int d\mathbf{q}\, e^{-\beta U(\mathbf{q})} A(\mathbf{q})}{\int d\mathbf{q}\, e^{-\beta U(\mathbf{q})}} \tag{1}$$

where $U(\mathbf{q})$ is the potential energy function of the system.

Further suppose we have carried out K independent simulations that sample from the canonical ensemble (using such techniques as Metropolis Monte Carlo[33] or thermally controlled molecular dynamics) at corresponding inverse temperatures β_1, β_2, ..., β_K, some or all of which may be different from the temperature of interest. We denote the coordinates and potential energies sampled at a fixed time interval Δt from simulation k by the time series $\{\mathbf{q}_{kn}, U_{kn}\}_{n=1}^{N_k}$, where $U_{kn} = U(\mathbf{q}_{kn})$ and N_k is the number of configurations collected from simulation k.

We first consider the probability density function from which the configurations are generated in simulation k. For a simulation sampling from the canonical distribution, the probability of generating a configuration with potential energy in the interval dU about U at inverse temperature β is given by

$$p(U|\beta)\, dU = [Z(\beta)]^{-1}\, \Omega(U)\, dU\, e^{-\beta U} \tag{2}$$

with the normalizing constant $Z(\beta)$, often referred to as the *configurational partition function*, chosen to ensure that $p(U)$ integrates to unity. The quantity $\Omega(U)$ is the *potential energy density of states*, and $\Omega(U)\, dU$ represents the volume of configuration space with potential energy in the interval dU around U.

While the Boltzmann factor $e^{-\beta_k U}$ and normalization constant $Z(\beta_k)$ differ for each simulation k, the density of states $\Omega(U)$ is independent of temperature. Since the Boltzmann factor is a known function and the configurational partition function is simply a normalizing constant, knowledge of the density of states allows the potential energy probability density to be computed at *any* temperature. If the average of the observable A over all configurations with potential energy U is known, these can be combined to give the expectation at a desired inverse temperature β

$$\langle A \rangle_\beta = \frac{\int dU\, \Omega(U)\, e^{-\beta U} A(U)}{\int dU\, \Omega(U)\, e^{-\beta U}} \tag{3}$$

where $A(U)$ is defined as the average of A over all configurations with potential energy U

$$A(U') \equiv \frac{\int d\mathbf{q}\, \delta(U(\mathbf{q}) - U')\, A(\mathbf{q})}{\int d\mathbf{q}\, \delta(U(\mathbf{q}) - U')} \tag{4}$$

It is easily seen that substituting this expression into eq 3 recovers the configuration space average in eq 1.

Our aim is to obtain the best estimate of the density of states and the expectation of the observable by combining information from several simulations. Since each simulation samples an energy range determined by its temperature, our final estimate of the density of states will be more accurate if we account for the different uncertainties in the estimate obtained from each simulation. We will therefore need a separate estimate of the density of states and its corresponding uncertainty from *each* simulation.

2.2. Obtaining an Estimate of the Density of States from Each Simulation. To obtain an estimate of the density of states from each simulation, we first need a way of mathematically expressing the form of the observed probability density function $p(U)$. While it may be possible to assume a particular functional form for this density, this would generally be inexact. A better approach is to use a *nonparametric density estimator* (see, for example, ref 27 for an overview) that makes no prior assumptions as to the true functional form of $p(U)$. Kumar et al.,[2] as Ferrenberg and Swendsen[18,19] earlier, chose a histogram-based estimator, in which the range of sampled energies is discretized into a set of nonoverlapping bins of equal width. While there are a number of more sophisticated smooth nonparametric estimators,[28] the histogram estimator is simpler and more efficient to apply.

Accordingly, we construct an estimate of the probability density function $p(U|\beta)$ on a set of M points, labeled U_m, that span the sampled potential energy range and are spaced ΔU apart. We denote the estimate of $p(U|\beta)$ at U_m by $p_m(\beta)$and the corresponding estimate of the density of states $\Omega(U_m)$ by Ω_m.

$$p_m(\beta) \equiv p(U_m|\beta) = [Z(\beta)]^{-1} \Omega_m e^{-\beta U_m} \tag{5}$$

The normalization factor $Z(\beta)$ can then be approximated by a discretized integration

$$Z(\beta) = \int dU\, \Omega(U)\, e^{-\beta U} \approx \sum_{m=1}^{M} \Delta U\, \Omega_m\, e^{-\beta U_m} \qquad (6)$$

where we have made the assumption that the integrand, $\Omega(U)\, e^{-\beta U}$, does not change significantly over the bin width ΔU. As the density of states $\Omega(U)$ increases with U and the Boltzmann factor $e^{-\beta U}$ decreases, their product is expected to vary less rapidly than either term individually.

We define $\psi_m(U)$ as the indicator or characteristic function for the energy bin of width ΔU centered about U_m

$$\psi_m(U) = \begin{cases} 1 & \text{if } U \in [U_m - \Delta U/2,\, U_m + \Delta U/2) \\ 0 & \text{otherwise} \end{cases} \qquad (7)$$

and the time series defined by this indicator function as $\{\psi_{mkn}\}_{n=1}^{N_k}$, where $\psi_{mkn} \equiv \psi_m(U_{kn})$. We denote the count of configurations from simulation k that fall in energy bin m—the "histogram" from which the weighted histogram analysis method derives its name—by H_{mk}, and see that it can be computed by

$$H_{mk} = \sum_{n=1}^{N_k} \psi_{mkn} \qquad (8)$$

We will also use the total number of configurations over all simulations that fall in energy bin m, which we term H_m:

$$H_m = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \psi_{mkn} \qquad (9)$$

Note that throughout our discussion, pairs of variables that only differ by the number of written subscripts, such as H_m and H_{mk}, represent similar quantities related in this way.

We can estimate $p_m(\beta_k)$ by the number of configurations sampled from the simulation at temperature β_k with energies that fall in the bin centered about U_m:

$$p_m(\beta_k) \approx \frac{1}{\Delta U} \cdot \frac{H_{mk}}{N_k} \qquad (10)$$

Equating this with the definition of p_m from eq 5 and rearranging terms, we can obtain an estimate of Ω_{mk}, the density of states at energy U_m from simulation k, which we will denote by $\hat{\Omega}_{mk}$:

$$\hat{\Omega}_{mk} = \frac{1}{\Delta U} \cdot \frac{H_{mk}}{N_k} \cdot \frac{Z(\beta_k)}{e^{-\beta_k U_m}}$$

$$= \frac{H_{mk}}{N_k\, \Delta U \exp[f_k - \beta_k U_m]} \qquad (11)$$

In the last step, we have replaced the partition function $Z(\beta_k)$ by an exponentiated dimensionless free energy $f_k \equiv -\ln Z(\beta_k)$. Each independent simulation k contributes an estimate of the density of states $\hat{\Omega}_{mk}$ for energy bin m. Each of these estimates in turn carries a statistical uncertainty $\delta^2\hat{\Omega}_{mk}$, determined primarily by the number of uncorrelated samples of the energy bin. (Expressions for $\delta^2\hat{\Omega}_{mk}$ will be derived later in section 2.5.) We will combine these individual estimates $\hat{\Omega}_{mk}$ to produce a single optimal estimator $\hat{\Omega}_m$ in a such way that the statistical uncertainty in the resulting estimate is minimized, giving more weight to the $\hat{\Omega}_{mk}$ with smaller uncertainties. To do this, we must first briefly review the maximum-likelihood method for combining independent measurements with associated uncertainties into an optimal estimate and also consider the uncertainty in a mean computed from a set of correlated observations.

2.3. Optimal Estimator from Independent Observations and Associated Uncertainties. Suppose we have K independent observations or measurements of some random variable X denoted $x_1,...,x_K$, each with corresponding squared uncertainty $\delta^2 x_k$, defined by

$$\delta^2 x_k \equiv \langle (x_k - \langle x_k \rangle)^2 \rangle = \langle x_k^{\,2} \rangle - \langle x_k \rangle^2 \qquad (12)$$

where $\langle \cdot \rangle$ here denotes the expectation over repeated measurements or experimental trials. We can then write $\hat{X}$, the optimal estimator for $\langle X \rangle$ in the sense of minimizing $\delta^2\hat{X}$, by a weighted sum of the individual estimates

$$\hat{X} = \frac{\sum_{k=1}^{K} [\delta^2 x_k]^{-1}\, x_k}{\sum_{k=1}^{K} [\delta^2 x_k]^{-1}} \qquad (13)$$

Note that observations with smaller uncertainties get greater weight, and if all the uncertainties are equal, the weight is simply $1/K$, as would be expected.

The uncertainty in the resulting estimate is simply given by

$$\delta^2\hat{X} = \{\sum_{k=1}^{K} [\delta^2 x_k]^{-1}\}^{-1} \qquad (14)$$

These are standard formulas that come from maximum likelihood considerations.[29]

2.4. Statistical Uncertainty in the Estimator for Correlated Time Series Data. We briefly review the estimation of statistical uncertainty for a time series of correlated measurements. (See ref 20 for an early exposition of this method as applied to the analysis of Monte Carlo simulations of spin systems, ref 21 for the analysis of molecular dynamics simulations, or ref 23 for a recent general illustration.)

Suppose we have a time series of correlated sequential observations of the random variable X denoted $\{x_n\}_{n=1}^{N}$ that come from a stationary, time-reversible stochastic process. Our estimate for the expectation of X is given by the time average

$$\hat{X} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad (15)$$

but the statistical uncertainty is more complicated than in the independent observation case

$$\delta^2\hat{X} \equiv \langle (\hat{X} - \langle \hat{X} \rangle)^2 \rangle = \langle \hat{X}^2 \rangle - \langle \hat{X} \rangle^2$$

$$= \frac{1}{N^2}\sum_{n,n'=1}^{N} [\langle x_n x_{n'}\rangle - \langle x_n\rangle\langle x_{n'}\rangle]$$

$$= \frac{1}{N^2}\sum_{n=1}^{N} [\langle x_n{}^2\rangle - \langle x_n\rangle^2] + \frac{1}{N^2}\sum_{n\neq n'=1}^{N} [\langle x_n x_{n'}\rangle - \langle x_n\rangle\langle x_{n'}\rangle] \tag{16}$$

In the last step, we have split the sum into two sums—a term capturing the variance in the observations and a remaining term capturing the correlation between observations. Using the properties of stationarity and time-reversibility, we can further manipulate this to obtain

$$\delta^2\hat{X} = \frac{1}{N}[\langle x_n{}^2\rangle - \langle x_n\rangle^2] + \frac{2}{N}\sum_{t=1}^{N-1}\left(\frac{N-t}{N}\right)[\langle x_n x_{n+t}\rangle - \langle x_n\rangle\langle x_{n+t}\rangle]$$

$$\equiv \frac{\sigma_x{}^2}{N}(1+2\tau) = \frac{\sigma_x{}^2}{N/g} \tag{17}$$

where the variance $\sigma_x{}^2$, statistical inefficiency g, and autocorrelation time τ (in units of the sampling interval) are given by

$$\sigma_x{}^2 \equiv \langle x_n{}^2\rangle - \langle x_n\rangle^2 \tag{18}$$

$$\tau \equiv \sum_{t=1}^{N-1}\left(1-\frac{t}{N}\right)C_t \tag{19}$$

$$g \equiv 1 + 2\tau \tag{20}$$

with the discrete-time normalized fluctuation autocorrelation function C_t defined as

$$C_t \equiv \frac{\langle x_n x_{n+t}\rangle - \langle x_n\rangle^2}{\langle x_n{}^2\rangle - \langle x_n\rangle^2} \tag{21}$$

The quantity $g \equiv (1 + 2\tau) \geq 1$ can be thought of as a *statistical inefficiency*, in that N/g gives the effective number of *uncorrelated* configurations contained in the time series. The statistical inefficiency will depend on the time interval at which configurations are collected for analysis; longer intervals will reduce the statistical inefficiency, which will approach unity as the sampling interval exceeds the correlation time. Practically, we use our best estimates for the variance $\sigma_x{}^2$ and autocorrelation function C_t to compute an estimate of the statistical uncertainty $\delta^2\hat{X}$.

2.5. Optimal Estimate of the Density of States. We now construct an optimal estimator of the density of states Ω_m from the individual estimates obtained from the K independent canonical simulations. From the results of section 2.3,we can write this estimator and its corresponding uncertainty as

$$\hat{\Omega}_m = \frac{\sum_{k=1}^{K}[\delta^2\hat{\Omega}_{mk}]^{-1}\,\hat{\Omega}_{mk}}{\sum_{k=1}^{K}[\delta^2\hat{\Omega}_{mk}]^{-1}} \tag{22}$$

$$\delta^2\hat{\Omega}_m = \{\sum_{k=1}^{K}[\delta^2\hat{\Omega}_{mk}]^{-1}\}^{-1} \tag{23}$$

The results of section 2.4 show us how to write the $\delta^2\hat{\Omega}_{mk}$, the uncertainty in our estimate of the density of states for energy bin m from simulation k. In eq 11 above, we see that this uncertainty comes only from $\delta^2 H_{mk}$, the uncertainty in the histogram count for the energy bin, since all other terms are known with certainty:

$$\delta^2\hat{\Omega}_{mk} = \frac{\delta^2 H_{mk}}{\{N_k\,\Delta U\exp[f_k - \beta_k U_m]\}^2} \tag{24}$$

H_{mk}, the histogram count from simulation k, can be written as a time average of the indicator function ψ_m over the correlated configurations collected from the simulation:

$$H_{mk} = N_k \cdot \frac{1}{N_k}\sum_{n=1}^{N_k}\psi_{mkn} \tag{25}$$

We can use the result of section 2.4 above to obtain an expression for $\delta^2 H_{mk}$, the uncertainty in the histogram count:

$$\delta^2 H_{mk} = N_k{}^2 \cdot \frac{\sigma^2{}_{mk}}{N_k/g_{mk}}$$

$$= g_{mk}\,N_k\,(\langle\psi_{mk}{}^2\rangle - \langle\psi_{mk}\rangle^2)$$

$$= g_{mk}\,N_k\langle\psi_{mk}\rangle(1 - \langle\psi_{mk}\rangle)$$

$$= g_{mk}\,\langle H_{mk}\rangle\left(1 - \frac{\langle H_{mk}\rangle}{N_k}\right) \tag{26}$$

where, because $\psi_m(U)$ is an indicator function (eq 7), $[\psi_m(U)]^2 = \psi_m(U)$. If the histograms are sparsely populated, a reasonable assumption if there are a sufficient number of histogram bins spanning the energy range sampled by each simulation, then $\langle H_{mk}\rangle/N_k \ll 1$, and we can further simplify this to

$$\delta^2 H_{mk} \approx g_{mk}\,\langle H_{mk}\rangle \tag{27}$$

The statistical inefficiency g_{mk} here reflects the number of configurations required for an uncorrelated sampling of the energy bin. This will, in general, depend on the bin index, bin width, and temperature. This dependence was omitted in the original Kumar et al. presentation.[2] At higher temperatures, the correlation time, and hence the statistical inefficiency, is expected to be smaller as the simulation can move through configuration space more easily. The structure of the energy landscape may cause the simulation to be stuck in certain regions of configuration space for different times, hence the dependence on energy bin index is also potentially important.

The expectation $\langle H_{mk}\rangle$ should be replaced by our best estimate of the histogram count for energy bin m at temperature β_k, which could be obtained from our yet-to-be-determined optimal estimate of the density of states $\hat{\Omega}_m$:

$$\langle H_{mk}\rangle = N_k\, p_m(\beta_k)\, \Delta U$$
$$\approx N_k\, \Delta U\, \hat{\Omega}_m \exp[f_k - \beta_k U_m] \tag{28}$$

Substituting this expression back into eqs 27 and 24, we obtain

$$\delta^2\hat{\Omega}_{mk} = \frac{g_{mk}\, N_k\, \Delta U\, \hat{\Omega}_m \exp[f_k - \beta_k U_m]}{\{N_k\, \Delta U \exp[f_k - \beta_k U_m]\}^2}$$
$$= \frac{\hat{\Omega}_m}{g_{mk}{}^{-1}\, N_k\, \Delta U \exp[f_k - \beta_k U_m]} \tag{29}$$

Using eq 29 for the uncertainty in the density of states associated with simulation k and eq 22 for the best estimate of the density of states, along with eq 11, we obtain

$$\hat{\Omega}_m = \frac{\sum_{k=1}^{K} g_{mk}{}^{-1} H_{mk}}{\sum_{k=1}^{K} g_{mk}{}^{-1}\, N_k\, \Delta U \exp[f_k - \beta_k U_m]} \tag{30}$$

Everything in the above expression can be easily evaluated, except for the f_k, which depends on the $\hat{\Omega}_m$ through

$$f_k = -\ln \sum_{m=1}^{M} \hat{\Omega}_m\, \Delta U\, e^{-\beta_k U_m} \tag{31}$$

The f_k may therefore be solved for self-consistency by iteration of eqs 30 and 31 starting from an arbitrary choice, such as $f_k = 0$.

The statistical uncertainty in $\hat{\Omega}_m$ is given by eq 14:

$$\delta^2\hat{\Omega}_m = \frac{\hat{\Omega}_m}{\sum_{k=1}^{K} g_{mk}{}^{-1}\, N_k\, \Delta U \exp[f_k - \beta_k U_m]} \tag{32}$$

We note that the relative uncertainty in this estimate is given by

$$\frac{\delta^2\hat{\Omega}_m}{\hat{\Omega}_m^2} = [\sum_{k=1}^{K} g_{mk}{}^{-1}\, H_{mk}]^{-1} \tag{33}$$

which is equal to $H_m{}^{-1}$, the inverse of the total number of configurations from all simulations in energy bin m, if all g_{mk} are unity. This is reasonable, since the uncertainty in our estimate for Ω_m should diminish as more independent samples are collected in energy bin m.

2.6. Estimating an Observable at the Temperature of Interest. Using the estimate of the density of states obtained above, we can obtain an estimate for the expectation of any configuration function $A(\mathbf{q})$ at an arbitrary temperature by writing analogous equations to eqs 3 and 4 where we have discretized the energy U

$$\langle A\rangle_\beta \approx \frac{\sum_{m=1}^{M} \hat{\Omega}_m\, \Delta U\, e^{-\beta U_m} A_m}{\sum_{m=1}^{M} \hat{\Omega}_m\, \Delta U\, e^{-\beta U_m}} \tag{34}$$

where

$$A_m = \frac{\int d\mathbf{q}\, A(\mathbf{q})\, \psi_m(U(\mathbf{q}))}{\int d\mathbf{q}\, \psi_m(U(\mathbf{q}))} \tag{35}$$

A_m, the mean of observable A over all configurations with potential energies consistent with energy bin m, can be best approximated by pooling configurations from *all* K simulations that have energies in bin m

$$\hat{A}_m = H_m{}^{-1} \sum_{k=1}^{K} \sum_{n=1}^{N_k} \psi_{mkn}\, A_{kn} \tag{36}$$

where H_m (eq 9) is the total count of configurations in energy bin m from all simulations. Substituting this expression for A_m into eq 34 above produces an estimator $\hat{A}(\beta)$ for $\langle A\rangle_\beta$

$$\hat{A}(\beta) = \frac{\sum_{m=1}^{M} \hat{\Omega}_m\, \Delta U\, e^{-\beta U_m} A_m}{\sum_{m=1}^{M} \hat{\Omega}_m\, \Delta U\, e^{-\beta U_m}}$$
$$= \frac{\sum_{m=1}^{M} \hat{\Omega}_m\, e^{-\beta U_m} [H_m{}^{-1} \sum_{k=1}^{K} \sum_{n=1}^{N_k} \psi_{mkn}\, A_{kn}]}{\sum_{m=1}^{M} \hat{\Omega}_m\, e^{-\beta U_m} [H_m{}^{-1} \sum_{k=1}^{K} \sum_{n=1}^{N_k} \psi_{mkn}]}$$
$$= \frac{\sum_{k=1}^{K} \sum_{n=1}^{N_k} w_{kn}(\beta)\, A_{kn}}{\sum_{k=1}^{K} \sum_{n=1}^{N_k} w_{kn}(\beta)} \tag{37}$$

where we have defined the per-configuration weights $w_{kn}(\beta)$ by

$$w_{kn}(\beta) = \sum_{m=1}^{M} \psi_{mkn}\, H_m{}^{-1} \hat{\Omega}_m\, e^{-\beta U_m} \tag{38}$$

where only one term of the sum will contribute—the bin m containing the energy U_{kn}—due to the presence of the indicator function ψ_{mkn}. Note that we only need to compute the weight w_{kn} up to a constant of proportionality because this constant drops out in the normalized sum in eq 37.

This relationship is significant in that we now have an expression for the canonical expectation of observable A in terms of a weighted sum over *all* of the data. These weights are determined by the temperature of interest from the WHAM equations and are simple functions of the count of

configurations with energies falling in a particular energy bin. The weights $w_{kn}(\beta)$ can be computed once for the temperature of interest and then used to calculate expectations of many observables.

It should be noted that our estimate of $\langle A \rangle_\beta$ will only be reasonable if the inverse temperature of interest β lies near or within the range of inverse temperatures sampled by the canonical simulations—the uncertainty in the estimate will increase as the temperature of interest deviates from the sampled range of temperatures (see refs 30 and 31 for an examination of this issue).

2.7. Statistical Uncertainty of the Estimator for the Expectation. If the observable of interest has a long correlation time compared to fluctuations in the potential energy (e.g., if the observable is a function of the large scale molecular conformation), then it is possible that the density of states Ω_m and dimensionless free energies $\{f_k\}$ may be sufficiently well-converged that they are not dominant contributors to the uncertainty in the estimate of the observable of interest. Instead, the long time-correlation in the observable means that there are many fewer effectively independent observations of the observable than stored configurations. We may then use the following procedure.

We can rewrite the estimator $\hat{A}_\beta$ as a ratio of two random quantities X and Y

$$\hat{A}_\beta = \frac{\hat{X}}{\hat{Y}} \tag{39}$$

where

$$\hat{X} \equiv \sum_{k=1}^{K} \sum_{n=1}^{N_k} w_{kn}(\beta) A_{kn}$$

$$\hat{Y} \equiv \sum_{k=1}^{K} \sum_{n=1}^{N_k} w_{kn}(\beta) \tag{40}$$

Applying standard error propagation techniques for a function of random variables (see, e.g. ref 32), which amounts to a first-order Taylor series expansion of $\hat{A}$ about $\langle \hat{X} \rangle / \langle \hat{Y} \rangle$, we can estimate the uncertainty in $\hat{A}$ as

$$\delta^2 \hat{A} = \left[\frac{\hat{X}}{\hat{Y}}\right]^2 \left[\frac{\delta^2 \hat{X}}{\hat{X}^2} + \frac{\delta^2 \hat{Y}}{\hat{Y}^2} - 2\frac{\delta \hat{X} \delta \hat{Y}}{\hat{X}\,\hat{Y}}\right] \tag{41}$$

Here, the cross-term $\delta \hat{X} \delta \hat{Y} \equiv \langle (\hat{X} - \langle \hat{X} \rangle)(\hat{Y} - \langle \hat{Y} \rangle) \rangle$ is non-zero only if the random variables X and Y are correlated, in which case the term involving it in the equation above serves to reduce the uncertainty in the estimate of the ratio $\hat{A}$.

Recognizing that X and Y include contributions from K statistically independent simulations, we can collect these terms and write (dropping the hats)

$$X \equiv \sum_{k=1}^{K} N_k X_k$$

$$Y \equiv \sum_{k=1}^{K} N_k Y_k$$

$$X_k \equiv \frac{1}{N_k} \sum_{n=1}^{N_k} w_{kn} A_{kn}$$

$$Y_k \equiv \frac{1}{N_k} \sum_{n=1}^{N_k} w_{kn} \tag{42}$$

where the argument β has been omitted for notational convenience. Because the K individual simulations are *independent*, the uncertainties required in eq 41 are given by

$$\delta^2 X = \sum_{k=1}^{K} N_k^2 \delta^2 X_k$$

$$\delta^2 Y = \sum_{k=1}^{K} N_k^2 \delta^2 Y_k$$

$$\delta X \delta Y = \sum_{k=1}^{K} N_k^2 \, \delta X_k \delta Y_k \tag{43}$$

These uncertainties involve the correlated data of simulation k and can be estimated by standard correlation analysis methods[21,23] or by block transformation methods,[22] though the latter method requires some modification to estimate the uncertainty cross-term $\delta X_k \delta Y_k$.

To compute the uncertainties by correlation analysis methods as in section 2.4, we first define new observables $x_{kn} = w_{kn} A_{kn}$ and $y_{kn} = w_{kn}$ and compute the uncertainties

$$\delta^2 X_k = \frac{\sigma^2_{k,x;x}}{N_k / g_{k,x;x}}$$

$$\delta^2 Y_k = \frac{\sigma^2_{k,y;y}}{N_k / g_{k,y;y}}$$

$$\delta X_k \delta Y_k = \frac{\sigma^2_{k,x;y}}{N_k / g_{k,x;y}} \tag{44}$$

These uncertainties involve (co)variances of the type $\sigma^2_{k,x;y}$, estimated for each replica by

$$\hat{\sigma}_{k,x;y}{}^2 = \frac{1}{N_k - 1} \sum_{n=1}^{N_k} (x_{kn} - \hat{X}_k)(y_{kn} - \hat{Y}_k) \tag{45}$$

The statistical inefficiencies of the form $g_{k,x;y}$ are computed by

$$g_{k,x;y} \equiv 1 + 2\tau_{k,x;y} \tag{46}$$

$$\tau_{k,x;y} \equiv \sum_{t=1}^{N_k - 1} \left(1 - \frac{t}{N_k}\right) C_{kt,x;y} \tag{47}$$

with the correlation function for simulation k computed by taking advantage of stationarity and time-reversibility:

$$C_{kt,x;y} \approx \frac{1}{2\hat{\sigma}_{k,x;y}{}^2}\frac{1}{(N_k - t)} \times \sum_{n=1}^{N_k - t}[(x_{kn} - \hat{X}_k)(y_{kn+t} - \hat{Y}_k) + (y_{kn} - \hat{Y}_k)(x_{kn+t} - \hat{X}_k)] \quad (48)$$

See section 5.2 for a discussion on efficiently computing the integrated correlation time τ from $\hat{C}_{kt;x,y}$.

3. Simulated and Parallel Tempering

3.1. WHAM for Simulated Tempering (STWHAM). In a simulated tempering simulation,[10,3] a single simulation is conducted in which configurations are sampled from a *mixed-canonical* ensemble.[16] In practice, a simulation algorithm that samples from the canonical ensemble is used to generate configurations, and at regular intervals attempts are made to change the temperature among a discrete set of choices $\beta_1,...,\beta_L$. The probability of accepting a proposed temperature change is given by the Metropolis-like criterion

$$P(\beta_l \rightarrow \beta_{l'}) = \min\{1, \exp[-(\beta_{l'} - \beta_l)U + (a_{l'} - a_l)]\} \quad (49)$$

where the constants $\{a_l\}_{l=1}^L$ are specified beforehand and chosen, often by tedious exploratory simulations, to attempt to achieve near-equal visitation of each temperature and, hopefully, potential energy. The optimal choice of $\{a_l\}$ is given by the dimensionless free energies $\{f_k\}$ in the equation above, and proposed temperature changes are usually between neighboring temperatures because the exchange probability diminishes with increased temperature separation. Use of the above criterion for accepting or rejecting proposed temperature changes ensures that, if the configurations were originally distributed from the equilibrium distribution at the old temperature, they are also distributed from the canonical distribution at the new temperature.

As a result of this procedure, the system spends a fraction of time in each of a number of different temperatures. Since we know the number of times each temperature was visited, we can write the probability density for energy bin m as a weighted sum of the canonical probability density functions at these different temperatures

$$p_m = \sum_{l=1}^{L}\frac{N_l}{N}\cdot\frac{\Omega_m e^{-\beta_l U_m}}{Z(\beta_l)} \quad (50)$$

where N_l/N is the fraction of configurations generated at inverse temperature β_l over the course of the simulation. As above, we introduce the Helmholtz free energy $f_l \equiv -\ln Z(\beta_l)$, which allows us to write

$$p_m = \sum_{l=1}^{L}\frac{N_l}{N}\Omega_m \exp[f_l - \beta_l U_m] = \Omega_m \sum_{l=1}^{L}(N_l/N)\exp[f_l - \beta_l U_m] \quad (51)$$

We can approximate p_m as before using our histogram count, H_m, the number of configurations with potential energy in the bin centered about U_m:

$$p_m \approx \frac{1}{\Delta U}\cdot\frac{H_m}{N} \quad (52)$$

Rearranging and including our definition of f_l, we obtain the coupled set of equations for estimating the density of states

$$\hat{\Omega}_m = \frac{H_m}{\sum_{l=1}^{L} N_l \Delta U \exp[f_l - \beta_l U_m]} \quad (53)$$

$$f_l = -\ln\sum_{m=1}^{M}\hat{\Omega}_m \Delta U e^{-\beta_l U_m} \quad (54)$$

These equations are similar to eqs 30 and 31 for the canonical ensemble WHAM if the configurations are grouped by the temperature at which they were generated but lacking statistical inefficiency terms since we are not combining data from multiple simulations.

The uncertainty in $\hat{\Omega}_m$ is then given by

$$\delta^2\hat{\Omega}_m = \frac{\delta^2 H_m}{\{\sum_{l=1}^{L} N_l \Delta U \exp[f_l - \beta_l U_m]\}^2} = \frac{g_m \langle H_m\rangle}{\{\sum_{l=1}^{L} N_l \Delta U \exp[f_l - \beta_l U_m]\}^2} \quad (55)$$

where, as in eq 27, we assume the histograms are sparsely populated and introduce the statistical inefficiency g_m to estimate the histogram uncertainty.

The estimate for the expectation of the total histogram count in energy bin m is given by the sampling probability

$$\langle H_m\rangle = N\Delta U p_m \approx N\Delta U \hat{\Omega}_m \sum_{l=1}^{L}(N_l/N)\exp[f_l - \beta_l U_m] \quad (56)$$

which gives the final estimate for the uncertainty as

$$\delta^2\hat{\Omega}_m = \frac{\hat{\Omega}_m}{\sum_{l=1}^{L} g_m{}^{-1} N_l \Delta U \exp[f_l - \beta_l U_m]} \quad (57)$$

Following the approach in section 2.6, we can again write the estimator in the form of a weighted sum over configurations

$$\hat{A}(\beta) = \frac{\sum_{n=1}^{N} w_n(\beta) A_n}{\sum_{n=1}^{N} w_n(\beta)} \quad (58)$$

$$w_n(\beta) = \sum_{m=1}^{M}\psi_{mn} H_m{}^{-1}\hat{\Omega}_m e^{-\beta U_m} \quad (59)$$

where again, only one term contributes to the sum in the expression for the weight w_n. The statistical uncertainty in this estimate, as in section 2.7, can be computed by eq 41, where X and Y are now given by

$$X \equiv \frac{1}{N}\sum_{n=1}^{N} w_n A_n$$

$$Y \equiv \frac{1}{N}\sum_{n=1}^{N} w_n \tag{60}$$

These uncertainties are simply computed as in eqs 44−48, without the subscript k as there is only one simulation instead of many. The quantities $\hat{X}$ and $\hat{Y}$ no longer correspond to canonical averages, since they are the expectations over the simulated tempering trajectory which spends a different amount of time at each of the L temperatures—it is a mixed canonical average. The sample mean over the trajectory provides the best estimator for these quantities.

Here, the statistical inefficiency g_m appearing in eq 57 and the inefficiencies required in applying eqs 44−48 are computed from the correlation functions computed over the simulated tempering trajectory, which includes unphysical jumps in temperature. It is worth noting that expressions for $\langle A \rangle_\beta$ given in formulations by Okamoto and co-workers (e.g. eq 24 of ref 14) instead contain a statistical inefficiency for each *temperature*. In principle, one could account for a temperature-dependent statistical inefficiency, since one might expect correlation times to be different at each temperature, but, in practice, the limited number of configurations sampled between temperature changes is likely too short to allow temperature-dependent correlation times to be computed. Additionally, a temperature-dependent treatment does not account for the correlation between configurations sampled before and after a temperature swap. The derivation presented here assumes the statistical inefficiency g_m depends only on the energy bin m, which causes these factors to cancel out of our estimator for $\langle A \rangle_\beta$ in eqs 58 and 59.

3.2. WHAM for Parallel Tempering or Independent Simulated Tempering Simulations (PTWHAM). In a parallel tempering (or replica-exchange among temperatures) simulation,[12,4] it was recognized that the constants a_k needed in the simulated tempering simulation to ensure equal sampling of temperatures could be eliminated if multiple simulated tempering simulations were conducted in parallel and the temperature changes of two simulations were coupled together into a temperature swap between the replicas. In practice, a number K of *replicas* are simulated independently at inverse temperatures $\beta_1,...,\beta_K$ using some simulation method that samples from the canonical distribution. At given intervals, an attempt is made to exchange the temperatures of two replicas i and j, with the exchange accepted with probability

$$\begin{aligned} P_{\mathrm{exch}} &= \min\{1, \exp[-(\beta_j - \beta_i)U_i + (a_j - a_i)]\} \times \\ &\quad \min\{1, \exp[-(\beta_i - \beta_j)U_j + (a_i - a_j)]\} = \\ &\quad \min\{1, \exp[-(\beta_j - \beta_i)(U_i - U_j)]\} \end{aligned} \tag{61}$$

where β_i is the current inverse temperature of replica i, and U_i is the corresponding potential energy.

Because of this exchange procedure, each replica executes a more or less random walk in temperature, eliminating the need to perform exploratory simulations to determine the parameters $\{a_i\}_{i=1}^{K}$ required for simulated tempering. Each replica simulation is nearly independent, as the correlation between configurations of different replicas introduced by the exchange of temperatures is minimal. The dominant contribution to statistical uncertainties will almost certainly be due to the variance and temporal correlation in the value of the observable of interest within each replica, which reduces the effective number of independent samples. We can therefore analyze a parallel tempering simulation as a set of *independent* simulated tempering simulations, each with a number L of accessible temperatures, with L equal to the number of replicas K. Below, we derive an analogue of the Kumar et al. WHAM procedure for the treatment of K independent simulated tempering simulations (replicas) each capable of visiting L temperatures, allowing this method to also treat simulations generated by procedures such as REST.[3] We make use of the sampling distribution for simulated tempering described above and properly account for the correlation within each replica, eliminating the need to artificially reorder configurations from parallel tempering simulations by temperature.

We can use the simulated tempering eqs 53 and 57 above to write the estimator and uncertainty for the density of states obtained from each replica k as

$$\hat{\Omega}_{mk} = \frac{H_{mk}}{\sum_{l=1}^{L} N_{kl} \Delta U \exp[f_l - \beta_l U_m]} \tag{62}$$

$$\delta^2 \hat{\Omega}_{mk} = \frac{\hat{\Omega}_m}{g_{mk}^{-1} \sum_{l=1}^{L} N_{kl} \Delta U \exp[f_l - \beta_l U_m]} \tag{63}$$

where we have added the index k to denote the *replica* from which the data are generated. H_{mk} therefore denotes the number of configurations sampled with potential energy in energy bin m from replica k, and N_{kl} denotes the number at temperature β_l from replica k. g_{mk} is the statistical inefficiency computed from replica k for energy bin m.

Again using the optimal combination rule of eq 13, we obtain the optimal estimate for the density of states

$$\hat{\Omega}_m = \frac{\sum_{k=1}^{K} g_{mk}^{-1} H_{mk}}{\sum_{k=1}^{K} g_{mk}^{-1} \sum_{l=1}^{L} N_{kl} \Delta U \exp[f_l - \beta_l U_m]} \tag{64}$$

and the statistical uncertainty from eq 14:

$$\delta^2\hat{\Omega}_m = \left\{ \sum_{k=1}^{K} \left[\frac{g_{mk}^{-1} \sum_{l=1}^{L} N_{kl}\, \Delta U \exp[f_l - \beta_l U_m]}{\Omega_m} \right] \right\}^{-1} = \frac{\hat{\Omega}_m}{\sum_{k=1}^{K} g_{mk}^{-1} \sum_{l=1}^{L} N_{kl}\, \Delta U \exp[f_l - \beta_l U_m]} \quad (65)$$

We can rewrite eq 64 as

$$\hat{\Omega}_m = \frac{H_m^{\mathrm{eff}}}{\sum_{l=1}^{L} N_{ml}^{\mathrm{eff}}\, \Delta U \exp[f_l - \beta_l U_m]} \quad (66)$$

where $H_m^{\mathrm{eff}} \equiv \sum_{k=1}^{K} g_{mk}^{-1} H_{mk}$ is the effective number of independent samples in energy bin m from all replicas, and $N_{ml}^{\mathrm{eff}} \equiv \sum_{k=1}^{K} g_{mk}^{-1} N_{kl}$ is an effective number of independent samples at temperature β_l from all replicas.

To compute the estimator of the expectation for an observable A, we apply the same technique in section 2.6 above and write the expectation as a weighted sum over configurations

$$\hat{A}(\beta) = \frac{\sum_{k=1}^{K} \sum_{n=1}^{N_k} w_{kn}(\beta)\, A_{kn}}{\sum_{k=1}^{K} \sum_{n=1}^{N_k} w_{kn}(\beta)} \quad (67)$$

where the weights are given by

$$w_{kn}(\beta) \equiv \sum_{m=1}^{M} \psi_{mkn}\, H_m^{-1} \hat{\Omega}_m e^{-\beta U_m} \quad (68)$$

As in eqs 38 and 59, the sum over m reduces to a single term, the one with energy bin index appropriate for configuration n of replica k. A_{kn} is the value of the observable A for configuration n of replica k and $H_m = \sum_{k=1}^{K} \sum_{n=1}^{N} \psi_{mkn}$, the total number of configurations from all replicas with potential energy in bin m.

Again, if the observable of interest has a correlation time that is long compared to fluctuations in the potential energy, we may compute the dominant contribution to the statistical uncertainty $\delta^2 A(\beta)$ by eqs 41−48, with the important distinction that k now indexes the *replicas*, rather than the temperatures. The correlation times are, as in the simulated tempering case, computed over the nonphysical replica trajectories; because the replicas perform random walks in temperature, these times are likely to be shorter than the correlation time for this observable computed from a canonical simulation at the lowest temperature. These replica correlation times properly capture the correlation between successive snapshots generated by a sampling method like Metropolis Monte Carlo or molecular dynamics, and their use in estimating the uncertainty is the primary novel result of this paper. Collecting configurations from all replicas into pseudotrajectories of constant temperature, as suggested in previous attempts to apply the method to parallel tempering simulations,[4] would give correlation times that are erroneously short and make the incorrect assumption that these pseudotrajectories are statistically independent.

4. Applications

4.1. One-Dimensional Model Potential. To validate the STWHAM and PTWHAM methods described above for estimating expectations and corresponding uncertainties, we consider a one-dimensional model potential where canonical expectations can be computed directly and a large quantity of simulation data can be obtained in order to verify our uncertainty formulas. We use an asymmetric double well potential, given by

$$U(q) = (q-1)^2(q+1)^2 + 0.1\, q \quad (69)$$

All simulations utilize the Metropolis Monte Carlo method[33] with the trial displacement Δq uniformly distributed on the interval $[-0.2, +0.2]$ to generate a series of configurations which are sampled every 10 move attempts, resulting in highly correlated data. In the following simulations, we estimate the expectation $\langle q \rangle_{\beta^*}$ at $\beta^* = 4$, where the integrated correlation time of q is rather long—approximately 130 samples. The initial conformation was chosen uniformly on the interval $[-1.8, +1.8]$ and the first 10^5 steps discarded to equilibration.

Four types of simulations were performed: a standard canonical Metropolis Monte Carlo (MMC) simulation at $\beta = \beta^* = 4$, as described above; a set of four independent canonical (4MMC) simulations with inverse temperatures β exponentially spaced in the range 1−4 ($\beta \approx \{4, 2.52, 1.59, 1\}$); a simulated tempering simulation (ST) with the same four possible temperatures and analytically computed optimal weights; and a parallel tempering (PT) simulation with replicas at the same four temperatures.

All simulations were conducted for 5×10^7 steps each (per replica, if multiple replicas are used), generating 5×10^6 samples (per replica). The data were then divided into 500 sequential blocks of 10^4 configurations (per replica) each, whose expectations were verified to be statistically independent by computing the correlation between expectations in neighboring blocks. The standard deviation of the set of expectations computed from each block is indicative of the statistical uncertainty in simulations of a single block length—10^4 samples (per replica)—and the difference between the mean of these estimates and the expectation computed from the potential directly is indicative of the bias. Expectations and uncertainties for each block were computed using the code appearing in listing 1 of the Supporting Information.

To assess the performance of the uncertainty estimate for each block, we compute the fraction of blocks for which the true magnitude of the deviation from the mean of the block expectations is smaller than a multiplicative constant σ times the estimated uncertainties, for $\sigma \in [0.1, 3]$. This fraction is related to a confidence interval if compared to the error function Gaussian integral (Figure 1). For example, for our uncertainty estimates to be meaningful, we expect the difference between the true mean and our estimate to be within one standard deviation σ approximately 66% of the

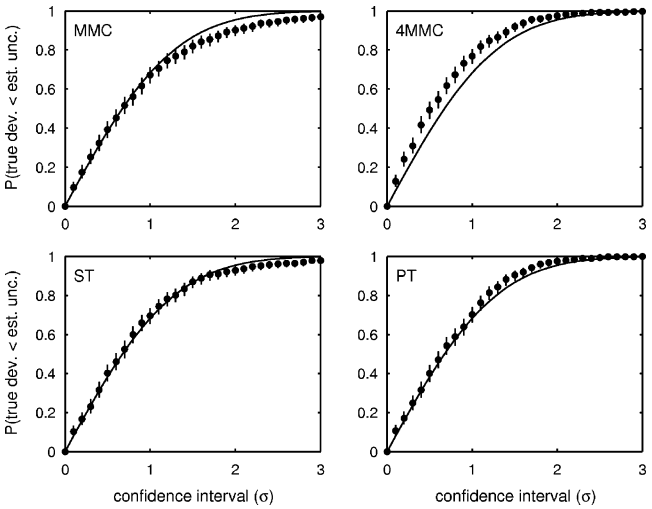


Figure 1. Confidence curves for Metropolis Monte Carlo simulations on the 1D model potential. The fraction of statistically independent blocks for which the true uncertainty (the deviation of the estimated expectation over the block from the mean of the block estimates) is less than a multiplier of the predicted 1σ uncertainty (here plotted as the independent variable is shown). The solid curve shows the fraction expected to fall within the interval for the normal distribution. Ideally, the curves would coincide. The results are shown for (MMC) a single Metropolis Monte Carlo simulation at $\beta = 4$; (4MMC) a set of four independent canonical simulations spanning the range $\beta = 1-4$; (ST) a simulated tempering simulation spanning $\beta = 1-4$; (PT) a parallel tempering simulation with four replicas spanning $\beta = 1-4$. Uncertainties, with 95% confidence intervals shown here as vertical bars, were computed as described in Appendix B.

time. It is readily apparent from the figure that the computed uncertainty estimate computed for each block is in fact quite good. Additionally, the bias is small—less than 10% of the magnitude of the statistical uncertainty in the cases studied (data not shown).

4.2. Alanine Dipeptide in Implicit and Explicit Solvent. To illustrate the utility and verify the correctness of the PTWHAM procedure described above for simulations of biological interest, we demonstrate their use in the analysis of parallel tempering simulations of alanine dipeptide in implicit and explicit solvent. A similar strategy to the 1D model system described above was adopted, with a long simulation partitioned into short blocks (here, 2 ns/replica per block) whose expectations were verified to be statistically independent by the same procedure described above.

Using the LEaP program from the AMBER7 molecular mechanics package,[34] a terminally blocked alanine peptide (sequence ACE-ALA-NME, see Figure 2) was generated in the extended conformation. For the explicit solvent system, the peptide was solvated with 431 TIP3P water molecules[35] in a truncated octahedral simulation box whose dimensions were chosen to ensure a minimum distance to the box boundaries from the initial extended peptide configuration of 7 Å. Peptide force field parameters were taken from the parm96 parameter set.[36] For the implicit solvent simulation, the Generalized Born method of Tsui and Case[37] (corresponding to the flag igb=1) was employed with radii from AMBER6, along with a surface area penalty term of the default 5 cal mol^{-1} $Å^{-2}$. Covalent bonds to hydrogen were constrained with SHAKE using a tolerance of 10^{-8} Å.[38] Long-range electrostatics for the explicit solvent simulation were treated by the particle-mesh Ewald (PME) method[39] with default settings.

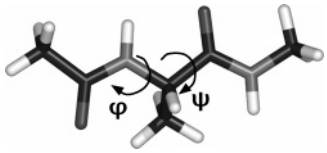


Figure 2. Terminally blocked alanine peptide with (ϕ,ψ) torsions labeled.

Each system was first subjected to 50 steps of steepest descent energy minimization, followed by 1000 steps of conjugate gradient optimization. To equilibrate the explicit solvent system to the appropriate volume, a 100 ps molecular dynamics simulation was performed with the temperature adjusted to 300 K and the pressure to 1 atm by the Berendsen weak-coupling algorithm[40] with temperature and pressure relaxation time constants of 1 and 0.2 ps, respectively. The simulation box was fixed at the final size obtained from this equilibration step, with a volume of 13 232 $Å^3$, in all subsequent simulations.

A parallel tempering (or replica-exchange among temperatures) molecular dynamics simulation[4] was conducted using a parallel Perl wrapper for the sander program. [A copy of this Perl wrapper to perform replica-exchange simulations using AMBER7 and AMBER8 can be obtained from URL http://www.dillgroup.ucsf.edu/~jchodera/code/rex.] Replica temperatures were exponentially distributed over the range 273−600 K, with 10 replicas required for the implicit solvent simulation (yielding an exchange acceptance probability between neighboring temperatures of approximately 75%) and 40 replicas for the explicit solvent simulation (yielding an acceptance probability of approximately 50%). All momenta were reassigned from the Maxwell-Boltzmann distribution at the appropriate replica temperature after each exchange attempt. Between exchanges, constant-energy, constant-volume molecular dynamics was carried out for the explicit solvent simulation, while the implicit solvent simulation utilized Langevin dynamics with a friction coefficient of 95 ps^{-1} to mimic the viscosity of water. All dynamics utilized a 2 fs time step. The algorithm used to select pairs of replicas for temperature exchange attempts starts from the highest-temperature replica and attempts to swap the configuration for the next-lowest temperature replica using a Metropolis-like criteria and proceeds down the temperatures in this manner. On the next iteration, swapping attempts start from the lowest temperature and proceed upward, and this alternation in direction is continued in subsequent pairs of iterations.

Starting all replicas from the minimized or volume-equilibrated configuration described above, 100 iterations were conducted with 1 ps between exchange attempts to equilibrate the replicas to their respective temperatures. This equilibration run was followed by a production run with 20

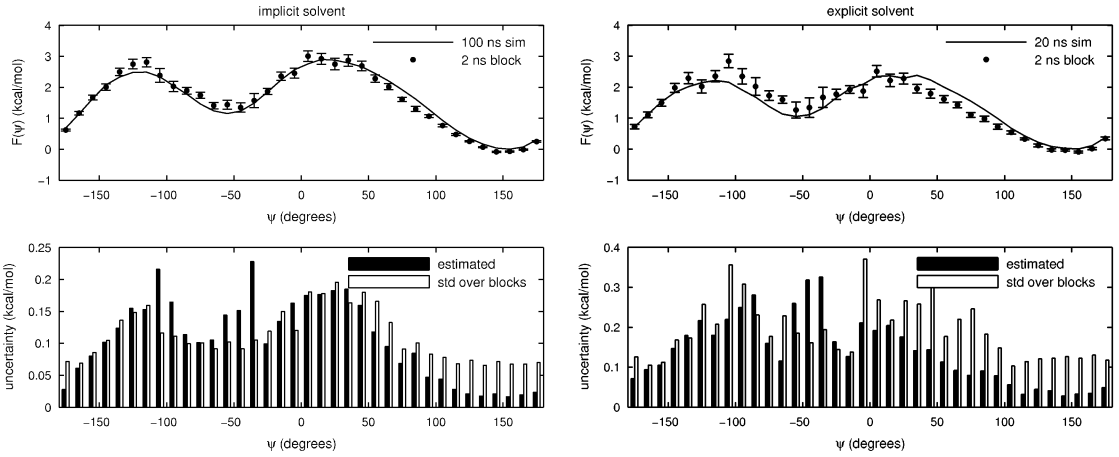


Figure 3. Potential of mean force in ψ for implicit and explicit solvent parallel tempering simulations. Left: implicit solvent; right: explicit solvent. Upper panels: The potential of mean force in the ψ torsion angle at 300 K. The solid line shows the PMF estimated from the entire simulation, while the filled circles show the estimated PMF uncertainty using the method described in the text for a single 2 ns/replica block. Lower panels: The computed uncertainties for the same 2 ns block (filled bars) along with the average uncertainty expected for a simulation 2 ns/replica in length, estimated from the standard deviation of the PMFs computed from all nonoverlapping blocks of length 2 ns in the full simulation (open bars). All uncertainties are shown as one standard deviation.

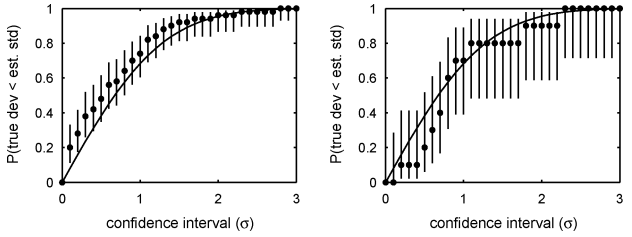


Figure 4. Confidence curves for implicit and explicit solvent parallel tempering simulations. As in Figure 1, the fraction of statistically independent 2 ns blocks for which the true uncertainty is less than a multiplier of the predicted 1σ uncertainty is shown. The observable used is an indicator function for the α_R configuration. Left: implicit solvent (statistics over 50 blocks); right: explicit solvent (statistics over 10 blocks).

ps between exchange attempts, giving a total of 100 ns/replica for the implicit solvent production run and 20 ns/replica for the explicit solvent run. Solute configurations and potential energies were saved from the production run every 1 ps. Expectations and uncertainties were again estimated using listing 1 appearing in the Supporting Information.

Over 2 ns blocks of simulation time (containing 2000 configurations/replica in each block), we computed the probability of the peptide occupying the α_R conformation at 300 K, with α_R here defined as $-105 \leq \phi < 0$ and $-124 \leq \psi < 28$. This corresponds to configurations that would be classified as right-handed alpha-helical. To validate the uncertainty estimates, confidence curves of the type description in section 4.1 were computed and are shown in Figure 4. Though the confidence intervals are larger because the data contain fewer independent blocks, the uncertainty estimates are still good indicators of the expected deviation from the true expectation.

The potential of mean force (PMF) for the ψ torsion angle at 300 K was also computed and is shown in Figure 3. The computed PMF and uncertainty for a representative block is depicted in the top panel, along with the PMF computed using the entire trajectory. The deviations of the block PMFs from the whole-simulation estimate fall within the 1σ uncertainty bars to the expected degree. In the lower panel, the uncertainties computed from the representative block are compared to the standard deviation of the PMF computed from all blocks, which should be indicative (to within an order of magnitude) of the uncertainty expected from a simulation of the block length. These too compare favorably.

It is important to note that our neglect of the uncertainty in the dimensionless free energies, $\{f_l\}$, is only reasonable if the correlation time of the observable of interest is much longer than that of the potential energy. When this condition is satisfied, the dominant contribution to the uncertainty in the computed expectation of the observable is due to the small number of effectively independent samples of this observable present in the simulation data. To demonstrate that this is the case for systems of interest, we have assessed the relative contribution of the neglected uncertainty in the $\{f_l\}$ to the uncertainty of the estimated probability of the α_R conformation of the alanine dipeptide system considered here. The resulting contribution is 10 times smaller than the uncertainty due to the time correlation treated above for the explicit solvent system and 100 times smaller for the implicit solvent system. [The impact of the uncertainty in the $\{f_l\}$ on the uncertainty in the estimated observable was computed in the following manner: We first computed estimates of the $\{f_l\}$ over all uncorrelated 2 ns/replica blocks of simulation data to form a pool of dimensionless free energies that represent the typical uncertainty in a simulation of this length. Next, for each 2 ns/replica block, we computed the standard deviation in the estimated α_R probability when all $\{f_l\}$ in this pool were substituted into the WHAM equations. The mean of this standard deviation over all blocks then provides an estimate of the magnitude of the impact of typical

uncertainties in the $\{f_l\}$ on the observable of interest.] However, if the observable has a correlation time comparable to that of the potential energy (e.g., if the expectation of the potential energy itself is of interest), then the uncertainty due to imperfect knowledge of the $\{f_l\}$ can be comparable to the uncertainty due to the correlation in the observable. In these cases where correlation times are comparable, an algorithm that combines our approach with the T-WHAM method of Gallicchio et al.,[24] which explicitly treats the uncertainty in the $\{f_l\}$ when the potential energy samples are uncorrelated, may provide a superior estimate of the uncertainty in the estimate of the observable.

We further note that pathological cases may arise where simulations at neighboring temperatures may have poor energy overlap, resulting in large uncertainties in some of the $\{f_l\}$. Fortunately, these cases are easily detected by examination of the exchange acceptance rates between neighboring temperatures, where they will be conspicuously low, and detectable early in the simulation. Such cases are easily remedied by adjusting the temperature spacing or by the addition of more replicas at intermediate temperatures.

5. Practical Considerations

Several issues of great importance to successful implementation of the algorithm have received little discussion in the literature.

5.1. Choice of Bin Width and Number of Bins. There is a bias-variance tradeoff in the choice of energy histogram width. As the energy bin width increases, the uncertainty in our histogram estimator for $p_m(\beta)$, the probability density for energy bin m, decreases. At the same time, one expects the resulting estimate of the density of states Ω_m to become increasingly biased, especially considering the dependence of $p(U)$ on the rapidly varying exponential Boltzmann factor $e^{-\beta U}$. Because of this, a reasonable assumption might be that the bin width ΔU should be chosen such that $\Delta U \ll k_B T$. However, if the bin size is too small, the uncertainty in the estimate for the $p_m(\beta)$ will be large. One possibility might be to use a data based choice of histogram bin width, as in Wand,[41] which uses concepts from nonparametric density estimation in attempting to minimize the mean integrated square error (MISE) to the true probability density.

For the alanine dipeptide simulations described in section 4.2 above, we find that the estimated probability of occupying the α_R region of conformation space is largely insensitive to the number of bins used to discretize the sampled potential energy range. In fact, the variation in the computed expectation is well within the statistical uncertainty over the range of 50−5000 bins (corresponding to a range of bin widths of 0.5−50 $k_B T$).

5.2. Computing Integrated Correlation Times. Estimating the correlation time τ, defined above in eqs 19 and 21, can be difficult when one is confronted with noisy correlation functions. While ensuring trajectories are many times longer than the longest correlation times is necessary for an accurate estimate, even if this is achieved, performing the straightforward sum over the entirety of the correlation function C_t as in eq 19 is almost always a poor choice, as the uncertainty in the computed correlation function grows approximately linearly with the lag time t.[42] Even for trajectories many times longer than the correlation length, this sum will be dominated by contributions from the noisy tail, likely resulting in large errors or even negative values for the computed correlation time τ. Janke proposes a self-consistent approach where the summation is performed only out to lag times of 6τ, after which the correlation function is assumed to be negligible.[23] Evertz contends that this approach produces incorrect results,[43] instead proposing an exponential fit to the tail of the correlation function and use of this fit to evalute the summand when the correlation function is dominated by noise. Neither solution is both stable and straightforward to apply, so we instead truncate the sum when the normalized fluctuation correlation function C_t first crosses zero, since it is likely unphysical for the correlation function to be negative for most observables. [Velocity autocorrelation functions, where there is often a clear negative peak at short times, are an obvious exception.] The zero crossing is an indication that the statistical uncertainty dominates the signal and that the remainder of the correlation function should be considered indistinguishable from zero.

For most systems and observables, the correlation function will decay rapidly at first and then slowly, approximately exponentially for large t. To avoid the expense of computing C_t at each value of t while still obtaining reasonably accurate integrated correlation times for observables with very different decay time scales, we use an adaptive integration scheme in which the correlation function C_t is computed only at times $t_i = 1 + i(i-1)/2$, where $i = 1, 2, 3, \ldots$. In computing the correlation time τ, the sum in eq 19 is now performed only over the t_i terms, with each term weighted by $t_{i+1} - t_i$, with $t_1 = 1$. This approach ensures high time resolution at small t when C_t is likely to be rapidly changing but avoids the expense of computing C_t at every t in the slowly decaying tails. We find the accuracy of this approach to be acceptable—differences typically amount to at most 10%.

5.3. Neglect of Bin Statistical Inefficiencies g_{kn}. It is often assumed or stated without justification that the energy bin statistical inefficiencies g_{mk} appearing in eqs 30 and 64, representing the number of snapshots required for a statistically independent sampling of the energy bin, are all equal or equal to unity.[2,4,1] All g_{mk} will be equal to unity only if the $\{\psi_{mkn}\}_{n=1}^{N_k}$ are uncorrelated. To test this assumption, we have computed the statistical inefficiencies for the systems mentioned in section 4 above. To our knowledge, this is the first time a test of this claim has been reported in the literature. Indeed, for the explicit solvent system studied in section 4.2 above, we find large differences in the statistical inefficiencies for the same replica but different energy bins, sometimes differing up to two orders of magnitude. Similarly, for the same energy bin, the statistical inefficiencies from different replicas can differ by up to 2 orders of magnitude.

In the limit that our parallel tempering simulation is very long—long enough for each replica to execute an unrestricted random walk through all temperatures and explore all relevant regions of configuration space—each replica can be considered to be equivalent. In this case, the statistical inefficiencies should be independent of replica index k, and

we can write g_m as the statistical inefficiency for energy bin m. Applied to eqs 64 and 65, this yields a new set of expressions:

$$\hat{\Omega}_m = \frac{\sum_{k=1}^{K} H_{mk}}{\sum_{k=1}^{K}\sum_{l=1}^{L} N_{kl}\,\Delta U \exp[f_l - \beta_l U_m]} \tag{70}$$

$$\delta^2\hat{\Omega}_m = \frac{\hat{\Omega}_m}{g_m^{-1}\sum_{k=1}^{K}\sum_{l=1}^{L} N_{kl}\,\Delta U \exp[f_l - \beta_l U_m]} \tag{71}$$

The quantity $\sum_{k=1}^{K} N_{kl}$, simply represents the total number of configurations stored from any replica at temperature β_l. For a parallel tempering simulation, where each temperature must be populated by exactly one replica at all times, this is simply N, the total number of configurations stored per replica. Additionally, $H_m \equiv \sum_{k=1}^{K} H_{mk}$ is the total number of configurations over all replicas (and hence temperatures) with energy in bin m. This gives

$$\Omega_m = \frac{H_m}{N\sum_{l=1}^{L}\exp[f_l - \beta_l U_m]} \tag{72}$$

which is identical to the WHAM result for independent canonical simulations for the case where all g_{mk} are identical or unity. If there is no correlation in the data—that is, all configurations are independent—it does not matter whether we apply this analysis to the original data or collect the configurations by temperature and apply WHAM equations for independent canonical simulations. This expression is in fact identical to the one used in many published works that have previously attempted to use the weighted histogram analysis method for the analysis of parallel tempering simulations, such as refs 4 and 1.

Under these same assumptions, the correlation functions for any observable A should also be identical for each replica. One can therefore average estimates of the unnormalized correlation functions $\langle A_n A_{n+t}\rangle$ over the replicas and use optimal estimates of the mean and variance computed over all of the replicas to obtain an optimal estimate of the statistical inefficiencies and uncertainties. An implementation illustrating this procedure is provided in the Supporting Information as listing 2.

Note, however, that the above assumptions of equivalence cannot be made in cases where the replicas are clearly *inequivalent*, such as in a simulated tempering replica-exchange (STREM) simulation.[3,14] In that case, the expression above will only be recovered if the time between samples is so long that all the g_{mk} are unity.

5.4. The Statistical Inefficiency for the Cross-Correlation Term, $g_{k,wA;w}$. In computing the statistical inefficiency for the cross-correlation term, uncertainties in the computed integrated autocorrelation times due to insufficient data or approximations may cause the cross-correlation term to dominate and the estimate of the square uncertainty of $\hat{A}$ (eq 41) to be negative. Clearly, this should not be allowed to occur, as the squared-uncertainty should be a strictly positive quantity.

The statistical inefficiencies g should obey the following relation (derived in Appendix A):

$$g_{x;y} \le (g_x\, g_y)^{1/2}\,\frac{\sigma_x\sigma_y}{|\sigma_{x;y}{}^2|} \tag{73}$$

This is often violated when the correlation function is noisy and can lead to negative estimates of the squared uncertainty when the cross-correlation term dominates. In these cases, we find it best to simply limit $g_{x;y}$ to its maximum allowed value computed from the right-hand side of eq 73. Since the autocorrelation times are usually shorter than the cross-correlation time, it is believed that these estimates will be better than the integrated cross-correlation time.

6. Conclusion

We have presented extensions of the weighted histogram analysis method, STWHAM and PTWHAM, for the analysis of one or more independent simulated tempering or parallel tempering simulations. The method provides not only estimators of canonical expectations but also estimators for the statistical uncertainties in the resulting estimates. We hope that, with the availability of the provided example code, workers using these simulation techniques will provide uncertainty estimates so that the statistical significance of results obtained from them can be assessed. We have shown that the estimator for the expectation has small bias and produces excellent uncertainty estimates for both a 1D model system and a solvated biomolecular system in implicit and explicit solvent.

While other workers had attempted to apply WHAM to simulated or parallel tempering data in the past,[3,4] the key advance here is the consideration of the correlated nature of the configurations sampled by each replica as it performs a pseudorandom walk in temperature, allowing a proper assessment of the true number of independent samples present in the data. This produces correct optimal estimators and makes possible the estimation of statistical uncertainties. This method can be extended to the analysis of other generalized-ensemble simulations, such as the multicanonical method (MUCA),[5−9] by consideration of replica correlation times as the system samples various energy levels biased by the estimate of the density of states. Still, it is important to point out that, while we consider the contribution from time-correlation of the observable to the uncertainty estimate, we currently neglect the contribution of the uncertainty in the per-configuration weights (which originates from the uncertainty in the density of states) to the estimate of the expectation—we assume it is negligible and await a more complete treatment of the uncertainty in cases where it is not.

Appendix A: Relation for Statistical Inefficiencies

Consider a random process where we make a series of N time-correlated measurements of two (possibly correlated)

observables X and Y, resulting in the time series $\{x_n, y_n\}_{n=1}^N$. We estimate the quantity $Z = \langle X\rangle/\langle Y\rangle$ from our sample means and wish to compute the uncertainty in our estimate, defined as

$$\delta^2 Z \equiv \sigma_Z{}^2 = \langle (Z - \langle Z\rangle)^2\rangle = \langle Z^2\rangle - \langle Z\rangle^2 \tag{74}$$

To first order about $\langle X\rangle/\langle Y\rangle$, this uncertainty is given by

$$\sigma_Z{}^2 = \left[\frac{\langle X\rangle}{\langle Y\rangle}\right]^2 \left[\frac{\sigma_X{}^2}{\langle X\rangle^2} + \frac{\sigma_Y{}^2}{\langle Y\rangle^2} - 2\frac{\sigma_{X;Y}{}^2}{\langle X\rangle\langle Y\rangle}\right] \tag{75}$$

where $\sigma_{X;Y}{}^2$ denotes the (not necessarily positive) covariance of the expectations of X and Y. The Schwartz inequality requires that this covariance obey the relation

$$|\sigma^2{}_{X;Y}| \leq \sigma_X \sigma_Y \tag{76}$$

(see, for example, ref 44). Given this, we note

$$\left|\frac{\sigma_{X;Y}{}^2}{\sigma_X \sigma_Y}\right| \leq 1 \tag{77}$$

Correlation analysis gives estimators for these quantities as

$$\sigma_X{}^2 = \sigma_x{}^2/(g_x{}^{-1}N)$$

$$\sigma_Y{}^2 = \sigma_y{}^2/(g_y{}^{-1}N)$$

$$\sigma_{X;Y}{}^2 = \sigma_{x;y}{}^2/(g_{x;y}{}^{-1}N) \tag{78}$$

where $\sigma_x{}^2$ denotes the sample variance of the observations $\{x_n\}_{n=1}^N$ and g denotes the statistical inefficiency obtained from the autocorrelation time, i.e.

$$g_x = 1 + 2\tau_x \geq 1 \tag{79}$$

Combining eqs 77 and 78 gives

$$\frac{|\sigma_{x;y}{}^2|}{\sigma_x\, \sigma_y} \cdot \frac{g_{x;y}}{(g_x\, g_y)^{1/2}} \leq 1 \tag{80}$$

where we have moved the statistical inefficiencies and variances out of the absolute value, as they are always positive. Finally, we obtain an upper bound for $g_{x;y}$:

$$g_{x;y} \leq (g_x\, g_y)^{1/2} \frac{\sigma_x \sigma_y}{|\sigma_{x;y}{}^2|} \tag{81}$$

In using numerical methods to estimate the statistical inefficiencies g from finite trajectories, this inequality may not hold, sometimes leading to negative squared uncertainties. Limiting the estimated $g_{x;y}$ by capping it at this value will prevent this from occurring.

Appendix B: Uncertainty Estimates for Confidence Curves

To estimate the uncertainties in Figures 1 and 4, a Bayesian inference scheme was used. Since the expectations computed from each block are independent, the number of blocks n that fall inside the given scaled deviation σ is described by a binomial distribution with parameter θ, true (unknown) probability that the blocks fall within the given deviation σ:

$$P(n|\theta) = \frac{N!}{n!(N-n)!}\theta^n(1-\theta)^{N-n} \tag{82}$$

We can write the posterior distribution for the probability p given the observed number of blocks n within the given deviation using Bayes' rule

$$p(\theta|n) \propto p(n|\theta)\, p(\theta) \tag{83}$$

where $p(\theta)$ is the prior distribution for the parameter θ. If we choose the prior $p(\theta)$ to be a Beta distribution with hyperparameters α and β, given by

$$p(\theta|\alpha,\beta) = [B(\alpha,\beta)]^{-1}x^{\alpha-1}(1-x)^{\beta-1} \tag{84}$$

where $B(\alpha,\beta)$ is the beta function, then the posterior $p(\theta|\, n)$ will also be a Beta distribution with parameters $n + \alpha$ and $(N - n) + \beta$, as the Beta distribution is a conjugate prior to the Binomial distribution. We take the hyperparameters α and β to be unity to make the prior distribution uniform, resulting in a posterior $\theta \sim \text{Beta}(n + 1, N - n + 1)$. A 95% central confidence interval, corresponding to the location where the cumulative distribution function for the Beta distribution reaches the values of 0.025 and 0.0975, was plotted.

Acknowledgment. J.D.C. was supported by a Howard Hughes Medical Institute and an IBM predoctoral fellowship. W.S. acknowledges support from NSF MRSEC Center on Polymer Interfaces and Macromolecular Assemblies DMR - 0213618, and K.D. acknowledges the support of NIH grant GM34993. The authors would like to thank Libusha Kelly, Bosco K. Ho, and M. Scott Shell (University of California, San Francisco) for critical reading of this manuscript as well as the anonymous referee who raised an excellent point regarding our neglect of the uncertainty in the dimensionless free energies. This manuscript was strengthened as a result of their input.

Supporting Information Available: Source code Listings 1 and 2 for reference implementations of PTWHAM in Fortran 90/95, assuming inequivalent and equivalent replicas, respectively, as well as an archive containing a compilable version of Listing 1 and a test data set. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Okamoto, Y. *J. Mol. Graphics Modell.* **2004**, *22*, 425−439.

(2) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011−1021.

(3) Mitsutake, A.; Okamoto, Y. *Chem. Phys. Lett.* **2000**, *332*, 131−138.

(4) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(5) Berg, B. A.; Neuhaus, T. *Phys. Lett. B* **1991**, *267*, 249−253.

(6) Berg, B. A.; Neuhaus, T. *Phys. Rev. Lett.* **1992**, *68*, 9−12.

(7) Hansmann, U. H. E.; Okamoto, Y. *J. Comput. Chem.* **1993**, *14*, 1333−1338.

(8) Nakajima, N.; Nakamura, H.; Kidera, A. *J. Phys. Chem. B* **1997**, *101*, 817−824.

(9) Yaşar, F.; Çelik, T.; Berg, B. A.; Meirovitch, H. *J. Comput. Chem.* **2000**, *21*, 1251−1261.

(10) Marinari, E.; Parisi, G. *Europhys. Lett.* **1992**, *19*, 451−458.

(11) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. *J. Chem. Phys.* **1992**, *96*, 1776−1783.

(12) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140−150.

(13) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *J. Chem. Phys.* **2003**, *118*, 6664−6675.

(14) Mitsutake, A.; Okamoto, Y. *J. Chem. Phys.* **2004**, *121*, 2491−2504.

(15) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **2000**, *2329*, 261−270.

(16) Fischer, A.; Cordes, F.; Schütte, C. *J. Comput. Chem.* **1998**, *1999*, 1689−1697.

(17) Sanbonmatsu, K. Y.; Garcia, A. E. *Proteins: Struct., Funct., Genet.* **2002**, *46*, 225−234.

(18) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1998**, *61*, 2635−2638.

(19) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195−1198.

(20) Müller-Krumbhaar, H.; Binder, K. *J. Stat. Phys.* **1973**, *8*, 1−23.

(21) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637−649.

(22) Flyvbjerg, H.; Petersen, H. G. *J. Chem. Phys.* **1989**, *91*, 461−466.

(23) Janke, W. Statistical analysis of simulations: Data correlations and error estimation. In *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*; Grotendorst, J., Marx, D., Murmatsu, A., Eds.; John von Neumann Institute for Computing: 2002; Vol. 10.

(24) Gallicchio, E.; Andrec, M.; Felts, A. K.; Levy, R. M. *J. Phys. Chem. B* **2005**, *109*, 6722−6731.

(25) Gelman, A.; Rubin, D. B. *Stat. Sci.* **1992**, *7*, 457−472.

(26) Souialle, M.; Roux, B. *Comput. Phys. Commun.* **2001**, *135*, 40−57.

(27) Silverman, B. W. *Density Estimation for Statistics and Data Analysis;* Chapman and Hall: New York, 1986.

(28) Willard, K. E.; Connelly, D. P. *Comput. Biomed. Res.* **1992**, *25*, 17−28.

(29) Cowan, G. *Statistical Data Analysis;* Oxford University Press: 1998.

(30) Ferrenberg, A. M.; Landau, D. P.; Swendsen, R. H. *Phys. Rev. E* **1995**, *51*, 5092−5099.

(31) Newman, M. E. J.; Palmer, R. G. *J. Stat. Phys.* **1999**, *97*, 1011−1026.

(32) Taylor, J. R. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*, 2nd ed.; University Science Books: 1996; Chapter 3.11, pp 73−79.

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

(34) Case, D. A. et al. *AMBER7*; 2002.

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

(36) Kollman, P. A.; Dixon, R.; Cornell, W.; Vox, T.; Chipot, C.; Pohorille, A. The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of *a* b initio calculations and experimental data. In *Computer Simulation of Biomolecular Systems*; Wilkinson, A., Weiner, P., van Gunsteren, W. F., Eds.; Kluwer/Escom: 1997; Vol. 3.

(37) Tsui, V.; Case, D. A. *J. Am. Chem. Soc.* **2000**, *122*, 2489−2498.

(38) Ryckaert, J.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.

(39) Darden, T. A.; York, D. M.; Pedersen, L. G. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(40) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(41) Wand, M. P. *Am. Stat.* **1997**, *51*, 59.

(42) Zwanzig, R.; Ailawadi, N. K. *Phys. Rev.* **1969**, *182*, 280−283.

(43) Evertz, H. G. *Adv. Phys.* **2003**, *52*, 1−66.

(44) Casella, G.; Berger, R. L. *Statistical inference;* Duxbury Press: Belmont, CA, 1998.

CT0502864

JCTC Journal of Chemical Theory and Computation

Stereoelectronic Substituent Effects in Saturated Main Group Molecules: Severe Problems of Current Kohn−Sham Density Functional Theory

S. Grimme,* M. Steinmetz, and M. Korth

Theoretische Organische Chemie, Organisch-Chemisches Institut der Universität Münster, Corrensstrasse 40, D-48149 Münster, Germany

Received August 7, 2006

Abstract: The Hartree−Fock method, two common density functionals (PBE and B3LYP), and two new functionals (B97-D and B2PLYP) together with very large AO basis sets are used to compute the isomerization energies for substituted (R=H, F, Cl) branched to linear alkanes and silanes. The results of accurate SCS-MP2 computations are taken as reference. These reactions are an important test of how nonlocal electron correlation effects on medium-range lengths scales in saturated molecules are treated by approximate quantum chemical methods. It is found that the unacceptably large errors observed previously for hydrocarbons persist also for the here considered more polar systems. Although the B97-D and B2PLYP functionals provide improved energetics, the problem is not fully solved, and thus these systems are suggested as mandatory benchmarks for future density functionals.

Kohn−Sham density functional theory (KS-DFT) is now the most widely used method for electronic structure calculations in condensed matter physics and quantum chemistry.[1,2] This success mainly results from significant 'robustness', i.e., providing reasonably accurate predictions for many properties of various molecules and solids.[3] However, a general drawback of many common density functionals (DF) is that they cannot be systematically improved, and thus, empirical tests in general chemistry applications are mandatory. Some fundamental failures of common DF are known quite well such as the self-interaction error or that they cannot really describe long-range electron correlations that are responsible for van der Waals (vdW) forces.[4−6] Especially the vdW problem has attracted some attention in the DFT community in recent years (see e.g. refs 7−14) as these interactions play an important role in many chemical systems such as for the structures of DNA and proteins, the packing of crystals, the formation of aggregates, host−guest systems, or the orientation of molecules on surfaces.

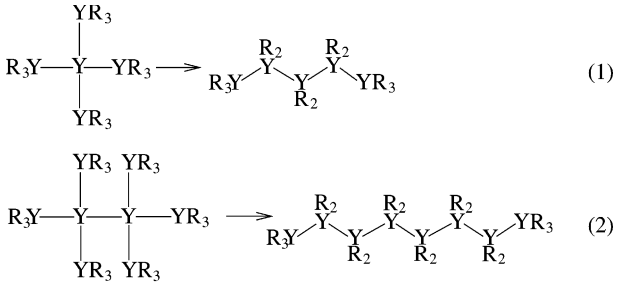
In a recent publication[15] it was shown, however, that current DFT also has problems to describe other, seemingly simple electron correlations that are the reason for stereoelectronic substituent effects.[16] In hydrocarbons, all common DF cannot describe the energetic consequences of medium-range but nonlocal electron correlations originating from different (perfectly) localized σ-orbitals. The effect of such correlations is important when the size or shape of molecules change (for related examples see refs 17−20) although it has nothing to do with 'size-consistency' problem of some approximate quantum chemical methods. Previously,[15] the most simple case, i.e., the isomerizations of branched to linear alkanes (see Scheme 1, Y = C, R = H, denoted as 5CH and 8CH, respectively) has been considered. It was found, that e.g. for the isomerization to *n*-octane (reaction 2), no current DF could provide even the right sign (endergonic) for the isomerization energy ΔE. Similar problems appear in standard semiempirical approaches such as MNDO or AM1 but could be eliminated by pairwise-distance-directed-Gaussian (PDDG) modifications of the interaction potential.[21]

In this study we investigate the obvious question if the problem only appears for hydrocarbons or if it is of more general importance. We extend our computations here to saturated systems of other elements and have chosen as new test molecules halogen-substituted alkanes (Y=C, R=F) and silanes (Y=Si; R=H, F, Cl), respectively. Polychlorinated

* Corresponding author fax: (+49)251-83-36515; e-mail: grimmes@uni-muenster.de.

10.1021/ct600224b CCC: $37.00
Published on Web 11/29/2006

Scheme 1. Investigated Molecules



Y=C, Si
R=H, F, Cl

Table 1: Computed[a] Isomerization (Branched → Linear) Energies ΔE (in kcal mol^{-1})

molecule	HF	B3LYP	PBE	B97-D	B2PLYP	SCS-MP2
5CH	0.1	1.3	1.6	3.4	2.1	3.0
5CF	10.4	9.8	11.5	11.5	12.5	14.1
8CH	−11.5	−8.4	−5.5	2.2	−3.5	1.4
8CF	−5.0	−5.0	−1.5	−0.9	2.6	7.4
5SiH	1.8	1.8	2.5	3.7	3.2	4.8
5SiF	26.7	24.7	26.5	26.7	28.1	30.9
5SiCl	2.0	3.3	6.0	6.2	7.9	11.0
8SiH	2.0	2.2	4.7	8.5	6.5	11.3
8SiF	49.9	45.7	50.3	52.7	54.1	61.5
8SiCl	−15.1	−8.3	−1.4	−1.8	1.8	10.2

[a] Single point calculations on MP2/TZVP optimized geometries and employing the cQZV3P AO basis.

alkanes are excluded because repulsive steric interactions between the chlorine atoms are very dominant in these systems. Compared to the previously studied alkanes, the average distance between the σ-bonds (or substituents) is larger in the silanes, and furthermore, the effect of bond polarity can be investigated quite systematically by comparison of R = H and R = F. We also present first results from a recently developed generalized gradient approximation (GGA) type functional (termed B97-D[22]) that is correct at least in the case of the hydrocarbons. It is based on a reparametrization of Beckes ansatz from 1997[23] but now explicitly includes long-range electron correlations by atom-pair wise dispersion corrections of the form $C_6 \cdot R^{-6}$ as in the DFT-D method.[8] In this new approach double-counting effects of electron correlation are avoided by construction, and the density functional description is restricted to the short-range electron correlations.

Experimental isomerization enthalpies are unfortunately not available for the fluoroalkanes and all silanes. However, recent experience showed that the SCS-MP2 method[24] with large AO basis sets as used here (of polarized quadruple-ζ quality, see Methods) is accurate to within 0.5 kcal mol^{-1} for the ΔE values of hydrocarbons. We expect for the bigger systems studied here where also the absolute ΔE values are generally larger an error of about $\pm$(1−2)kcal mol^{-1}. This is in any case accurate enough to judge the quality of approximate functionals that in our case yield large errors of 5−15 kcal mol^{-1}.

The results for the linear to branched form isomerization energies are shown in Table 1, and the errors with respect to the SCS-MP2 values taken as reference are shown graphically in Figure 1 (for convenience only errors for reaction 2 are shown). As a quantum chemical method we employ also (uncorrelated) Hartree−Fock (HF) to see how a large fraction of the electron correlation contribution to ΔE is typically recovered by DFT. As standard functionals we consider the popular B3LYP[25,26] and PBE[27] forms and note in passing that many other functionals (e.g. the meta-GGA TPSS[28]) behave quite similar. Besides the new B97-D GGA, the recently developed 'fifth-rung' virtual-orbital dependent hybrid B2PLYP[14] is used. This functional contains explicitly interpair correlations that seem to be necessary for an accurate description of such isomerization processes.[15]

Before discussing the errors of the investigated methods that is the main point of the present work, a brief look at some general aspects and trends of the ΔE values in the series of systems seems appropriate. First, for all systems the ΔE values are positive meaning that the branched form is more stable. Furthermore, the ΔE values decrease (the branched forms are destabilized) when the core of the molecules is increased from five to eight atoms for Y = C, while they increase for Y = Si (R=H, F) or are almost constant (R=Cl). The different behavior of the alkanes opposed to the silanes can be explained by the much longer Si−Si bonds which lead to much smaller steric interactions than in alkanes, and consequently, the intrinsic effects of electron correlation (that stabilize the branched forms) are more obvious in the silanes. In the chlorosilanes 5SiCl and 8SiCl we once again can observe the steric effects as here the ΔE values decrease as in the case of Y = C. This is clearly due to the larger size of Cl and the longer Si−Cl bonds that cause more steric interference in the branched forms.

A second point concerns the dependence on the bond polarity. For both, Y = C and Y = Si, the ΔE values increase for R = Cl, F compared to R = H (with the exception of 8SiCl, see above). When comparing R = H and R = F (where steric effects are of minor importance), the ΔE values are larger by about a factor of 6 for R = F. From an analyses of the errors of the different methods (see below) and the additive property of the effects, we assign this behavior to an unfavorable arrangement of the bond dipoles in the linear forms compared to the branched ones. This view is corroborated by the energy contribution to ΔE from the point-charge model in common force-fields.

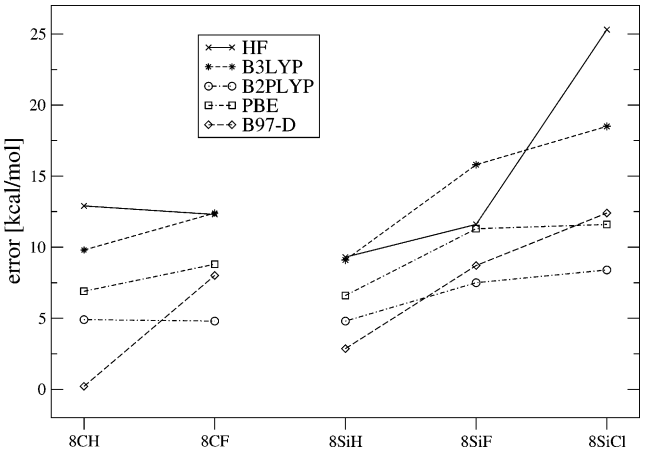


Figure 1. Errors with respect to SCS-MP2 as reference for isomerization energies (reaction 2) with different quantum chemical methods.

For the discussion of the errors of the different density functionals considered we restrict ourselves to reaction 2 because the effects are qualitatively the same but more pronounced compared to reaction 1. Inspection of Figure 1 shows clearly that the methods behave quite similar for carbon and silicon, and thus, the problems appear to be of general importance. The errors are huge, i.e., >10 kcal mol^{-1} for HF and B3LYP, between 5 and 10 kcal mol^{-1} for PBE, and between 0 and 10 kcal mol^{-1} for B97-D and B2PLYP. Even more important, for 8CH all methods except B97-D predict the wrong sign for ΔE, and for 8CF and 8SiCl only B2PLYP is right. Errors on the order of 5−15 kcal mol^{-1} are typical for e.g. the atomization energies of medium-sized molecules where the electronic structure changes a lot but are completely unacceptable for a simple isomerization process where the number and type of bonds and the hybridization state of all atoms remains the same.

The much larger errors of the HF and B3LYP methods for 8SiCl can be attributed to missing attractive vdW interactions that are better accounted for by the other methods. This reasoning also applies for the in general larger errors provided by HF and B3LYP. As discussed in ref 15, intramolecular vdW effects contribute significantly to ΔE, but are not the main reason for the failures. This is also found here for substituents other than hydrogen. In passing we note that also for the halogen substituted systems the amount of nonlocal HF exchange in a density functional has a minor effect on its performance.

The new functionals B97-D and B2PLYP that both try to account for the relevant interactions by empirical and orbital-dependent terms, respectively, perform better than all other functionals. However, there is still room for improvement, i.e., the errors for B97-D increase for R = F, Cl and the B2PLYP errors (although less system-dependent) are still on the order of 4−8 kcal mol^{-1}.

The most important finding of the present investigation is that the errors of common density functionals are comparable for a wide range of polar and nonpolar systems. The explanation of the problem[15] due to nonlocal electron correlations between localized σ-bonds at intermediate distances is fully supported by the present study. It is a very reasonable conjecture that the DFT problems in the description of such isomerization processes, that are *not* element-specific but related to a change in the shape of molecules, indicate only the tip of an iceberg. In many quantum chemical routine studies, these problems may be buried by (counter-effective) basis set incompleteness effects or the fact that only small model systems are studied. For the further systematic development of quantum chemistry and DFT in particular, however, it seems important to solve these problems, and thus, we suggest the here considered systems as mandatory benchmarks for future density functionals.

Methods

The SCS-MP2 and DFT calculations have been performed with slightly modified versions of the TURBOMOLE suite of programs.[29] As AO basis, triple-ζ (TZV) or quadruple-ζ (QZV) sets of Ahlrichs et al.[30] have been employed. In all perturbative treatments (MP2, SCS-MP2, and B2PLYP) the RI-approximation for the two-electron integrals has been used,[31] and all electrons have been correlated. As RI-auxiliary basis, the sets of Weigend et al.[32] that were optimized for the cc-pVQZ AO basis have been employed. The geometries were fully optimized at the MP2/TZV(d,p) level, and single-point calculations on these structures were performed with a quadruple-ζ (cQZV3P) AO basis set that includes (3d2f/2pd) polarization functions and (2s2p2d1f) core-polarization/correlation functions taken from ref 33. This basis set provides results to within $\pm$0.1 kcal mol^{-1} of the basis set limit for the (zero-point vibrational energy exclusive) isomerization energies ΔE in the case of hydrocarbons.[15]

Note Added in Proof. At the proof stage of this letter, important related work came to our attention. Schreiner et al. (*Org. Lett.* **2006**, *8*, 3635−3638) and Wodrich et al. (*Org. Lett.* **2006**, *8*, 3631−3634) present other examples of isomerization, where common density functionals fail badly. Zhao et al. propose a new meta-hybrid functional (M05-2X, *J. Chem. Theory Comput.* **2006**, *2*, 364) that seems to solve the here discussed problems at least for medium-sized hydrocarbons.

Acknowledgment. This work was supported by the Deutsche Forschungsgemeinschaft in the framework of the SFB 424 ("Molekulare Orientierung als Funktionskriterium in chemischen Systemen"). The authors thank C. Mück-Lichtenfeld for technical assistance and helpful discussions.

References

(1) Kohn, W.; Sham, L. J. *Phys. Rev. A* **1965**, *140*, 1133.

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

(3) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory;* Wiley-VCH: New York, 2001.

(4) Hobza, P.; Sponer, J.; Reschel, T. *J. Comput. Chem.* **1995**, *11*, 1315−1325.

(5) Allen, M.; Tozer, D. J. *J. Chem. Phys.* **2002**, *117*, 11113.

(6) Kristyan, S.; Pulay, P. *Chem. Phys. Lett.* **1994**, *229*, 175−180.

(7) Zimmerli, U.; Parrinello, M.; Koumoutsakos, P. *J. Chem. Phys.* **2004**, *120*, 2693−2699.

(8) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463−1476.

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

(10) Basanta, M. A.; Dappe, Y. J.; Ortega, J.; Flores, F. *Europhys. Lett.* **2005**, *70*, 355−361.

(11) Angyan, J. G.; Gerber, I. C.; Toulouse, J. *Phys. Rev. A* **2005**, *72*, 012510.

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

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

(14) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108−16.

(15) Grimme, S. *Angew. Chem.* **2006**, *118*, 4571−4575. Grimme, S. *Angew. Chem., Int. Ed.* **2006**, *45*, 4460−4464.

(16) We use the term 'stereoelectronic' in a broader sense as e.g. some organic chemists who restrict it to conformational (e.g. gauche or anomeric) effects.

(17) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374−7383.

(18) Redfern, P. C.; Zapol, P.; Curtiss, L. A.; Raghavachari, K. *J. Phys. Chem. A* **2000**, *104*, 5850−5854.

(19) Yao, X.-Q.; Hou, X.-J.; Jiao, H.; Xiang, H.-W.; Li, Y.-W. *J. Phys. Chem. A* **2003**, *107*, 9991−9996.

(20) Check, C. E.; Gilbert, T. M. *J. Org. Chem.* **2005**, *70*, 9828−9834.

(21) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002,** *23*, 1601−1622.

(22) Grimme, S. *J. Comput. Chem.* In press.

(23) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554−8560.

(24) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095−9102.

(25) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(26) Stephens, P. J.; Devlin, F. J.; Chablowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623−11627.

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

(28) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003,** *91*, 146401.

(29) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989,** *162*, 165−169. Ahlrichs, R., et al. *TURBOMOLE (Vers. 5.6)*; Universität Karlsruhe, 2003. See, also: http://www.turbomole.com.

(30) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829−5835. Weigend, F.; Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *119*, 12753−12762. The basis sets are available from the TURBOMOLE homepage via the FTP Server Button (in the subdirectories basen, jbasen, and cbasen). See http://www.turbomole.com.

(31) Weigend, F.; Häser, M. *Theor. Chem. Acc.* **1997**, *97*, 331−340.

(32) Weigend, F.; Köhn, A.; Hättig, C. *J. Chem. Phys.* **2002**, *116*, 3175−3183.

(33) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1995**, *103*, 4572. Peterson, K. A.; Dunning, T. H. *J. Chem. Phys.* **2002**, *117*, 10548.

CT600224B

JCTC Journal of Chemical Theory and Computation

Electrostatically Embedded Many-Body Expansion for Large Systems, with Applications to Water Clusters

Erin E. Dahlke and Donald G. Truhlar*

Department of Chemistry and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455-0431

Received August 4, 2006

Abstract: The use of background molecular charge to incorporate environmental effects on a molecule or active site is widely employed in quantum chemistry. In the present article we employ this practice in conjunction with many-body expansions. In particular, we present electrostatically embedded two-body and three-body expansions for calculating the energies of molecular clusters. The system is divided into fragments, and dimers or trimers of fragments are calculated in a field of point charges representing the electrostatic potential of the other fragments. We find that including environmental point charges can lower the errors in the electrostatically embedded pairwise additive (EE-PA) energies for a series of water clusters by as much as a factor of 10 when compared to the traditional pairwise additive approximation and that for the electrostatically embedded three-body (EE-3B) method the average mean unsigned error over nine different levels of theory for a set of six tetramers and one pentamer is only 0.05 kcal/mol, which is only 0.4% of the mean unsigned net interaction energy. We also test the accuracy of the EE-PA and EE-3B methods for a cluster of 21 water molecules and find that the errors relative to a full MP2/aug′-cc-pVTZ calculation to be only 2.97 and 0.38 kcal/mol, respectively, which are only 1.5% and 0.2%, respectively, of the net interaction energy. This method offers the advantage over some other fragment-based methods in that it does not use an iterative method to determine the charges and thus provides substantial savings for large clusters. The method is convenient to adapt to a variety of electronic structure methods and program packages, it has N^2 or N^3 computational scaling for large systems (where N is the number of fragments), it is easily converted to an O(N) method, and its linearity allows for convenient analytic gradients.

1. Introduction

Many computational chemistry applications require one to calculate accurate energies for very large molecules such as proteins, for large molecular clusters or nanoparticles that have hundreds or even thousands of atoms or for condensed-phase extended systems (such as liquid solvents or high-pressure solids in the earth's mantle). Such requirements have led to research on developing quantum mechanical electronic structure methods to efficiently and accurately describe large systems.[1−45] Correlated quantum mechanical methods, such as post-Hartree−Fock wave function theory[46] (WFT) or density functional theory[47] (DFT), while able to give highly accurate results, are limited—in their original computational formulation—to relatively small systems containing only tens or, in the case of DFT, hundreds or a few thousand atoms. If one is interested in studying a system with tens or hundreds of thousands of atoms, conventional quantum mechanical methods are useless, and for systems with only 50−100 atoms, they are often so computationally demanding as to be impractical with the time and resources available for a given project. Molecular mechanics,[48,49] on the other hand, while able to handle the large number of atoms present in extended systems does not provide the predictive quantitative accuracy needed to investigate many of the questions of interest, such as chemical reactions or phase equilibria under conditions (for example, the high-pressure phases of the

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

10.1021/ct600253j CCC: $37.00
Published on Web 11/04/2006

earth's mantle) where there is not enough data for a specific parametrization.

Perhaps the simplest of all approximations that is frequently invoked is to expand the system as a sum of many-body terms and then to truncate the series after only a few terms. If the sum is truncated after inclusion of the two-body terms (interactions of two monomers), then one is said to have made the pairwise additive approximation, while if three-body terms are also included, then one is said to have made the three-body approximation. The benefit of making such approximations is that the total energy of the system can be written as a combination of the energy of all the fragments (called monomers), dimers, and, in the case of the three-body approximation, trimers within the structure. This reduces a single very large and expensive calculation to a large number of small and computationally efficient calculations, and it can allow quantum mechanics methods to be used for large systems where it would be otherwise impractical. Many-body expansions are applicable both to covalently connected monomers, such as the peptide residues in a protein, and to noncovalently connected monomers, such as water molecules in ice, liquid water, or water clusters. Additionally, these methods can be highly parallelized because each fragment can be calculated on a separate processor, making the calculation quite fast. Unfortunately, while the pairwise additive approximation may give qualitatively correct results, it does not give good quantative results, even for small clusters with highly accurate pair potentials.[50] While the inclusion of the three-body terms helps to improve quantitative accuracy,[51] the results may still be insufficiently accurate for systems in which many-body effects play an important role, such as water clusters.[52]

As a result of these shortcomings, some workers have tried to modify the pairwise and three-body approximations by adding additional terms to the monomer, dimer, or trimer Hamiltonian that account for the electrostatic field of the other atoms in the system.[16,17] These methods, which include the pair interaction molecular orbital method[16] (PIMO) and the fragment molecular orbital method[17,24,29,30] (FMO) of Kitaura et al., add terms to the fragment and dimer Hamiltonians that account for the electrostatic potential of the other fragments in the system. These methods involve an iterative procedure to determine the electrostatic potential that occurs in the fragment, dimer, and, in some cases, trimer Hamiltonians. In this procedure one divides the system into the fragments and provides an initial guess for the electron distribution of each fragment. The Schrödinger equation is then solved for all fragments in the system to obtain both the fragment energies and also a new electron density distribution. The new guess is plugged back into the Hamiltonian, and the cycle is repeated until the electron density distributions are converged. FMO calculations in which a unit of two consecutive peptide residues are treated as a monomer have been shown to reproduce ab initio molecular orbital energies of polypeptides to within 2 kcal/mol; however, one might wonder if there is a way to attain similar accuracy without the use of an iterative procedure, particularly for noncovalently connected monomers where one can easily separate the system into fragments without having to cut through any covalent bonds. Some approximations to the electrostatic potential have been made[24] in which the electrostatic potential was treated using a system of Mulliken and fractional point charges, depending on the separation of the fragments, and the expression for the total energy was written in terms of a density difference matrix so that only the net contribution of the electrostatic energy was included. Since this method led to no significant loss of accuracy when compared to the original FMO method, it prompts one to wonder if there is a simpler way to approximate the electrostatic potential, without sacrificing accuracy.

Recent work by Jiang et al.[43] has modified the molecular fragmentation with conjugated caps (MFCC) method of Zhang et al.[25,26] so that the energy of each fragment is calculated in a field of point charges representing the other charge centers. Their method, called electrostatic field-adapted molecular fragmentation with conjugated caps (EFA-MFCC), shows errors within a few millihartees of the ab initio energy for a set of model peptides and biological molecules. In the EFA-MFCC method the total energy is written as the sum of the energies of the capped fragments minus the energy of the conjugated caps (see refs 25, 26, and 43). In light of the discussion of many-body expansions presented in the Introduction, one could consider the energy expression of MFCC and EFA-MFCC to be an electrostatically embedded one-body expansion of the total energy. In light of the good results obtained by Jiang et al. one might ask whether using higher order many-body expansions (i.e., pairwise additive or three-body) in simpe electrostatic fields would be even more accurate, particularly in light of the work of Kitaura et al.[29,30] who have successfully used a many-body expansion with more complicated ways of representing the electrostatic potential (see previous paragraph and refs 24 and 28−30).

In this work, inspired by the successes of the many-body expansions within in a self-consistent field representing the electrostatic potential of the other particles in the work of Federov and Kitaura[29,30] and similar in some respects to the work of Jiang et al.[43] in which a simpler electrostatic embedding is applied to fragments, we combine the use of a many-body expansion with the use of prespecified point charges to represent the electrostatic field of the other molecules. To assess the accuracy of the method we have used it to calculate the binding energies of a series of water clusters ranging in size from trimer to pentamer. Because water clusters are known to have very large many-body effects,[52] this should provide a good test of the capabilities of the method.

2. Theory

Consider a system of N interacting units, called monomers. In the present paper we limit the treatment to noncovalently connected monomers, but, like the fragment molecular orbital method,[29] it can be extended to covalently connected monomers.

Without any approximation, the total energy of the system can be written as

$$V = V_1 + V_2 + V_3 + ... + V_N \quad (1)$$

where

$$V_1 = \sum_i^N E_i \tag{2}$$

$$V_2 = \sum_{i<j}^N (E_{ij} - E_i - E_j) \tag{3}$$

$$V_3 = \sum_{i<j<k}^N [(E_{ijk} - E_i - E_j - E_k) - (E_{ij} - E_i - E_j) - (E_{ik} - E_i - E_k) - (E_{jk} - E_j - E_k)] \tag{4}$$

with higher-order terms defined analogously, and where E_i, E_{ij}, E_{ijk}, ... are energies of embedded monomers, embedded dimers, embedded trimers, and so forth. By an embedded n-mer, we mean one whose energy is calculated in a field representing the $N - n$ other particles. If one retains only the first two terms of eq 1, the total energy of the system can be approximated as

$$E_{\mathrm{EE-PA}} = \sum_{i<j}^N E_{ij} - (N-2)\sum_i^N E_i \tag{5}$$

where EE-PA denotes the electrostatically embedded pairwise additive method. If one includes the first three terms in eq 1, the total energy becomes

$$E_{\mathrm{EE-3B}} = \sum_{i<j<k}^N E_{ijk} - (N-3)\sum_{i<j}^N E_{ij} + \frac{(N-2)(N-3)}{2}\sum_i^N E_i \tag{6}$$

where EE-3B denotes the electrostatically embedded three-body method. (EE-MB denotes an electrostatically embedded many-body approximation of unspecified order.) If one calculates the energies using eqs 5 and 6 without the presence of the point charges, then one gets the conventional pairwise additive (PA) and three-body (3B) energies, as discussed in the Introduction. We emphasize that if the series of eq 1 is not truncated, the result is exact and is independent of whether embedding is employed; however, we will show that the rate of convergence of the series (i.e., the accuracy if one truncates after a given order of many-body terms) depends strongly on embedding.

In the present study, the electrostatic embedding is carried out by using point charges, as in ref 43, to represent the other $N - n$ particles for each n-mer. Two possible ways to obtain the set of point charges needed for these methods are considered in this work:

A. Determine a charge representation for the entire cluster; then, for each monomer, dimer, or trimer, represent the other $N - 1$, $N - 2$, or $N - 3$ water molecules with the charges from this full-system charge calculation.

B. Determine the charges for a gas-phase monomer and, for each monomer, dimer, or trimer, represent the other $N - 1$, $N - 2$, or $N - 3$ monomers with the gas-phase monomer point charges.

In order to compare method A to method B we have used the AM1[53] Mulliken charges[54] to carry out both (denoted AM1 and AM1M for methods A and B, respectively). Since for a very large system the calculation of the charges for the entire cluster may become very expensive we are especially interested in method B (monomer charges), and test the monomer based charge scheme, using three other choices for the monomer charges: B3LYP[55−57]/6-31G*[58] Mulliken charges (denoted B3LYPM), B3LYP/6-31G* class IV charges[59] (denoted CM4M), and TIP3P[61] molecular mechanics charges (denoted TIP3P). (Note that Mulliken charges are class II charges, and CM4 is a class IV charge model.[60]) All AM1 calculations were carried out using the Gaussian03[62] software package, while the B3LYP/6-31G* CM4 and Mulliken charges were calculated using the MN GSM software program, version 6.0.[63] (Note that the B3LYP Mullliken charges can be calculated with any electonic structure package employing B3LYP that can calculate Mulliken charges.)

The procedure for the electrostatically embedded many-body method can be summarized as follows:

1. Determine the partial atomic charges for all N molecules in the cluster by one of the methods described in the previous paragraph.

2. Calculate the energy of each n-body cluster in the expansion (for EE-PA, one has $n = 1$ and 2; for EE-3B, one has $n = 1$, 2, and 3) by representing the atoms of the other $N - n$ molecules with nuclear-centered point charges corresponding to the partial atomic charges determined in step 1.

3. Calculate the total energy of the system using the appropriate equation (i.e., eq 5 for EE-PA, eq 6 for EE-3B).

We note that the internal geometries (O−H bond distances and H−O−H bond angles) of the monomers for those clusters taken from simulations vary within the cluster, and in some cases they are also quite different from those clusters obtained from gas-phase optimizations (the bond lengths and angles for the optimized clusters tend to be similar). Thus in principle, many of the monomers in these clusters have slightly different charges, and indeed when one uses method A the charges are different. One could also laboriously calculate all the slightly different monomer charge sets, but when we use the monomer charges we use a simpler procedure—all charges correspond to the equilibrium structure of the monomer (taken to have the O−H distances of 0.9572 Å and a bond angle of 104.52 degrees).

3. Results and Discussion

In previous work[50] we have tested the ability of the pairwise additive approximation to reproduce full density functional calculations of binding energies of small water clusters ranging in size from trimer to pentamer. The database used in that work consists of a set of eight trimers, six tetramers, and a pentamer; the structures are taken from Monte Carlo and molecular mechanics simulations of bulk water and ice as well as from gas-phase optimizations. We have chosen to use the same set of 15 clusters in this work, and we refer the reader to ref 50 for more information about the database. All tests of the many-body and electrostatically embedded many-body results are tests of how well these approximations can reproduce the total energy from full calculations at a given level of theory not how well they agree with higher-

level calculations. For example EE-PA and EE-3B calculations of the energy with PBE1W/6-311+G(2d,2p) monomers, dimers, and trimers are compared to full PBE1W/6-311+G-(2d,2p) calculations on the whole cluster, which we refer to as the true energies.

While the main subject of the present work is the use of prespecified point charges in a many-body expansion, one might be interested in seeing the results for electrostatic embedding with only the one-body term. For the interested reader these results and a brief discussion have been placed in the Supporting Information.

For each of the 15 clusters we have calculated the true energies at a given level of electronic structure theory and basis set, the pairwise additive energies at the same level of theory and basis set, and the electrostatically embedded pairwise additive energies with that level and basis set. For the EE-PA energies we tested the five charge approximations described above: AM1, AM1M, B3LYPM, CM4M, and TIP3P. Each of these five energies were calculated at five different levels of theory with one or two basis sets. The levels of electronic structure theory are BLYP,[55,56] B3LYP,[55−57] PBE,[64] PBE1W,[49] and MP2.[65] Four of these are methods that were also used in ref 50; they include three generalized gradient approximations (BLYP, PBE, PBE1W), or GGAs, and one hybrid GGA (B3LYP), that also includes a percentage (20%) of Hartree−Fock exchange. The fifth level of theory is MP2 (second-order Møller−Plesset perturbation theory[65]), both to illustrate the electrostatically embedded many-body expansion for a WFT method and because MP2 is known to be highly accurate for small water clusters.[66,67] For each of the density functionals we use the MG3S basis set (which for water is the same as the 6-311+G(2df,2p)[58] basis set), and also the basis set, different for each functional, that was determined previously[50] to be optimal for small water clusters. Three of the functionals (BLYP, B3LYP, and PBE1W) use a Pople basis set,[58] while one functional (PBE) uses an augmented Dunning basis set;[68] in particular we consider the BLYP/6-31+G(d,p), B3LYP/6-31+G(d,2p), PBE/aug-cc-pVTZ, and PBE1W/6-311+G(2d,2p) levels of theory. For the MP2 calculations we use the cc-pVTZ[69] basis set on hydrogen and the aug-cc-pVTZ basis set on oxygen; this combination will be referred to as the aug′-cc-pVTZ basis set. All density functional and MP2 calculations were carried out with the *Gaussian03* program.[62]

Before discussing the results, it is useful to provide some characterization of the clusters. Seven of the eight trimers are bound with respect to three optimized monomers for all nine levels of theory tested. The net binding energies for these seven trimers range, depending on electronic structure level and the particular trimer, from 0.2 to 17.3 kcal/mol, with an average of 13.0 kcal/mol. The only exception, taken from a Monte Carlo simulation of liquid water, is predicted to be unbound by as much as 3.2 kcal/mol by MP2/aug′-cc-pVTZ and to be bound at most by 2.1 kcal/mol with BLYP/6-31+G(d,p). The tetramers are all bound with respect to four optimized monomers; depending on the electronic structure level and the particular tetramer, the net binding energy ranges from 1.9 to 30.8 kcal/mol, with an average binding energy of 13.6 kcal/mol. The pentamer, in contrast, taken from a simulation of liquid water, is unbound by 8.8 to 16.4 kcal/mol, depending on the electronic structure method.

Table 1 shows the mean errors of the pairwise additive and electrostatically embedded pairwise additive energies, relative to the true energies calculated at the same level of theory. We can see that even for the worst EE-PA result, the one with AM1M charges, the average mean unsigned error is reduced by a factor of 2. Moreover, the best two methods, EE-PA-B3LYPM and EE-PA-TIP3P, show a 10-fold reduction in error. For the EE-PA-AM1 and EE-PA-AM1M calculations we see improved performance of the full cluster calculation over the gas-phase monomer charges; however, the difference in the average mean unsigned error in the nine methods is only 0.15 kcal/mol indicating that the use of monomer charges is sufficient for this application. The best overall result comes from the use of the TIP3P charges, which were parametrized (along with two Lennard-Jones parameters) to improve the energetics and liquid density of simulations of liquid water.[60] We note that Jiang et al.[43] saw a similar result, in that while the energy obtained is not independent of the charges used, the choice of charge model does not greatly influence the quality of the result obtained.

As the two best results, EE-PA-B3LYPM and EE-PA-TIP3P, use only the gas-phase monomer charges, which are related by $q_H = -q_O/2$, one might be able to use q_O as a variational parameter in order to obtain even better results. Optimization of q_O to minimize the average mean unsigned error of the nine methods led to $q_O = -0.872$. Table 2 compares the mean errors for the optimized charges (denoted QOPT) to the other EE-PA methods that use monomer charges. We find that optimizing q_O does not lead to significant improvement over the TIP3P charges. In fact, the average mean unsigned errors over all nine methods for the EE-PA approximation with the AM1M, B3LYPM, CM4M, TIP3P, and QOPT charges methods are 0.80, 0.21, 0.34, 0.19, and 0.19 kcal/mol, respectively. In light of the small difference in errors between the B3LYPM, TIP3P, and QOPT methods, and in the interest of developing a method that is generally transferable to other systems, we have decided to use the B3LYP/6-31G* monomer Mulliken charges (B3LYPM) because (unlike the TIP3P charges) they are well defined for any system of interest and thus they avoid the need to optimize a new set of charges for each system of interest.

Table 3 compares the three-body and electrostatically embedded three-body energies to the true energies for the tetramers and pentamer. Based on the results for the electrostatically embedded pairwise additive methods we test only the B3LYP/6-31G* monomer Mulliken (B3LYPM) charges. In all cases we see that the EE-3B method gives a mean unsigned error of less than 0.1 kcal/mol; in the case of MP2/aug′-cc-pVTZ we see a mean unsigned error of only 0.01 kcal/mol. Again we see a minimum of a 2-fold decrease in the error upon inclusion of the point charges and as much as a 10-fold improvement for MP2. The average mean unsigned error for the nine methods is 0.05 kcal/mol, compared to an average mean unsigned error of 0.31 kcal/

Table 1. Mean Errors[a] (kcal/mol) in Binding Energies for the Pairwise Additive Approximation and the Electrostatically Embedded Pairwise Additive Approximation with Five Point-Charge Models

	PA			EE-PA											
				AM1			AM1M			B3LYPM			CM4M			TIP3P		
	MSE	MUE	RMSE	MSE	MUE	RMSE	MSE	MUE	RMSE	MSE	MUE	RMSE	MSE	MUE	RMSE	MSE	MUE	RMSE
BLYP/MG3S	1.99	2.11	2.66	0.74	0.76	1.01	0.91	0.94	1.23	0.21	0.21	0.32	0.42	0.42	0.58	0.14	0.15	0.24
PBE/MG3S	1.62	1.83	2.32	0.39	0.58	0.74	0.56	0.69	0.92	−0.13	0.21	0.32	0.07	0.28	0.38	−0.20	0.23	0.34
PBE1W/MG3S	1.58	1.80	2.28	0.35	0.56	0.72	0.52	0.67	0.89	−0.17	0.24	0.35	0.03	0.28	0.38	−0.24	0.26	0.38
B3LYP/MG3S	1.87	2.00	2.52	0.65	0.68	0.91	0.82	0.85	1.12	0.14	0.14	0.25	0.34	0.34	0.50	0.07	0.10	0.18
BLYP/6-31+G(d,p)	1.81	1.95	2.47	0.74	0.77	1.02	0.90	0.93	1.22	0.27	0.27	0.39	0.46	0.47	0.64	0.21	0.21	0.31
PBE/aug-cc-pVTZ	1.82	2.02	2.53	0.43	0.58	0.76	0.62	0.72	0.98	−0.19	0.25	0.36	0.05	0.27	0.37	−0.27	0.28	0.41
PBE1W/6-311+G(2d,2p)	1.59	1.80	2.29	0.36	0.56	0.72	0.53	0.67	0.90	−0.17	0.24	0.35	0.04	0.28	0.38	−0.23	0.26	0.38
B3LYP/6-31+G(d,2p)	1.84	1.97	2.49	0.68	0.71	0.95	0.84	0.87	1.15	0.17	0.17	0.29	0.37	0.38	0.54	0.11	0.13	0.21
MP2/aug′-cc-pVTZ[b]	1.93	2.04	2.55	0.65	0.68	0.88	0.82	0.85	1.10	0.13	0.13	0.20	0.33	0.33	0.45	0.07	0.09	0.13

[a] MSE, MUE, and RMSE denote mean signed, mean unsigned, and root-mean-squared errors, respectively. [b] aug′-cc-pVTZ denotes using the cc-pVTZ basis set on hydrogen and the aug-cc-pVTZ basis set on oxygen.

mol for the EE-PA scheme applied to the tetramers and pentamer. The average error of 0.05 kcal/mol in the EE-3B energies corresponds to a relative error of 0.4% when compared to the average unsigned net interaction energy of the tetramers and pentamer, which is 13.6 kcal/mol; the average mean unsigned error of 0.31 kcal/mol for the EE-PA scheme applied to this same set of clusters corresponds to 2.3%, which is also small enough to be useful for demanding applications.

To test the new methods for larger clusters, we have applied them to calculate the MP2/aug′-cc-pVTZ energy of a cluster of 21 water molecules (see Figure 1), taken from the Cambridge Cluster Database,[70] which corresponds to the global minimum-energy structure for $N = 21$ with the TIP5P[71] empirical potential.[72] We find that the EE-PA method gives an error of 2.97 kcal/mol relative to the true energy, while EE-3B gives an error of only 0.38 kcal/mol. These values are larger than the mean unsigned errors for the smaller clusters because the 21-mer has a much larger interaction energy, due to the large number of hydrogen bonds (see Figure 1). In particular the energy of the 21-mer is 203.2 kcal/mol lower than 21 separated gas-phase water monomers. The errors in the EE-PA and EE-3B net interaction energies are thus 1.5% and 0.2%, respectively.

To put these results into perspective Table 4 shows FMO results for DFT calculations on water clusters containing 16 and 32 water molecules.[30] In general we see that our errors are lower than those for the FMO methods with $N = 16$. Federov et al. assume that the error scales linearly with the system size,[29,30] indicating that for the same system size the method presented here should be more accurate than FMO. The FMO2 and FMO3 methods perform better for small-basis-set Hartree−Fock calculations, but we have not tested the EE-MB schemes for those levels of calculation, which are less accurate for polarization and which do not include correlation.

It is also interesting to note the wide range in performance that one sees in Table 4 with respect to varying basis set. While we see some dispersion in Tables 1 and 2, the spread is not as severe as what is seen in Table 4, even after adjusting for system size.

An important feature of the EE-PA and EE-3B methods is their favorable scaling in the limit of large system size. For example, the calculation of the EE-3B energy for the 21-mer took only ∼55 h using a single processor on an SGI Altix, while the full MP2/aug′-cc-pVTZ calculation took ∼86 h using eight processors on the same machine (a factor of 12.5 times longer). In theory, in the limit of large system size and independent of what method is used for the trimer, the EE-3B method scales as N^3 where N is the number of monomers. Thus the method shows great promise for allowing calculations of MP2 or CCSD(T) (coupled cluster theory with single and double excitations and quasiperturbative connected triples[73]) accuracy on large systems with N^3 scaling, whereas conventional MP2 and CCSD(T) scale as N^5 and N^7, respectively. The EE-PA method has even more favorable N^2 scaling. If one makes a further approximation of including only pairs or trimers within a certain cutoff distance of each other (for example, only neighboring

Table 2. Comparison of the Mean Errors (kcal/mol) for Four Kinds of Point Charges in the EE-PA Method

	AM1M[a]			B3LYPM[b]			CM4M[c]			TIP3P[d]			QOPT[e]		
	MSE	MUE	RMSE	MSE	MUE	RMSE	MSE	MUE	RMSE	MSE	MUE	RMSE	MSE	MUE	RMSE
BLYP/MG3S	0.91	0.94	1.23	0.21	0.21	0.32	0.42	0.42	0.58	0.14	0.15	0.24	0.07	0.09	0.16
PBE/MG3S	0.56	0.69	0.92	−0.13	0.21	0.32	0.07	0.28	0.38	−0.20	0.23	0.34	−0.26	0.26	0.39
PBE1W/MG3S	0.52	0.67	0.89	−0.17	0.24	0.35	0.03	0.28	0.38	−0.24	0.26	0.38	−0.30	0.30	0.44
B3LYP/MG3S	0.82	0.85	1.12	0.14	0.14	0.25	0.34	0.34	0.50	0.07	0.10	0.18	0.01	0.08	0.13
BLYP/6-31+G(d,p)	0.90	0.93	1.22	0.27	0.27	0.39	0.46	0.47	0.64	0.21	0.21	0.31	0.15	0.15	0.23
PBE/aug-cc-pVTZ	0.62	0.72	0.98	−0.19	0.25	0.36	0.05	0.27	0.37	−0.27	0.28	0.41	−0.35	0.35	0.48
PBE1W/6-311+G(2d,2p)	0.53	0.67	0.90	−0.17	0.24	0.35	0.04	0.28	0.38	−0.23	0.26	0.38	0.01	0.06	0.09
B3LYP/6-31+G(d,2p)	0.84	0.87	1.15	0.17	0.17	0.29	0.37	0.38	0.54	0.11	0.13	0.21	0.04	0.09	0.15
MP2/aug′-cc-pVTZ[f]	0.82	0.85	1.10	0.13	0.13	0.20	0.33	0.33	0.45	0.07	0.09	0.13	−0.30	0.30	0.43

[a] $q_O = -0.385115$, $q_H = 0.1925575$. [b] $q_O = -0.778311$, $q_H = 0.3891555$. [c] $q_O = -0.637$, $q_H = 0.319$. [d] $q_O = -0.834$, $q_H = 0.417$. [e] $q_O = -0.8972$, $q_H = 0.4486$. [f] aug′-cc-pVTZ using the cc-pVTZ basis set on hydrogen and the aug-cc-pVTZ basis set on oxygen.

Table 3. Mean Errors (kcal/mol) for Three-Body and Electrostatically Embedded Three-Body Binding Energies

	3B			EE-3B[a]		
	MSE	MUE	RMSE	MSE	MUE	RMSE
BLYP/MG3S	−0.07	0.18	0.29	−0.02	0.03	0.05
PBE/MG3S	−0.13	0.18	0.32	−0.08	0.08	0.11
PBE1W/MG3S	−0.14	0.17	0.32	−0.09	0.09	0.12
B3LYP/MG3S	−0.07	0.16	0.27	−0.02	0.03	0.04
BLYP/6-31+G(d,p)	−0.06	0.17	0.27	−0.02	0.03	0.04
PBE/aug-cc-pVTZ	−0.10	0.16	0.29	−0.07	0.07	0.09
PBE1W/6-311+G(2d,2p)	−0.13	0.17	0.32	−0.08	0.08	0.12
B3LYP/6-31+G(d,2p)	−0.07	0.16	0.25	−0.02	0.02	0.04
MP2/aug′-cc-pVTZ[b]	−0.05	0.14	0.21	−0.01	0.01	0.01

[a] Based on B3LYPM point charges. [b] aug′-cc-pVTZ denotes the use of the cc-pVTZ basis set on hydrogen and the aug-cc-pVTZ basis set on oxygen.

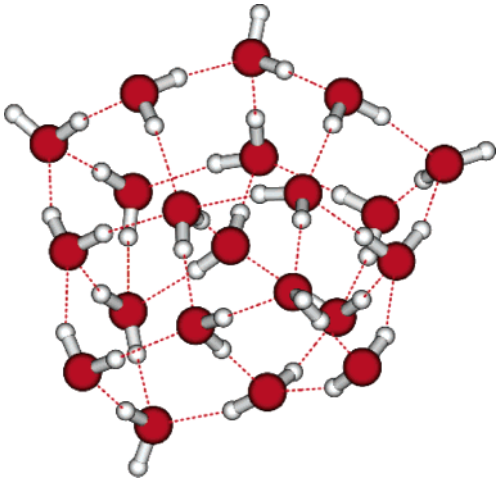


Figure 1. Geometry of 21-mer used to test EE-PA and EE-3B for large clusters.

monomers), the method scales linearly in N. The use of a cutoff has particular value if one is interested in using periodic boundary conditions, as it would allow for the study of a large periodic system with highly accurate quantum mechanical methods at a relatively low cost. This further approximation was not made in the present work but is an interesting topic for further study.

Another important issue to emphasize is simplicity. Since essentially all modern electronic structure packages allow calculations in the field of point charges (or the one-electron part of the Hamiltonians can be easily modified to allow this), the present EE-PA and EE-3B methods can be carried out by writing a simple script to drive virtually any electronic structure package. Furthermore, if a new variant of correlated WFT becomes available, for example a new coupled cluster approximation, it can immediately be employed for the n-mers of the electrostatically embedded many-body series without method-specific programming. An additional benefit is that performing an EE-PA or EE-3B calculation is equivalent to carrying out a full quantum mechanical calculation on the system, unlike combined QM/MM models[13,20,22] that treat only part of the system quantum mechanically.

Table 4. Errors (kcal/mol) in FMO2[a] and FMO3[b] Approximations to True Energies for Clusters of N Water Molecules[c]

	$N = 16$	$N = 32$
FMO2		
B3LYP/6-31G*	−9.83	−26.27
B3LYP/6-31++G**	−6.78	−23.58
FMO3		
B3LYP/6-31G*	0.51	1.85
B3LYP/6-31++G**	3.12	14.71

[a] FMO2 corresponds to the two-body fragment molecular orbital method with one water molecule per fragment. [b] FMO3 corresponds to the three-body fragment molecular orbital method with one water molecule per fragment. [c] All errors are taken from ref 30.

Finally we emphasize the linearity of the method. Equations 1−6 are all linear in the energies. Thus if analytic gradients or Hessians are available for the monomers, they are immediately available for the EE-MB energy. For example

$$\nabla E_{\mathrm{EE-PA}} = \sum_{i<j}^{N} \nabla E_{ij} - (N-2)\sum_{i}^{N} \nabla E_i \qquad (7)$$

In contrast to this simplicity, most other fragment-based methods require extra programming or even assumptions to obtain analytic gradients. Fast and accurate gradients are essential for dynamics calculations.

4. Summary and Conclusions

We present here a many-body series called the electrostatically embedded many-body expansion that merges the many-

body expansion, as developed extensively by Federov and Kitaura,[29,30] with the use of prespecified point charges on fragments as also proposed by Jiang et al.[43] We have implemented it using fixed monomer partial atomic charges for the electrostatic embedding, and we have demonstrated its accuracy for a set of water clusters ranging in size from trimer to 21-mer. For 8 trimers, 6 tetramers, and a single pentamer, the electrostatically embedded pairwise additive approximation yields a 10-fold reduction in error as compared to the conventional pairwise additive method, with an average mean unsigned error of approximately 0.21 kcal/mol when the B3LYP/6-31G* monomer Mulliken charges are used. When three-body terms are included with the same set of charges, the average mean unsigned error is only 0.05 kcal/mol, as compared to 0.17 kcal/mol if no electrostatic embedding is used, an improvement of more than a factor of 3. In the case of the 21-mer, the EE-PA approximation leads to an error in the total energy of only 2.97 kcal/mol, while the EE-3B approximation gives an error of only 0.38 kcal/mol relative to a full MP2/aug′-cc-pVTZ calculation.

While we have only tested the method up to the three-body terms, it is easily generalized to include as many of the *n*-body terms as desired. In the future we hope to apply this method to the calculation of binding energies for much larger clusters and to evaluate possible choices of charge model for other types of molecular clusters.

The MBPAC software package for running EE-PA and EE-3B calculations is available free of charge and can be downloaded at http://comp.chem.umn.edu/mbpac/.

Acknowledgment. This work was supported in part by the National Science Foundation under grant nos. CHE03-49122 and ITR-0428774.

Supporting Information Available: Results for truncating the many-body expansion after V_1. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Newton, M. D.; Boer, F. P.; Lipscomb, W. N. *J. Am. Chem. Soc.* **1966**, *88*, 2353.

(2) Degand, P.; Leroy, G.; Peeters, D. *Theor. Chim. Acta* **1973**, *30*, 243.

(3) Niessen, W. V. *Theor. Chim. Acta* **1973**, *31*, 111.

(4) Stechel, E. B. In *Domain-Based Parallelism and Problem Decomposition in Computational Science and Engineering*; Keyes, D. R., Saad, Y., Truhlar, D. G., Eds.; SIAM: Philadelphia, PA, 1995; p 217.

(5) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170.

(6) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Seiber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357.

(7) Humbel, S.; Sieber, S.; Morokuma, K. *J. Chem. Phys.* **1996**, *105*, 1959.

(8) Lee, T.-S.; York, D. M.; Yang, W. *J. Chem. Phys.* **1996**, *105*, 2744.

(9) Svensson, M.; Humbel, S.; Morokuma, K. *J. Chem. Phys.* **1996**, *105*, 3654.

(10) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chem. Phys. Lett.* **1996**, *253*, 268.

(11) Challacombe, M.; Schwegler, E. *J. Chem. Phys.* **1997**, *106*, 5526.

(12) Sierka, J.; Sauer, J. *Faraday Discuss. Chem. Soc.* **1997**, *106*, 41.

(13) *Combined Quantum Mechanical and Molecular Mechanical Methods*; Gao, J., Thompson, M. A., Eds.; ACS Symp. Ser. 712; American Chemical Society: Washington, DC, 1998.

(14) Galli, G. *Comput. Mater. Sci.* **1998**, *12*, 242.

(15) Scuseria, G. E.; Ayala, P. Y. *J. Chem. Phys.* **1999**, *111*, 8330.

(16) Kitaura, K.; Sawai, T.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *312*, 319.

(17) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701.

(18) Amara, P.; Field, M. J.; Alhambra, C.; Gao, J. *Theor. Chem. Acc.* **2000**, *104*, 336.

(19) Papanicolaou, N. I.; Kallinteris, G. C.; Evangelakis, G. A.; Papaconstantopoulos, D. *Comput. Mater. Sci.* **2000**, *17*, 224.

(20) Sherwood, P. In *Modern Methods and Algorithms of Quantum Chemistry*; Grotendurst, J., Ed.; NIC Ser. 3; Neumann Institute for Computing: Jülich, 2000; p 285.

(21) Nakano, T.; Kaminuma, T.; Sato, T.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chem. Phys. Lett.* **2000**, *318*, 614.

(22) Gao, J.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467.

(23) Titmuss, S. J.; Cummins, P. L.; Rendall, A. P.; Bliznyuk, A. A.; Gready, J. E. *J. Comput. Chem.* **2002**, *23*, 1314.

(24) Nakano, T.; Kaminuma, T.; Sato, T.; Fukuzawa, K.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chem. Phys. Lett.* **2002**, *351*, 475.

(25) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599.

(26) Zhang, D. W.; Chen, X. H.; Zhang, J. Z. H. *J. Comput. Chem.* **2003**, *24*, 1846.

(27) Assfeld, X.; Ferré, N.; Rivail, J.-L. *Theor. Chem. Acc.* **2004**, *111*, 328.

(28) Federov, D. G.; Olson, R. M.; Kitaura, K.; Gordon, M. S.; Koseki, S. *J. Comput. Chem.* **2004**, *25*, 872.

(29) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *120*, 6832.

(30) Fedorov, D. G.; Kitaura, K. *Chem. Phys. Lett.* **2004**, *389*, 129.

(31) Re, S.; Morokuma, K. *Theor. Chem. Acc.* **2004**, *112*, 59.

(32) Miyazaki, T.; Bowler, D. R.; Choudhury, R.; Gillan, M. J. *J. Chem. Phys.* **2004**, *121*, 6186.

(33) Staszewska, G.; Staszewski, P.; Schultz, N. E.; Truhlar, D. G. *Phys. Rev. B* **2005**, *71*, 45423.

(34) Wada, M.; Sakurai, M. *J. Comput. Chem.* **2005**, *26*, 160.

(35) He, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 31103.

(36) Deev, V.; Collins, M. A. *J. Chem. Phys.* **2005**, *122*, 154102.

(37) Monard, G.; Bernal-Uruchurtu, van der Vaart, A.; Merz, K. M., Jr.; Ruiz-López, M. F. *J. Phys. Chem. A* **2005**, *109*, 3425.

(38) Federov, D. G.; Kitaura, K. *J. Chem. Phys.* **2005**, *123*, 134103.

(39) Lin, H.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 3991.

(40) Federov, D. G.; Kitaura, K.; Li, H.; Jensen, J. H.; Gordon, M. S. *J. Comput. Chem.* **2006**, *27*, 976.

(41) Ishida, T.; Federov, D. G.; Kitaura, K. *J. Phys. Chem. B* **2006**, *110*, 1457.

(42) Canfield, P.; Dahlborn, M. G.; Hush, N. S.; Reimers, J. R. *J. Chem. Phys.* **2006**, *124*, 24301.

(43) Jiang, N.; Ma, J.; Jiang, Y. *J. Chem. Phys.* **2006**, *124*, 114112.

(44) He, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *124*, 184703.

(45) Vashista, P.; Kalia, R. K.; Nakano, A. *J. Phys. Chem. B* **2006**, *110*, 3727.

(46) Pople, J. A. *Rev. Mod. Phys.* **1999**, *71*, 1267.

(47) Kohn, W. *Rev. Mod. Phys.* **1999**, *71*, 1253.

(48) Bowen, J. P.; Allinger, N. L. *Rev. Comput. Chem.* **1991**, *2*, 81.

(49) Dinur, U.; Hagler, A. T. *Rev. Comput. Chem.* **1991**, *2*, 99.

(50) Dahlke, E. E.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *109*, 15677.

(51) Elrod, M. J.; Saykally, R. J. *Chem. Rev.* **1994**, *94*, 1975.

(52) Xantheas, S. S. *J. Chem. Phys.* **1994**, *100*, 7523.

(53) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

(54) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833.

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

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

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

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

(59) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Theor. Chem. Acc.* **2005**, *113*, 133.

(60) Storer, J. W.; Giesen, D. J.; Cramer. C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87.

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

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

(63) Chamberlin, A. C.; Kelly, C. P.; Thompson, J. D.; Xidos, J. D.; Li, J.; Hawkins, G. D.; Winget, P. D.; Zhu, T.; Rinaldi, D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G.; Frisch, M. J. *MN-GSM, version 6.0*; University of Minnesota: Minneapolis, MN 55455-0431, 2006.

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

(65) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

(66) Xantheas, S. S.; Burnham, C. J.; Harrison, R. J. *J. Chem. Phys.* **2002**, *116*, 1493.

(67) Xantheas, S. S.; Aprà, E. *J. Chem. Phys.* **2004**, *120*, 823.

(68) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.

(69) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.

(70) Wales, D. J.; Doye, J. P. K.; Dullweber, A.; Hodges, M. P.; Naumkin, F. Y.; Calvo, F.; Hernández-Rojas, J.; Middleton, T. F. *Cambridge Cluster Database*. http://www-wales.ch.cam.ac.uk/CCD.html (accessed March 8. 2006).

(71) Mohoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910.

(72) James, T.; Wales, D. J.; Hernández-Rojas, J. *Chem. Phys. Lett.* **2005**, *415*, 302.

(73) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.

CT600253J

JCTC Journal of Chemical Theory and Computation

A Density Functional Study of Methanol Clusters

Susan L. Boyd*,† and Russell J. Boyd‡

Department of Chemistry, Mount Saint Vincent University, Halifax, N.S. B3M 2J6, Canada, and Department of Chemistry, Dalhousie University, Halifax, N.S. B3H 4J3, Canada

Received September 20, 2006

Abstract: The potential energy surfaces of methanol clusters, $(CH_3OH)_n$, $n = 2-12$, have been studied using density functional theory at the B3LYP/6-31G(d) and higher levels of theory. Cyclic clusters in which n methanol molecules are joined in a ring structure formed by n hydrogen bonds are shown to be more stable than structures of the same number of methanol molecules where one or more methanol molecules are outside the ring and are hydrogen-bonded to oxygens of methanols in rings of $n - 1$, $n - 2$, and so forth. So-called chain structures are generally even less stable. Furthermore, the hydrogen-bonding energy per methanol molecule of the n-ring clusters is shown to converge to an asymptotic value of about 27 kJ/mol at B3LYP/6-311+G-(d,p)//B3LYP/6-31G(d) after five to six methanols are included in the cluster. As expected, there are many minima on the potential energy surfaces of the methanol clusters, the number increasing rapidly with n. A cyclic cluster of five to six methanol molecules appears to be sufficient to mimic liquid behavior as far as vibrational frequencies are concerned.

Introduction

Like water, liquid methanol displays anomalous behavior, including a high boiling point, due to hydrogen bonding. Unlike water, whose bulk properties are entirely governed by hydrogen bonding, methanol has both hydrogen bonding and hydrophobic interactions. The latter are due to the presence of the methyl group, which imposes a constraint on the hydrogen-bonding network in condensed phases.

It is well-known that each oxygen in water is able to act as a donor for two hydrogen bonds and as an acceptor for two hydrogen bonds; therefore, in the solid state, the coordination number of oxygen is four. Liquid water is much less ordered than the solid, but nonetheless, the "structure" and properties are dominated by hydrogen bonding. The oxygen atom in methanol is limited to being a single hydrogen-bond donor. In principle, each oxygen in methanol may be an acceptor for two hydrogen bonds, but it is much more probable that a given oxygen atom will be an acceptor for only one hydrogen bond. The latter point is supported by experimental studies of crystalline methanol which clearly show that the methanol molecules are connected by hydrogen bonds to form one-dimensional structures.[1,2] This simple difference between methanol and water has important consequences for the structure of liquid methanol.

The investigation of the structure of liquid methanol has a long history and remains an active research topic.[3,4] In his classic treatise, Pauling[5] proposed that the molecules in liquid methanol form cyclic hexamer structures. Numerous experimental and theoretical studies have been designed to test Pauling's hypothesis. Sarkar and Joarder[6] have provided support on the basis of an analysis of earlier neutron diffraction[7,8] and X-ray scattering[9,10] experiments. Recent resonant soft X-ray emission spectroscopy experiments by Kashtanov et al.,[11] supported by density functional theory (DFT), suggest that liquid methanol consists of combinations of rings and chains of methanol molecules linked with hydrogen bonds and is dominated by hexamers and octamers. Wilson et al.[12] have used X-ray absorption spectroscopy and DFT-computed spectra to arrive at very similar conclusions. Thus, there appears to be a convergence of conclusions from experimental studies on the structure of liquid methanol.

* Corresponding author fax: 902-457-6134; e-mail: susan.boyd@msvu.ca.

† Mount Saint Vincent University.

‡ Dalhousie University.

10.1021/ct6002912 CCC: $37.00
Published on Web 11/15/2006

Additional information on the electronic and thermodynamic properties of liquid methanol has been obtained from a variety of experimental studies.[13−19]

Many theoretical studies of liquid methanol have been reported. The early Monte Carlo work of Jorgensen,[20] the classical molecular dynamics (MD) simulation of Haughney et al.,[21] and the latest ab initio MD simulations[22−24] all favor the existence of chains, but there is no agreement on the average chain length. It has been suggested[11] that MD simulations based on potentials derived from the methanol dimer do not predict ring structures because of significant differences between the polarization in the dimer and larger rings. Potentials that provide a better description of hydrogen bonding in ring structures are required for future MD simulations.

The red-shifted hydroxyl-stretching mode in liquid methanol has been the subject of several studies,[25,26] including the first ab initio MD simulation of liquid methanol,[27] while others have attempted to account for thermodynamic properties.[28−33] Many of the studies have employed hybrid quantum mechanical/molecular mechanical (QM/MM) methods.[34,35] Despite the many MD and QM/MM studies of liquid methanol, it is surprising that so few studies by conventional electronic structure methods of methanol clusters in the gas phase have been published. The most comprehensive study to date is the work of Mó et al.[36] on the methanol trimer, in which high-level calculations showed that the most favorable structure is a ring with one methyl group above the ring and two below the ring. Mandado et al.[37] have studied cooperative effects in the methanol trimer. Vener and Sauer[38] have reported an MP2 and B3LYP study of two cyclic isomers of the methanol tetramer. B3LYP/6-31+G(d) calculations on clusters as large as the octamer have been reported by Ludwig;[39] by combining the DFT calculations with a quantum cluster equilibrium model, he concludes that liquid methanol is dominated by cyclic and/or lasso structures. El-Shall et al.[40] have reported Monte Carlo simulations of $(CH_3OH)_n$, $n = 2-9$, clusters.

Despite the extensive literature on the structures of methanol clusters, only the trimer surface has been explored thoroughly.[36] For the methanol tetramer, we have located 12 stationary points on the surface corresponding to minima, whereas only two have been reported previously.[38] In this paper, we use a reliable level of electronic structure theory to study methanol clusters as large as the dodecamer, $(CH_3OH)_{12}$. Our goal is to complement earlier work on clusters and to stimulate MD simulations with improved potentials.

Nomenclature

We have considered various types of clusters.

(i) A cluster of n methanol molecules, $(CH_3OH)_n$, where all n molecules are joined in a single ring structure with n hydrogen bonds, is termed an nr-$n$mer, for example, the methanol 6r-hexamer or the 6r-hexamer.

(ii) Systems which contain a ring and branches from that ring, where the branches are x methanol "monomers" hydrogen-bonded to an oxygen in the ring, are called $((n - x)\mathrm{r} + x\mathrm{m})$-$n$mer; for example, a hexamer formed from a tetramer ring with two single methanol branches from the ring is called a $(4\mathrm{r} + 2\mathrm{m})$-hexamer.

(iii) We have also studied a few examples in which the branch is two or more methanols in length; that is, the branch is a "dimer", "trimer", and so forth; these are called $((n - x)\mathrm{r} + x/2\ \mathrm{d})$-$n$mer, and so forth, as in the case of a pentamer formed from a three-membered ring with one dimer branch, $(3\mathrm{r} + 1\mathrm{d})$-pentamer.

(iv) When there is reason to do so for clarification, we have used u and d as a subscript for up and down, respectively, to indicate the direction of the methyl groups relative to the (approximately) planar arrangement of the O−H bonds participating in the rings; thus, the lowest-energy trimer ring structure has a *udu* arrangement, as noted previously by Mó et al.,[36] and is called the $3\mathrm{r}_{udu}$-trimer.

(v) Sometimes, to distinguish between different "isomers", we have had to clarify where the branching occurs, as in $(4\mathrm{r}_{uddd} + 1\mathrm{m}_{d2})$-pentamer. This notation means that the single methanol branch on the four-membered ring of the pentamer occurs at the second "down" methanol in the ring.

(vi) Note also that some structures, in addition to normal H bonds, have a weak interaction between the O of a hydroxyl group and the H of a nearby methyl group. When this occurs, it is indicated, as in the $(3\mathrm{r} + 1\mathrm{m}_{\mathrm{O}\cdots\mathrm{HC}})$-tetramer. In this structure, the methanol outside the three-membered ring in a sense straddles the ring, via both a normal H bond and the extra weak O···HC interaction. The O···HC interaction closes a ring; the smallest example was described for the trimer by Mandado et al.[37]

(vii) Finally, we have considered structures where the methanols are H-bonded into chains; these are referred to as nc-$n$mers, for example, the 6c-hexamer, and all of these have one or more inter-O···HC actions, which are either "tight" as for the trimer O···HC ring of Mandado et al.[37] or "loose" when it is a tetramer that is closed by the O···HC interaction.

Clearly, as the value of n increases, the number of "isomers" for an $n$mer grows enormously; branched chain structures could exist, and so on. We have not attempted to optimize all possible structures for any $n$mers except the tetramers.

Computational Methods

All density functional theory calculations reported herein were carried out using the Gaussian 03 suite of programs.[41] The geometry optimizations on the $(CH_3OH)_n$, $n = 2-12$, clusters were performed using the B3LYP functional and the 6-31G(d) basis set. The B3LYP functional is composed of Becke's three-parameter hybrid exchange functional (B3),[42,43] as implemented in Gaussian 03,[44] and the correlation functional of Lee, Yang, and Parr (LYP).[45] Harmonic vibrational frequencies and zero-point vibrational energy (ZPVE) corrections were calculated at the same level of theory. A few additional calculations with larger basis sets were carried out as a check on the effects of basis set truncation. Relative energies were determined by single-point studies with the B3LYP functional and the 6-311+G(d,p) basis set, as discussed below.

Table 1. Selected Bond Lengths from Optimized Geometries of Methanol and Methanol Trimer Global Minimum Energy Structures, as Obtained with Various Basis Sets

	bond lengths (Å)[a]					
	methanol		$3r_{udu}$-trimer			
level	C−O	O−H	O···O	C−O	O−H	O···H
B3LYP/6-31G	1.452	0.978	2.634	1.449	1.002	1.721
			2.641	1.447	1.002	1.736
			2.627	1.447	1.001	1.717
			(2.634)	(1.448)	(1.002)	(1.725)
B3LYP/6-31G(d)	1.418	0.969	2.740	1.423	0.986	1.825
			2.753	1.421	0.987	1.844
			2.741	1.421	0.986	1.829
			(2.744)	(1.422)	(0.986)	(1.833)
B3LYP/6-31G(d,p)	1.418	0.965	2.743	1.422	0.982	1.833
			2.756	1.420	0.983	1.852
			2.745	1.421	0.982	1.836
			(2.748)	(1.421)	(0.982)	(1.840)
B3LYP/6-31+G(d)	1.425	0.969	2.765	1.426	0.984	1.863
			2.778	1.424	0.984	1.886
			2.766	1.424	0.983	1.866
			(2.770)	(1.425)	(0.984)	(1.872)
B3LYP/6-31+G(d,p)	1.426	0.965	2.770	1.425	0.980	1.873
			2.782	1.424	0.901	1.986
			2.771	1.424	0.979	1.876
			(2.774)	(1.424)	(0.953)	(1.912)
B3LYP/6-311+G(d,p)[b]	1.424	0.961	2.765	1.424	0.976	1.876
			2.778	1.422	0.976	1.874
			2.770	1.422	0.975	1.902
			(2.771)	(1.423)	(0.976)	(1.884)
MP2/6-311+G(d,p)[c]	1.421	0.959	2.761	1.422	0.972	1.879
			2.776	1.421	0.972	1.896
			2.763	1.421	0.973	1.867
			(2.767)	(1.421)	(0.972)	(1.881)

[a] Average values for the $3r_{udu}$-trimer bond lengths are in parentheses. [b] From Mó et al.[36] [c] From Mó et al.[36] Supporting Information.

Results and Discussion

A. Choice of the Computational Method and Basis Set. Given that our goal was to study larger clusters than those studied previously, our first objective was to identify an appropriate level of theory. For this purpose, we used the study of the methanol $3r_{udu}$-trimer potential energy surface by Mó et al.[36] as a reference. Our target was a basis set which yields bond lengths that agree to within 0.01 Å with the Mó et al. calculations and intermolecular distances that agree to within 0.05 Å. We assumed that the four internuclear distances, *r*(O···O), *r*(C−O), *r*(O−H), and *r*(O···H), were sufficient comparators for this purpose. Furthermore, our goal was to find a level of theory that yielded dissociation energies that agree to within 5 kJ/mol of the Mó et al. calculations for the trimer.

Before entering into a detailed discussion of the choice of the computational method and basis set, it should be noted that it would be desirable to use the MP2 method with a fairly large basis set, such as the aug-CC-pVDZ basis set. Unfortunately, with our computational facilities, we cannot get beyond the pentamer with MP2/aug-CC-pVDZ. Given that our goal is to be able to include the dodecamer and to treat all clusters at a uniform level, we were forced to compromise. Methods based on density functional theory offer the best alternative. It is well-known, however, that DFT methods fail to provide accurate descriptions of weak interactions. Such problems are most severe for dispersion-bound systems. But it is also well-established[46,47] that some of the commonly used functionals provide an adequate description of relatively strong hydrogen-bonded interactions, such as the O−H···O interactions that are present in liquid methanol.

For the purposes of the following discussion, we define the term "dissociation energy" as

$$D_0 = nE(\mathrm{CH_3OH}) - E(n\mathrm{mer}) \qquad n = 2-12$$

where the terms $E(n\mathrm{mer})$ and $E(\mathrm{CH_3OH})$ include the corresponding ZPVE (unscaled) but not the basis set superposition error or thermal corrections. (Our aim was to observe trends as the size of the methanol cluster increases, rather than to attain the experimental accuracy sought by Mó et al. for the small trimer system.)

Table 1 lists the key geometrical parameters of the optimized geometries of the $3r_{udu}$-trimer structure obtained with the B3LYP method and the 6-31G, 6-31G(d), 6-31G(d,p), 6-31+G(d), and 6-31+G(d,p) basis sets; these are compared with the same parameters obtained by Mó et al., who used

Table 2. Dissociation Energies (D_0) for the Methanol $3r_{udu}$-Trimer

	methanol		$3r_{udu}$-trimer		
level of theory	*E* (au)	ZPVE (au)	*E* (au)	ZPVE (au)	D_0 (kJ/mol)
Full Optimization Study					
B3LYP/6-31G	−115.679 50	0.051 07	−347.086 91	0.159 86	109.6
B3LYP/6-31G(d)	−115.714 41	0.051 48	−347.181 74	0.160 99	83.9
B3LYP/6-31G(d,p)	−115.723 96	0.051 41	−347.209 46	0.160 46	82.3
B3LYP/6-31+G(d)	−115.725 19	0.051 31	−347.205 02	0.160 05	61.2
B3LYP/6-31+G(d,p)	−115.734 87	0.051 23	−347.232 26	0.159 42	57.6
B3LYP/6-311+G(d,p)	−115.764 94	0.051 05	−347.322 09	0.158 82	56.7
single point, B3LYP/6-31G(d) geometry					
B3LYP/6-31G(d)	−115.714 41	0.051 48	−347.181 74	0.160 99	83.9
B3LYP/6-31G(d,p)	−115.723 94	0.051 48	−347.209 44	0.160 99	81.6
B3LYP/6-31+G(d)	−115.725 06	0.051 48	−347.204 07	0.160 99	58.6
B3LYP/6-31+G(d,p)	−115.734 74	0.051 48	−347.231 09	0.160 99	53.4
B3LYP/6-311+G(d,p)	−115.764 78	0.051 48	−347.320 46	0.160 99	51.4
Mó, Yáñez and Elguero[a]					
B3LYP/6-311++G(3df,2p)//6-311+G(d,p)	−115.773 09	0.051 04	−347.343 30	0.158 80	48.2

[a] From Mó et al.[36] Supporting Information, except *without* thermal energy, BSSE corrections, or scaling of ZPVE, to allow direct comparison with our data.

B3LYP/6-311+G(d,p) and MP2/6-311+G(d,p) optimized geometries. The B3LYP/6-31G structural parameters agree poorly with the B3LYP/6-311+G(d,p) results. The addition of d-type polarization functions on the heavy atoms eliminates most of the error, while further improvement in the basis set leads to more modest incremental improvements. The B3LYP/6-31G(d) bond lengths for the methanol monomer are within 0.01 Å of the B3LYP/6-311+G(d,p) ones of Mó et al.; for the trimer, C−O and O−H bonds are similarly within 0.01 Å. The intermolecular distances vary to a slightly greater extent: the B3LYP/6-31G(d) calculations yield shorter intermolecular distances, for example, an average O···H bond of 1.83 Å compared to the B3LYP/6-311+G(d,p) result of 1.88 Å, and an average O···O distance of 2.74 Å compared to 2.77 Å. In view of these observations and in order to make large clusters accessible with our computational facilities, we conclude that the B3LYP/6-31G(d) optimization of geometries of methanol clusters, which yields parameters within 0.01 Å for bond lengths and 0.05 Å for intermolecular distances, is adequate.

Table 2 lists the dissociation energy obtained for the $3r_{udu}$-trimer global minimum energy structure with each of the various basis sets listed. Given that we have aimed to attain dissociation energies for this trimer that agreed within 5 kJ/mol with those obtained in the comprehensive study of Mó et al., we have used single-point energies from the B3LYP/6-311+G(d,p) level of theory, calculated with the B3LYP/6-31G(d) geometry and frequency. This yields a stabilization energy of 51.4 kJ/mol for the trimer, which compares well with the 48.2 kJ/mol obtained by Mó et al. who used a much larger basis set, B3LYP/6-311++G(3df,2p), for their single-point study, together with a geometry optimized with B3LYP/6-311+G(d,p). (Note: the Mó et al. paper reports results calculated by including thermal energies, scaled ZPVE, and BSSE corrections; we have recalculated their data without these corrections, so that their results may be directly compared to ours.)

B. Geometries of the *n*mers. As noted above, the number of possible "isomers" for an *n*mer increases rapidly with *n*. To illustrate this, Figure 1 shows the structures of the 12 tetramers, both "top" and "side" views, for which minima (no imaginary frequencies) were found on the B3LYP/6-31G(d) surface. B3LYP/6-311+G(d,p) single-point energies of each of the conformers (including B3LYP/6-31G(d) ZPVE), relative to that of the lowest-energy one, are given in kilojoules per mole. The complete set of structures and geometrical parameters is available as Supporting Information.

C. Dissociation Energies of the Methanol Clusters. In Figure 2 are plotted the dissociation energies [B3LYP/6-311+G(d,p) single-point energies obtained with B3LYP/6-31G(d) geometries and ZPVE] per mole of methanol for the various ($n = 1-12$) methanol clusters. The lowest-energy species is plotted for each type of ring for a given *n*mer. For example, while four 4r-tetramer structures were found (*udud*, *uudd*, *uuud*, and *uuuu*), only the energy for the *udud* structure appears on the plot. Similarly, the lowest energy of the six (3r+1m)-tetramer structures is given, and so forth. There are many points worth noting.

(1) Clearly there is an increase in the dissociation energy per mole of methanol as the number of methanol molecules in the cluster increases to about six; that is, the larger clusters are more stable than the smaller ones until the cluster reaches about six methanol molecules. At that point, the D_0/mole of methanol essentially levels off.

(2) The curve for the largest rings of a given *n* in Figure 2 is not smooth as *n* increases. This may be due to the possibility that the lowest-energy ring species was not found in each case. The potential energy surface is very flat for the different conformers of the clusters, as has been noted by Mó et al. for the trimer,[36] and it is quite possible that, despite many searches in each case, the global minimum energy structure was not found for some of these clusters.

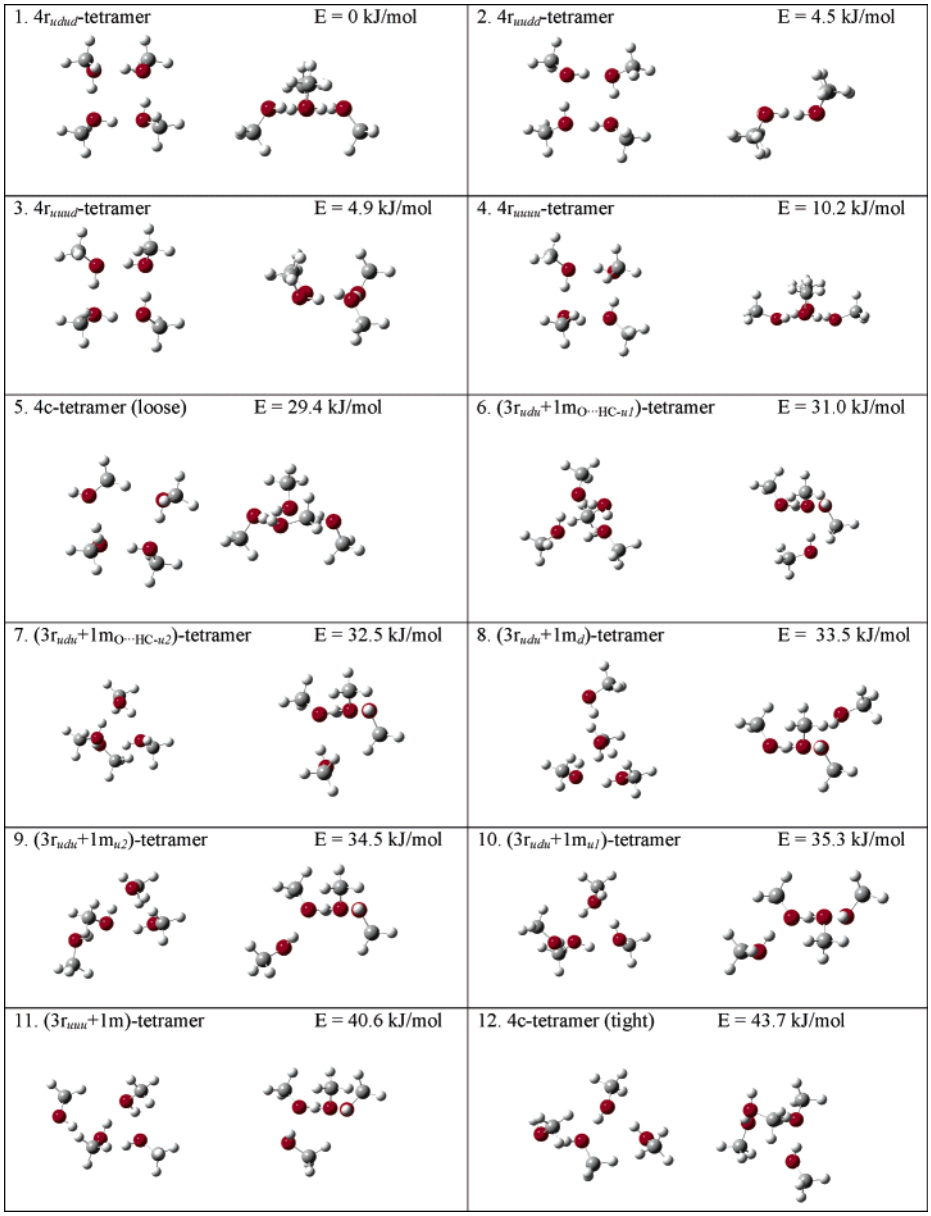


Figure 1. Minimum energy structures for the methanol tetramers. B3LYP/6-311+G(d,p) single-point energies, including B3LYP/6-31G(d) ZPVE, relative to that of the global minimum energy structure are indicated (kJ/mol).

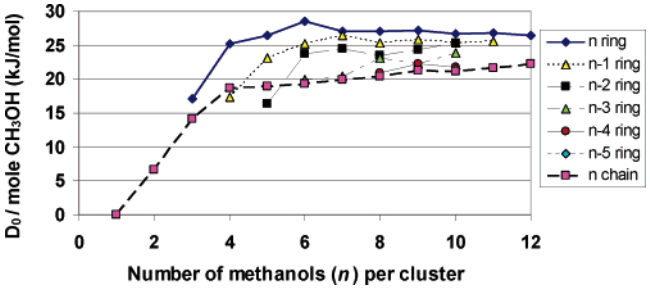


Figure 2. Dissociation energies of methanol clusters per mole of methanol as a function of the number of methanol molecules, at the B3LYP/6-311+G(d,p)//6-31G(d) level with B3LYP/6-31G(d) ZPVE included.

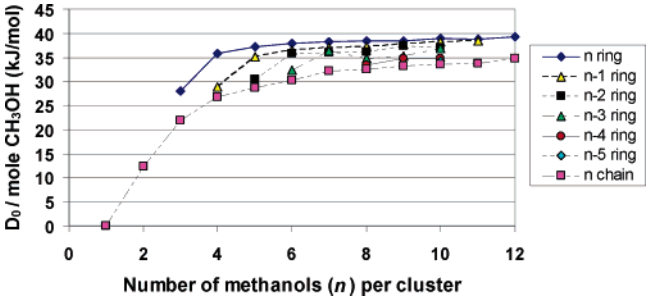


Figure 3. Dissociation energies of methanol clusters per mole of methanol as a function of the number of methanol molecules, at the B3LYP/6-31G(d) level with ZPVE corrections included.

Another explanation may be the fact that the energies were obtained with a larger basis set than the geometries and ZPVE. When B3LYP/6-31g(d) energies from optimized geometries and ZPVE are used, the curve is somewhat smoother (Figure 3). A third possibility is that the fluctuations may be real and may reflect, for example, that the 6r-hexamer, as predicted long ago by Pauling,[5] is the most stable cyclic methanol cluster.

Table 3. Dissociation Energy of Pentamer Clusters per Mole of Pentamer and per Mole of Methanol at B3LYP/6-311+G(d,p)//B3LYP/6-31G(d)

	cluster	*E* au	ZPVE au	ΔE conf[a] kJ/mol	D_0 kJ/mol	D_0/ mole CH_3OH kJ/mol
1	$5r_{ududu}$-pentamer "planar"	−578.885 23	0.268 26	0	132	26
2	$5r_{uuuud}$-pentamer	−578.884 88	0.268 15	0.91	132	26
3	$5r_{ududu}$-pentamer "envelope"	−578.883 39	0.268 97	4.83	126	25
4	$(4r_{udud} + 1m)$-pentamer	−578.879 86	0.269 28	14.10	116	23
5	$(4r_{udud} + 1m_{O}\cdot\cdot HC)$-pentamer	−578.879 31	0.269 20	15.54	114	23
6	$(4r_{uddd} + 1m_{u})$-pentamer	−578.877 50	0.268 45	20.30	112	22
7	$(4r_{uudd} + 1m)$-pentamer	−578.877 44	0.268 55	20.46	111	22
8	$(4r_{uddd} + 1m_{d2})$-pentamer	−578.877 38	0.268 27	20.60	112	22
9	$(4r_{uddd} + 1m_{d3})$-pentamer	−578.876 86	0.268 55	21.97	110	22
10	$(4r_{uuuu} + 1m)$-pentamer	−578.874 72	0.268 40	27.58	105	21
11	$(3r_{udu} + 1d)$-pentamer	−578.870 75	0.268 78	38.00	93	19
12	5c-pentamer (loose)	−578.870 29	0.267 81	39.22	94	19
13	5c-pentamer (tight)	−578.869 58	0.267 90	41.08	92	18
14	$(3r_{udu} + 2m)$-pentamer	−578.865 73	0.268 01	51.19	82	16

[a] ΔE conf = $(E + \text{ZPVE})_{\text{conformer}} - (\text{E} + \text{ZPVE})_{\text{minimum\_energy\_structure}}$ in kJ/mol.

(3) Both Figures 2 and 3 indicate that the maximum dissociation energy per mole of methanol for a cluster of *n* methanol molecules is obtained when the clusters contain all *n* molecules in a single ring. This is not surprising, because the *n*-ring clusters have the greatest opportunity for H bonding, with each methanol molecule acting as both donor and acceptor, hence, *n* donors and *n* acceptors. For example, consider the data in Table 3, which lists the D_0 for various pentamer clusters. Three different 5r-pentamer structures with zero imaginary frequencies were located; the "planar" one, in which the hydroxyl groups are approximately in a plane with the methyl groups alternating up and down, *ududu*, is the lowest-energy structure found. The three 5r-pentamer structures have D_0 in the vicinity of 126−132 kJ/mol of pentamer (25−26 kJ/mol of methanol). With smaller ring structures, the dissociation energies are progressively lower. The (4r + 1m)-pentamer structures have D_0 ranging from 105 to 116 kJ/mol of pentamer (21−23 kJ/mol of methanol); in this case, there are *n* donors but only $n - 1$ acceptor H bonds. The (3r + 1d)-pentamer has D_0 equal to 93 kJ/mol of pentamer (19 kJ/mol of methanol), while the (3r + 2m)-pentamer structure has a D_0 of 82 kJ/mol of pentamer (16 kJ/mol of methanol). The pentamer chain has D_0 equal to 92−94 kJ/mol of pentamer (18−19 kJ/mol of methanol). Similar trends were noted for all other clusters.

4. The data illustrated in Figure 2, which are based on B3LYP/-311+G(d,p) single-point energies, show that the dissociation energies per mole of methanol for the *n*mers converge to a value of about 27 kJ/mol of methanol. [In Figure 3, the energies converge to about 40 kJ/mol for the B3LYP/6-31G(d) energies. A major effect of the larger basis set is to lower the calculated dissociation energies, as is also evident from the basis set study shown in Table 2.]

5. The "loose" chain structures are more stable than trimer ring structures, which have strain. For example, for the pentamers, the loose 5c-pentamer has a higher D_0 than both the (3r + 2m)-pentamer and the (3r + 1d)-pentamer.

6. As *n* increases, the difference in D_0 per mole of methanol between the lowest-energy ring structure for a given *n* and the lowest-energy chain structure initially increases to a maximum at $n = 6$ then decreases steadily, as illustrated in Figure 4. Thus, the 6r-hexamer is substantially more stable than the 6c-hexamer, but beyond that, the D_0 per mole of methanol for the chain structures appears to be approaching those of the ring structures.

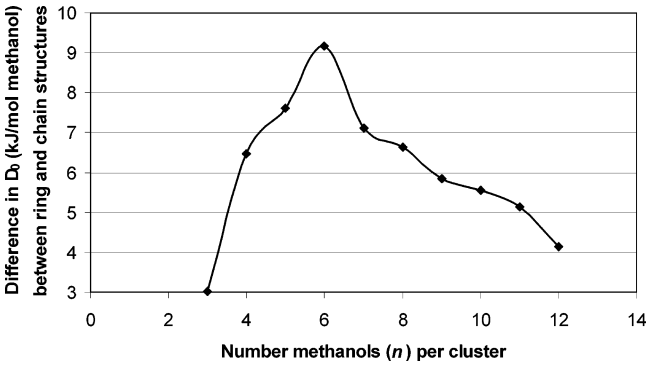


Figure 4. Difference in dissociation energies in kJ/mol of methanol [B3LYP/6-311+G(d,p)//B3LYP/6-31G(d)] between ring structures as a function of the number of methanols in the cluster.

D. Vibrational Frequency Shifts. Figure 5 is a plot of the average O−H vibrational frequency as a function of the number of methanols in the ring structures. (Average values were used, because the *n*mers have *n* O−H stretching absorptions.) It is well-known that gaseous methanol has a higher vibrational frequency for the O−H stretch than liquid methanol.[47] Clearly, even with the small basis set used in this study [6-31G(d)], the O−H vibrational frequency drops as the number of methanols in the cluster increases. The minimum frequency is essentially attained when there are five methanols in the ring, which provides further evidence to support the suggestion that five or six hydrogen-bonded methanols are sufficient to mimic liquid behavior.

Another observation from the O−H stretch frequency data is that stretches associated with methanols that are outside the ring have higher frequencies than those within the ring.

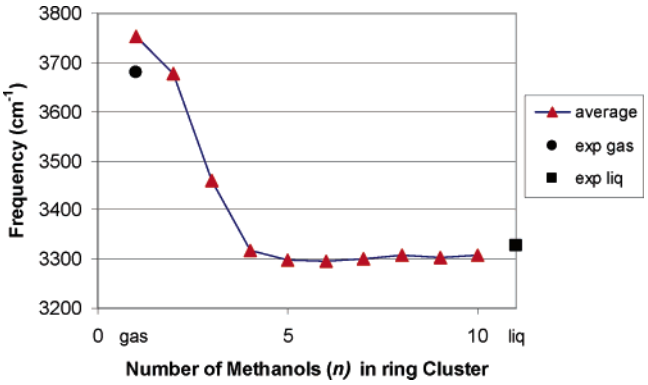


Figure 5. Average B3LYP/6-31G(d) O−H vibrational frequencies (cm^{-1}) for the methanol ring clusters as a function of *n*, compared to the experimental values in the gas and liquid[46] phases.

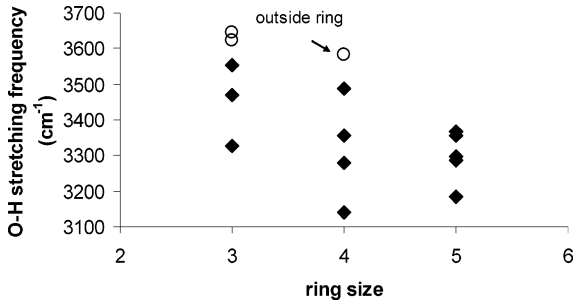


Figure 6. O−H stretching frequencies for three of the pentamer clusters as a function of the number of methanols within the ring. The structures are the lowest-energy ones in each of the respective groups, that is, the ($3r_{udu}$ + 2m)-pentamer, ($4r_{udud}$ + 1m)-pentamer, and $5r_{ududu}$-pentamer species.

Figure 6 shows the O−H stretching frequencies as a function of the number of methanol molecules within a ring for three pentamer clusters (the $5r_{ududu}$-pentamer, the ($4r_{udud}$ + 1m)-pentamer, and the ($3r_{udu}$ + 2m)-pentamer). In each case, the stretch associated with the methanol(s) with only donor H bonds has a significantly higher stretching frequency.

Conclusions

The potential energy surfaces of methanol clusters are very flat. Our computational study of a series of $(CH_3OH)_n$ clusters, where $n = 2-12$ methanol molecules, indicates that there are increasing numbers of minimum energy structures (no imaginary vibrational frequencies) as *n* increases. Furthermore, our study suggests that the dissociation energy per mole of methanol increases as *n* increases, up to about five to six methanols. Beyond that, the D_0/mole of methanol remains approximately constant at about 27 kJ/mol for B3LYP/6-311+G(d,p)//B3LYP/6-31G(d) energies, though this value is likely to be lower at a higher level of theory. Ring clusters of *n* molecules are more stable than branched-ring and chain ones; however, the difference in D_0/mole of methanol between ring and chain structures decreases as the value of *n* increases beyond six methanol molecules. On the basis of the results shown in Figures 2 and 3, it would appear that pentamer or hexamer structures approximate the bulk liquid behavior. This is supported by the results of O−H vibrational energies (Figure 5). It would be interesting to know if a significantly higher level of theory would yield the same value, that is, that five to six methanol molecules in a cluster is sufficient to mimic liquid behavior.

Acknowledgment. We gratefully acknowledge the hospitality of Professor Jesus M. Ugalde and his colleagues at The University of the Basque Country where much of this research was completed. We also thank Professors Otilia Mó and Manuel Yáñez of the Universidad Autónoma de Madrid for helpful discussions. The financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Killam Trusts is gratefully acknowledged.

Supporting Information Available: Archive entries for all geometry optimizations of the structures discussed in this paper are available free of charge via the Internet at http://pubs.acs.org.

References

(1) Torrie, B. H.; Weng, S.-X.; Powell, B. M. *Mol. Phys.* **1989**, *67*, 575.

(2) Tauer, K. J.; Lipscomb, W. N. *Acta Crystallogr.* **1952**, *5*, 606.

(3) Ladanyi, B. M.; Skaf, M. S. *Annu. Rev. Phys. Chem.* **1993**, *44*, 335.

(4) Tsuchida, E.; Kanada, Y.; Tsukada, M. *Chem. Phys. Lett.* **1999**, *311*, 236.

(5) Pauling, L. *The Nature of the Chemical Bond*; Cornell University Press: Ithaca, NY, 1960.

(6) Sarkar, S.; Joarder, R. N. *J. Chem. Phys.* **1993**, *99*, 2032.

(7) Montague, D. G.; Gibson, I. P.; Dore, J. C. *Mol. Phys.* **1981**, *44*, 1355.

(8) Tanaka, Y.; Ohtomo, N.; Arakawa, K. *Bull. Chem. Soc. Jpn.* **1984**, *57*, 644.

(9) Narten, A. H.; Habenschuss, A. *J. Chem. Phys.* **1984**, *80*, 3387.

(10) Magini, M.; Paschina, G.; Piccaluga, G. *J. Chem. Phys.* **1982**, *77*, 2051.

(11) Kashtanov, S.; Augustson, A.; Rubensson, J.-E.; Nordgren, J.; Ågren, H.; Guo, J.-H.; Luo, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2005**, *71*, 104205.

(12) Wilson, K. R.; Cavalleri, M.; Rude, B. S.; Schaller, R. D.; Catalano, T.; Nilsson, A.; Saykally, R. J.; Pettersson, L. G. M. *J. Phys. Chem. B* **2005**, *109*, 10194.

(13) Bertolini, D.; Cassettari, M.; Salvetti, G. *J. Chem. Phys.* **1983**, *78*, 365.

(14) Guillot, B.; Marteau, P.; Obriot, J. *J. Chem. Phys.* **1990**, *93*, 6148.

(15) Schulman, E. M.; Dwyer, D. W.; Doetschman, D. C. *J. Phys. Chem.* **1990**, *94*, 7308.

(16) Bertie, J. E.; Zhang, S. L.; Eysel, H. H.; Baluja, S.; Ahmed, M. K. *Appl. Spectrosc.* **1993**, *47*, 1100.

(17) Bertie, J. E.; Zhang, S. L. *J. Mol. Struct.* **1997**, *413*, 333.

(18) Morineau, D.; Guégan, R.; Xia, Y.; Alba-Simionesco, C. *J. Chem. Phys.* **2004**, *121*, 1466.

(19) Takamuko, T.; Maruyama, H.; Kittaka, S.; Takahara, S.; Yamaguchi, T. *J. Phys. Chem.* **2005**, *109*, 892.

(20) Jorgensen, W. L. *J. Am. Chem. Soc.* **1981**, *103*, 341.

(21) Haughney, M.; Ferrario, M.; McDonald, I. R. *J. Phys. Chem.* **1987**, *91*, 4934.

(22) Handgraaf, J.-W.; van Erp, T. S.; Meijer, E. J. *Chem. Phys. Lett.* **2003**, *367*, 617.

(23) Morrone, J. A.; Tuckerman, M. E. *J. Chem. Phys.* **2002**, *117*, 4403.

(24) Pagliai, M.; Cardini, G.; Righini, R.; Schettino, V. *J. Chem. Phys.* **2003**, *119*, 6655.

(25) Handgraaf, J.-W.; Meijer, E. J.; Gaigeot, M.-P. *J. Chem. Phys.* **2004**, *121*, 10111.

(26) Wojcik, M. J.; Hermansson, K.; Lindgren, J.; Ojamäe, L. *Chem. Phys.* **1993**, *171*, 189.

(27) Wang, J.; Boyd, R. J.; Laaksonen, A. *J. Chem. Phys.* **1996**, *104*, 7261.

(28) Jorgensen, W. L. *J. Phys. Chem.* **1986**, *90*, 1276.

(29) Pettitt, B. M.; Rossky, P. J. *J. Chem. Phys.* **1983**, *78*, 7296.

(30) Matsumoto, M.; Gubbins, K. E. *J. Chem. Phys.* **1990**, *93*, 1981.

(31) Svishchev, I. M.; Kusalik, P. G. *J. Chem. Phys.* **1994**, *100*, 5165.

(32) Palinkas, G.; Bako, I.; Heinzinger, K.; Bopp, P. *Mol. Phys.* **1991**, *73*, 897.

(33) Wick, C. D.; Dang, L. X. *J. Chem. Phys.* **2005**, *123*, 184503.

(34) Martin, M. E.; Sánchez, M. L.; Olivares del Valle, F. J.; Aguilar, M. A. *J. Chem. Phys.* **2002**, *116*, 1613.

(35) Morrone, J. A.; Tuckerman, M. E. *Chem. Phys. Lett.* **2003**, *370*, 406.

(36) Mó, O.; Yáñez, M.; Elguero, J. *J. Chem. Phys.* **1997**, *107*, 3592.

(37) Mandado, M.; Grañña, A. M.; Mosquera, R. A. *Chem. Phys. Lett.* **2003**, *381*, 22.

(38) Vener, M. V.; Sauer, J. *J. Chem. Phys.* **2001**, *114*, 2623.

(39) Ludwig, R. *Chem. Phys. Chem.* **2005**, *6*, 1369.

(40) El-Shall, M. S.; Wright, D.; Ibrahim, Y.; Mahmoud, H. *J. Phys. Chem. A* **2003**, *107*, 5933.

(41) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision B.05; Gaussian, Inc.: Pittsburgh, PA, 2003.

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

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

(44) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11263.

(45) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev.* **1987**, *37*, 785.

(46) Rankin, K. N.; Boyd, R. J. *J. Comput. Chem.* **2001**, *22*, 1590.

(47) Shimanouchi, T. *Tables of Molecular Vibrational Frequencies*; NSRDS−NBS 39; U.S. Department of Commerce: Washington, DC, 1972; Vol. 1.

CT6002912

JCTC Journal of Chemical Theory and Computation

The Optical Rotation of Glucose Prototypes: A Local or a Global Property?

Clarissa O. da Silva* and Benedetta Mennucci

Departamento de Química, UFRuralRJ, BR456 km 47, Rio de Janeiro, Brazil, and Dipartimento di Chimica e Chim. Ind., Università degli Studi di Pisa, via Risorgimento 35, Pisa, Italy

Received August 3, 2006

Abstract: In this work, we present a quantum-mechanical study of the optical rotation (OR) of model systems representing glucose prototypes with one and two chiral centers. The ONIOM method is used to evaluate the property and to analyze its local or global character. Different ONIOM partitions are tested and compared to better appreciate differences and similarities between mono- and bichiral prototypes. The local versus global character of OR is investigated and compared to other properties such as energies and nuclear magnetic shieldings, which have been deeply studied in previous applications of the ONIOM method.

1. Introduction

Optical rotation (OR) plays a significant role in carbohydrate research. Both the high number of chiral centers present in these compounds and the relative ease of experimentally obtaining the OR values have been important factors. While in the sugar industry the OR facilitates an efficient way to measure concentrations, the scientific community has especially employed the relationship between OR and the compound's structure.[1−10]

However, the development of theoretical [and especially quantum-mechanical (QM)] methods for the prediction of OR, from which carbohydrate research could greatly benefit, is quite difficult for different reasons. First of all, we have to consider the difficulty involved in the QM calculation of the optical rotation. Only in recent years have advances in the field of theoretical chemistry led to the development of new computational approaches for calculating OR at different levels of accuracy, including methods such as the Hartree−Fock (HF),[11] density functional theory,[12] and coupled cluster.[13] This development of the theoretical background has been followed by a large number of papers using theoretical approaches,[14−19] possibly including the effect of the solvent,[20,21] to elucidate the factors that determine OR.

Generally, these studies have been limited to chiral systems with a single chiral center. Saccharides, on the contrary, are systems with many chiral centers, and therefore, the complexity of the calculations is largely amplified. Additionally, such chiral centers can occur in different geometrical arrangements because saccharides present many conformations: all of these aspects make the calculation of OR for saccharides very difficult but, at the same time, a challenging test for QM approaches. Recently, we have presented a study on OR of glucose in aqueous solution showing that, indeed, QM methods can be reliably applied to the study of these kinds of systems (even including the effect of the solvent).[22] Obviously, the extension to other saccharides (and especially oligosaccharides) is not straightforward. While a reliable conformational analysis is still possible using either less accurate QM approaches[23] or molecular mechanics,[24] the calculation of OR necessarily requires an extended QM description, and thus it becomes computationally quite demanding.

One possible approach to make accurate calculations on such kinds of systems feasible is the use of hybrid or layered methods. These methods partition the system into regions, each of them treated with a different level of accuracy. The justification for this approach is that the various parts of the system contribute to the property we are interested in, in different ways, and therefore they require different computational levels. A powerful version of this kind of approach is the ONIOM (our own n-layered integrated molecular orbital) method.[25] Over the years, ONIOM has been used to predict energies and geometries of a large number of

* Corresponding author fax: +55-21-26822807; e-mail: clarissa-dq@ufrrj.br.

10.1021/ct600250w CCC: $37.00
Published on Web 11/15/2006

chemical systems,[26] also involving solvent effects.[27] In addition, it has been successfully applied in the prediction of various molecular properties such as nuclear magnetic resonance (NMR) shieldings.[28]

These previous studies suggest that ONIOM works well if the properties to be calculated are principally determined locally, that is, by a small part of a large system, with some corrections from the rest of the system. However, when the property is global, that is, the whole system largely contributes to its final value, ONIOM can still be used but in a different way. In these cases, only relative values can be reliably obtained, such as, for example, energy differences among different conformers or activation free energies.

Motivated by these results, an ONIOM procedure has been tried here to investigate two different but related aspects of optical rotation, one concerning its nature and the other its calculation. In particular, for the first time, the character of OR as a global or a local property is analyzed, and ONIOM is tested as a potential method to simulate OR in saccharide-like systems. If this would be the case in fact, it could represent an extremely useful tool to validate the structures selected by sampling potential energy surfaces.

The general ONIOM strategy, however, requires reconsideration in some aspects when applied to the calculation of OR in saccharide-like systems. In fact, because of the presence of multiple chiral centers, the preliminary identification of the part of the system which most contributes to the property required by any layered approach is not only not univocal but also questionable. In the following sections, the optical rotation of model systems representing glucose prototypes with one and two chiral centers are thus introduced, and different ONIOM partitions are tested and compared to better appreciate differences and similarities between the two classes of prototypes.

The paper is organized as follows: In section 2, we define the monochiral and bichiral prototypes studied and the ONIOM partition used, while in section 3, the calibration of the ONIOM method is performed in order to define the proper high and low levels of calculations. In section 4, the chosen ONIOM partitions and levels are applied to the calculation of the OR of the mono- and bichiral systems and the results compared with the high-level benchmarks. In section 5, two further ONIOM calculations of energy differences about the α and β anomers and NMR nuclear shieldings are presented and compared with the previous ones on OR to rationalize the results in terms of global and local character of the property. Finally, in section 6, some conclusive comments are reported.

2. The Definition of the Chiral Systems

In this study, we introduce simplified monosaccharide model systems, derived from glucose but with just one or two adjacent chiral centers; in the following, we shall refer to these systems as prototypes. By studying these simplified systems in which we have eliminated any conformational freedom (each prototype is considered as a rigid system) and reduced the number of chiral centers, we can simplify the analysis but still keep the same main structural and electronic aspects of the real systems.

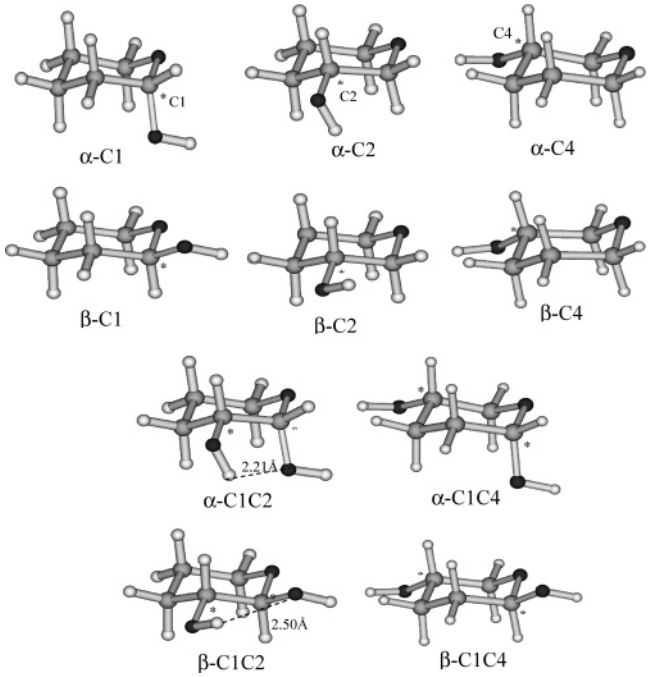


Figure 1. Glucose anomeric prototypes used in this work. The chiral centers are indicated by an asterisk. The hydrogen-bond distances are reported for the α and β C1C2 bicenter chiral prototypes.

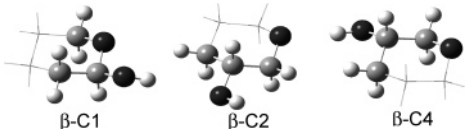


Figure 2. Scheme of the ONIOM partitions adopted to describe the OR in systems with just one chiral center.

Two sets of prototypes have been built starting from the two most abundant conformers of glucose in aqueous solution, one for the α anomer and one for the β anomer. These structures correspond to the GT conformation of the hydroxyl methyl group of each anomer of glucose. The geometries of the two conformers have been presented in a previous work on the OR.[22]

From both structures, three monocenter and two bicenter chiral systems are obtained by replacing the hydroxymethyl and all of the hydroxyl groups, except one (monocenter) or two (bicenter), with hydrogen atoms. The structures of the different prototypes we have thus obtained are reported in Figure 1. We note that among all possible structures which can be obtained from glucose we have selected those that keep the chiral center on the anomeric carbon (indicated as C1) and/or on its vicinal carbon (C2) and/or on its opposite carbon (C4). This choice has been dictated by the fact that, here, we are interested in analyzing the local character of the OR and thus starting from the main chiral center (the anomeric carbon); C2 and C4 represent the two extremes (the closest and farthest).

In Figures 2 and 3, the ONIOM two-layer partition adopted in this work is presented, but for simplicity's sake, only for the β anomers. Figure 2 shows the ONIOM partitions adopted for the monochiral prototypes, while Figure 3 shows those for the bichiral ones. The ball-and-stick representation

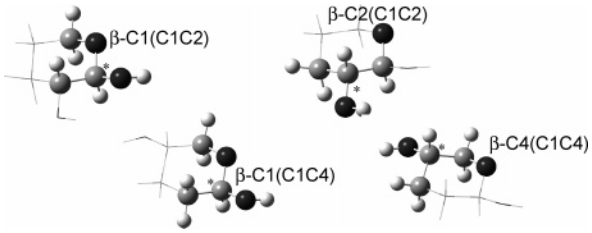


Figure 3. Scheme of the ONIOM partitions adopted to describe the OR in systems with two different chiral centers. In all systems, only one chiral center (*) is present in the HM layer.

regards the part of the system which is described at the high level and is usually indicated as the high-model (HM) layer in the ONIOM framework. All the link atoms, that is, the atoms that are bounded to the model to complete their valence when a bond is broken, are hydrogen atoms.

As can be seen from Figure 2, the HM layer has been chosen to include the proper chiral center of the monocenter chiral prototypes of Figure 1. We note that, for the C4 prototype, two different HMs are possible, one including the two carbons on the left of the chiral C4 and the other including one carbon and the oxygen on the right of C4; in the following analysis, we shall consider only the latter one because only this preserves the ring oxygen atom in the HM layer as the other C1 and C2 HM layers. The ring oxygen is in fact of fundamental importance in defining the nature of the sugarlike systems because it belongs to the moiety responsible for the anomeric effect.

The four structures of Figure 3 correspond to ONIOM calculations for the bichiral systems reported in Figure 1. They refer to ONIOM calculations where just one of the two chiral centers (that indicated in Figure 3 by an asterisk) present in the whole system is preserved in the HM layer. As already noted for C4, for C2, we can also define in principle two different HMs. Once again, we only consider that preserving the ring oxygen atom. We also note that both bichiral systems C1C2 and C1C4 present the same C1 HM layer; thus, in the Figure 3 and in the following analysis, we will use two labels, X−C1(C1C2) and X−C1(C1C4), to indicate such a layer in C1C2 and C1C4 systems, respectively.

3. Calibrating the Model

3.1. The QM Calculation of OR. The quantum-mechanical foundations of optical rotation are due to Rosenfeld,[29] who demonstrated that the electric dipole moment of a chiral molecule induced by a frequency-dependent electromagnetic field may be written as

$$\mu_a = \sum_b \left(\alpha_{ab} E_b - \beta_{ab} \frac{\partial B_b}{c \partial t} \right) \tag{1}$$

where E and B represent the applied, time-dependent electric and magnetic field vectors and the α tensor denotes the usual molecular dipole polarizability. The key quantity introduced by Rosenfeld is the electric dipole-magnetic dipole polarizability β; it is in fact its isotropic value which determines the OR of a chiral molecule.

From eq 1, it follows that, as the dipole μ and the polarizability α of the molecule can be properly defined in an ONIOM scheme,[22d] it is also possible to define an ONIOM β as

$$\beta^{\mathrm{ONIOM}} = \beta^{\mathrm{LR}} + \beta^{\mathrm{HM}} - \beta^{\mathrm{LM}} \tag{2}$$

where a two-layer ONIOM has been used.

From a computational point of view, the calculation of each $\beta^{\mathrm{level,system}}$ in eq 2 can be obtained, at least for variational wave functions in the limit of a static field, in terms of electric and magnetic field derivatives of the ground-state electronic wave function,[30] namely

$$\beta_{ab} = \frac{hc}{3\pi} \mathrm{Im} \left\langle \frac{\partial \Psi}{\partial E_a} \middle| \frac{\partial \Psi}{\partial B_b} \right\rangle \tag{3}$$

The derivatives of the ground-state electronic wave function, $\partial\Psi/\partial E$ and $\partial\Psi/\partial B$, are calculated using analytical derivative methods and field-dependent atomic orbitals, in particular, gauge-invariant (or including) atomic orbitals (also referred to as London orbitals).[31] This procedure has been generalized to the frequency-dependent case[32] by extending the analytical derivative method to its time-dependent analogue.[12,13]

As a matter of fact, in the following, the optical rotation will be analyzed in terms of the specific rotation $[\alpha]_{\mathrm{D}}$ instead of the isotropic electric dipole-magnetic dipole polarizability β; their simple relation is given by

$$[\alpha]_\nu = 13.43 \times 10^{-5} \frac{\beta \bar{\nu}^2}{M} \tag{4}$$

with β in atomic units, the molar mass M in grams per mole, and $\bar{\nu}$ (the wavenumber where optical rotation is measured) in centimeters^{-1}.

3.2. The Benchmarks and the Choice of the High Level. Taking into account the studies performed by Stephens et al.[12] where the effects of considering different levels of QM descriptions were deeply investigated, the first reasonable choice to calculate the optical rotation for the glucose prototypes studied in this paper is to adopt the B3LYP density functional approach with the aug-cc-pVDZ basis set. An important characteristic of the correlated consisted (cc) group of basis is the ability to generate a sequence of basis sets which converges toward the basis set limit. However, small basis sets (that are needed in the low-level ONIOM layer) belonging to this family of bases are not available, and thus the use of ONIOM becomes useless if two time-demanding calculations must be done in both layers.

Therefore, it is mandatory to introduce, for the low level, another kind of basis set not belonging to the cc groups of basis. For this reason, it is interesting to test an alternative to the aug-cc-pVDZ for the high level, which is more coherent with the type of the basis set to be adopted for the low level. Still following the analysis by Stephens et al., we see that a valid alternative is represented by the split valence 6-311++G(2d,2p) basis set.

This basis sets, however, is quite large, and thus, in the view of future applications of the method to real mono- and disaccharides, we have tried to find a substitute which could correctly reproduce the behavior of the larger basis set but

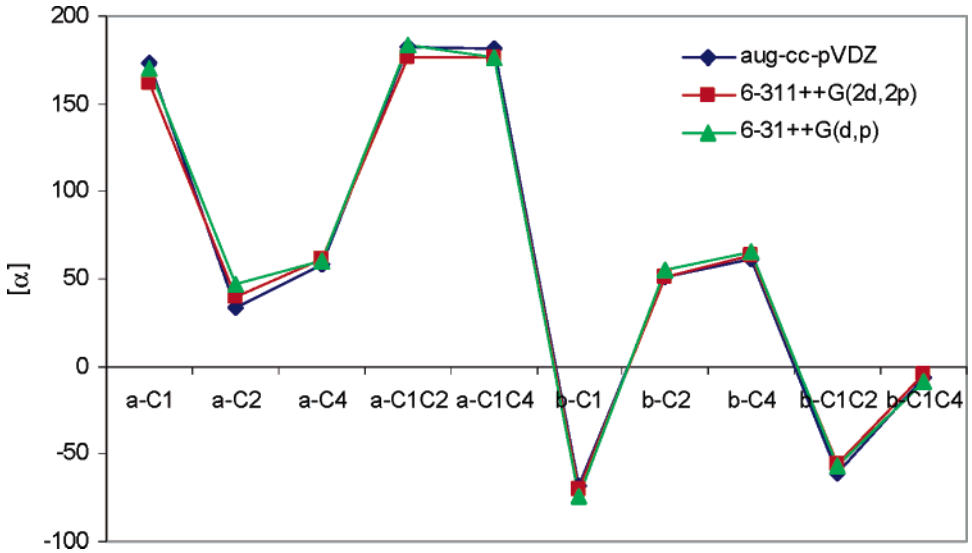


Figure 4. $[\alpha]_D^{20}$ values {in degrees/[dm(g/cm^3)]} for 10 prototypes of glucose, at the B3LYP level, in the gas phase obtained with three different basis sets.

at a much lower computational cost. The selected basis set is the 6-31++G(d,p). We note that also in this second set we have maintained the presence of diffuse functions on both heavy and hydrogen atoms, as required to correctly describe the OR property.

The results obtained for the 10 prototypes with these three basis sets are reported in Figure 4. A local version of Gaussian 03 computational code[33] has been used for all the calculations.

Looking at data reported in Figure 4, it can be seen that the three basis sets give very similar OR values, with a single exception for α-C2, for which the 6-31++G(d,p) gives an OR value 30% higher than aug-cc-pVDZ.

Following this analysis, we have fixed the high-layer level of the next ONIOM calculations to B3LYP/6-31++G(d,p).

3.3. *S* Test for Monochiral Systems: The Choice of the Low Level. A systematic procedure to test the reliability of the level/model combination in ONIOM has been devised using the monochiral systems. The descriptions tested are HF/3-21G, HF/3-21+G, and B3LYP/3-21+G. All of the calculations performed did not involve any reoptimization of the geometries.

In the present application of the ONIOM method, $[\alpha]_D^{20}$(high,real) is the target (or benchmark) calculation that $[\alpha]_D^{20}$(ONIOM) is trying to reach. Let us define the *S* value at a given level as

$$S(\text{level}) = [\alpha]_D^{20}(\text{level,real}) - [\alpha]_D^{20}(\text{level,model}) \quad (5)$$

The *S* value, *S*(level), is the difference between the *real* and the *model* system, or the effect of the substituent (second layer) evaluated at a given level. It is obvious that if *S*(low) → *S*(high), then $[\alpha]_D^{20}$(ONIOM) → $[\alpha]_D^{20}$(high,real). Namely, if the *S* value evaluated at the low level is the same as that at the high level, the ONIOM property is the same as the target property; that is, the error of the ONIOM extrapolation, $[\alpha]_D^{20}$(ONIOM) − $[\alpha]_D^{20}$(high,real), is zero.

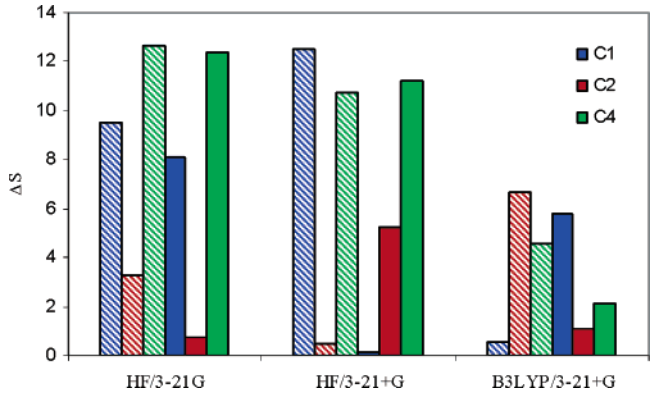


Figure 5. Δ*S* values for ONIOM(B3LYP/6-31++G(d,p):X) $[\alpha]_D^{20}$ {in degrees/[dm(g/cm^3)]} for the monocenter prototypes. Striped and full columns refer to α and *β* anomers, respectively.

Taking into account these considerations, we prefer to visualize the results in terms of the Δ*S* = *S*(low) − *S*(high) values, namely

$$\Delta S = S(\text{low}) - S(\text{high}) = (\text{LR} - \text{LM}) - (\text{HR} - \text{HM})$$
$$= \text{LR} - \text{LM} - \text{HR} + \text{HM} = (\text{LR} + \text{HM} - \text{LM}) - \text{HR} \quad (6)$$

where the quantity between the parentheses is exactly the integrated ONIOM property. In other words, the Δ*S* value is a straightforward and quantitative measurement of how much each OR ONIOM value is close to the respective benchmark. These quantities are reported in the graphs in Figures 5.

From Figure 5, it appears that the effects of the description of the low level (both the method and the basis set) are dependent on the conformation.[34] For example, the Δ*S* values when passing from HF/3-21G to HF/3-21+G increase for αC1 and decrease for *β*C1, while the opposite trend is found passing from HF/3-21+G to B3LYP/3-21+G. On the average, however, the B3LYP/6-31++G(d,p):B3LYP/3-21+G ONIOM combination significantly differs from the others giving small Δ*S* values for all prototypes.

Table 1. B3LYP/6-31++G(d,p) $[\alpha]_D^{20}$ Values {in degrees/[dm(g/cm^3)]} for All Prototypes

prototype	OR	prototype	OR
	Monochiral		
α-C1	170.53	*β*-C1	−74.02
α-C2	46.87	*β*-C2	55.53
α-C4	59.86	*β*-C4	64.9
	Bichiral		
α-C1C2	183.09	*β*-C1C2	−56.44
α-C1C4	176.47	*β*-C1C4	−8.66

Table 2. ONIOM(B3LYP/6-31++G(d,p):B3LYP/3-21+G) $[\alpha]_D^{20}$ Values {in degrees/[dm(g/cm^3)]} for the Different ONIOM Partitions Reported in Figure 2

	OR (degrees/[dm(g/cm^3)])			
	LR	HM	LM	ONIOM
α-C1	149.86	154.62	134.49	169.98
α-C2	45.13	17.61	9.23	53.50
α-C4	69.11	39.70	53.49	55.32
β-C1	−72.19	−86.24	−78.66	−79.77
β-C2	79.65	24.60	47.60	56.65
β-C4	74.11	47.56	54.67	67.00

It can thus be assumed that such a combination is the most suited for an ONIOM description of the present systems, and from now on, it will be the only one used.

4. The ONIOM Analysis

In this section, we report the analysis of the results in three sections, one referring to the high-level (or benchmark) calculations and two to the ONIOM calculations for OR of mono- and bichiral prototypes, respectively.

4.1. OR Benchmark. All of the OR ($[\alpha]_D^{20}$) benchmark values obtained for all of the prototypes are reported in Table 1.

A first aspect to be underlined is the different behavior of OR in α and *β* prototypes. In the α series, the OR values are all positive, with α-C2 and α-C4 presenting smaller values. On the contrary, in the *β* series, both positive and negative values are found. We note, however, that despite this different behavior, in both series, the monochiral C1 system presents OR values very close to those of the bichiral system C1C2, principally for the α conformers. One possible interpretation of this fact is that the chiral centers do not equally contribute to the net chiral behavior, but instead the C1 center plays the major role, because, even when the chirality of C2 is eliminated by saturation, the monochiral C1 prototype presents an OR value very close to that presented by the C1C2 bichiral system. Complimentarily, when the chirality of C1 is eliminated, the results obtained for the OR of the monochiral C2 prototype are not comparable to the corresponding bichiral values.

This analysis can be applied also to C1C4 bichiral. In this case, the chiral centers are away from each other, and thus we can expect that the elimination of one chiral center (to build the corresponding monochiral prototype) should not largely affect the other chiral center. In this case, we should more easily find if one of the two centers is dominant or if, instead, both of them similarly contribute to the OR. For α conformers, it appears that C1 plays the dominant role. In fact, the C1 monochiral prototype practically has the same OR value as the bichiral C1C4 system, in both phases. By contrast, in the *β* conformer, the chirality of both centers seems to be additive, because the bichiral C1C4 OR value is approximately reproduced by the sum of the values found for each monochiral prototype.

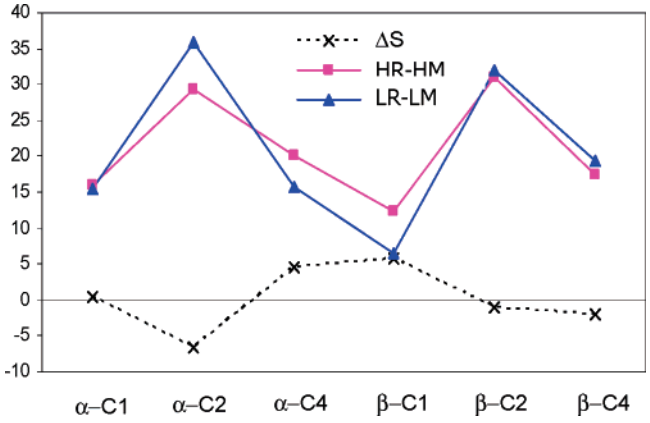


Figure 6. Differences (HR − HM) and (LR − LM) and Δ*S* values {in degrees/[dm(g/cm^3)]} for the different ONIOM partitions of the monochiral systems.

It is important to recall that the α and *β* C4 monochiral systems have very similar OR values because the anomerism has been eliminated. The only differences are due to small geometric changes.

4.2. The Monochiral Prototypes. In Table 2, we report the ONIOM OR values for the four different monochiral systems C1, C2, and C4 (see Figure 2).

From the comparison of the results reported in Tables 1 and 2, it appears that the ONIOM method can be reliably used to compute the OR property for monochiral systems once the high/low combination is accurately defined on the basis of the *S* test (see section 2.2). For all systems, the difference between ONIOM and benchmark values is always below 6°/[dm(g/cm^3)].

These results seem to suggest that OR behaves as a local property for these monochiral systems: one could observe that this is indeed a straightforward result as only one chiral center is present. This conclusion, however, is not correct as locality is not automatically determined by monochirality. Also, in the presence of a single chiral center, in fact, the OR can significantly depend on the structural and electronic properties of the whole molecule.

To verify if the good results obtained with ONIOM really imply a local character of the property or, instead, are due to a fortuitous cancellation of errors, we introduce here a further analysis.

In Figure 6, we report the Δ*S* value defined in eq 6 together with the differences between the real and the model systems (both at a low and at a high level of calculations).

The local character of OR is directly evaluated in terms of the difference between HR and HM descriptions. If the part of the system introduced in the model layer contains the whole effect responsible for the property, such a

Table 3. ONIOM(B3LYP/6-31++G(d,p):B3LYP/3-21+G) $[\alpha]_D^{20}$ and ΔS Values {in degrees/[dm(g/cm^3)]} for the Different Partitions of the Bichiral System (see Figure 3)

	LR	HM	LM	ONIOM	ΔS
α-C1(C1C2)	164.96	133.68	116.28	182.36	0.73
α-C2(C1C2)	164.96	15.22	7.98	172.20	10.89
α-C1(C1C4)	179.03	133.68	116.28	196.43	−19.96
α-C4(C1C4)	179.03	34.32	46.25	167.11	9.36
β-C1(C1C2)	−51.31	−74.56	−68.01	−57.86	1.42
β-C2(C1C2)	−51.31	21.27	41.15	−71.20	−14.76
β-C1(C1C4)	−4.84	−74.56	−68.01	−11.39	−2.73
β-C4(C1C4)	−4.84	41.12	47.27	−10.99	2.33

difference must be very small. From Figure 6, it can be seen that the "local" character decreases in the order C1, C4, and C2. We also note that the same ordering in "degree of locality" found from the HR − HM difference is obtained also in terms of LR − LM differences, generally a more accessible quantity for large systems when the RH calculations are prohibitive. However, the reliability of an ONIOM description depends not only on this degree of locality. As it can be seen from Figure 6, larger ΔS values (i.e., worse ONIOM descriptions) are obtained for α-C4 and β-C1, for which the HR − HM difference is relatively small.

Following this analysis, it seems that the performance of ONIOM to describe the OR of a chiral center depends on two different aspects, the "true locality" of the property for such a given chiral center and the different evaluation of the locality at low and high levels, or in other words the differences between HR − HM and LR − LM.

4.3. The Bichiral Prototypes. In Table 3, we report the ONIOM results for the bichiral systems C1C2 and C1C4 (see Figure 3), for the three ONIOM partitions considered.

In the previous section, we have shown that the ONIOM description is sufficiently accurate for monochiral systems and that the ΔS error found is always acceptable. On the contrary, for bichiral systems, the error associated with the ONIOM calculations can be large, as shown in Table 3, and it ranges from 2° to 20°/[dm(g/cm^3)], depending on the system.

This result is not completely unexpected as the bichiral systems introduce a new aspect, which was not present in the monochiral systems. This aspect is related to the elimination of one of the two chiral centers in the HM layer and the consequences of this on the resulting OR value. Because of the limited dimensions of the prototypes studied, it is not possible to define an ONIOM model system including both chiral centers. However, we can try to approximately account for them by summing up the ONIOM results of the two complementary model systems [e.g., α-C1-(C1C2) and α-C2(C1C2)], namely:

$$\begin{aligned}\text{ONIOMbi} &= \text{ONIOMCx(CxCy)} + \text{ONIOMCy(CxCy)} - \text{LR} \\ &= (\text{LR} + \text{HMx} - \text{LMx}) + (\text{LR} + \text{HMy} - \text{LMy}) - \text{LR}\end{aligned} \quad (7)$$

where the subtraction of the value found for the low real system has been introduced to avoid its double counting. The ΔS values corresponding to this approximate description

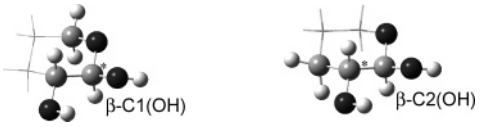


Figure 7. Scheme of the additional ONIOM partitions adopted to describe the OR in the C1C2 system.

can be easily obtained from Table 3; the results obtained are −6.5, −8.0, 21.3, and 8.9 {in degrees/[dm(g/cm^3)]} for α-C1C2, α-C1C4, β-C1C2, and β-C1C4, respectively. These values further confirm the nonlocality of the property because locality would also imply a limited coupling between the two chiral centers, and thus eq 7 (in which two separated ONIOM calculations, each one with a different model system, are combined) should reliably reproduce the benchmark.

To better analyze the real character of OR in these bichiral systems, we introduce now a different test based on the change of the ONIOM description as the model layer increases. We thus introduce two new ONIOM partitions, C1(OH) and C2(OH), in which the HM layer includes also the hydroxyl group bonded to the carbon atom adjacent to the chiral center kept active (see Figure 7 for a schematization of the β anomers).

One could expect that, if the OR for a bichiral system was a purely local property, the increase of the model system (for example, by including the OH groups) would imply an improvement in the description of the property.

From Figure 8, in which we report the ΔS values for the different ONIOM partitions used to describe the C1C2 bichiral system, one can see that it is not exactly what happens.

Passing from the partition C1(C1C2) to C1(OH), the HM becomes more "complete" but the ONIOM description becomes less accurate (the ΔS value increases) for both anomers.

To try to rationalize the results obtained for the bichiral systems, in the following section, we introduce two auxiliary levels of analysis.

5. A Further Analysis in Terms of Anomeric Energy Differences and NMR Shieldings

We introduce here two additional properties to be studied with ONIOM, one of global character, that is, the energy, and one of local character, that is, the nuclear shielding. Both these quantities have been deeply analyzed with ONIOM approaches in many previous papers. This experience will thus allows us to use them as a comparative tool to better appreciate the global/local character of the OR for bichiral systems and to rationalize the results obtained in the previous section.

5.1. Energy Differences. As far the energy is concerned, it is well-known that ONIOM can only be used to get relative values, and thus here we shall consider differences in the α and β anomers. In Figure 9, the (β − α) energy differences for different ONIOM partitions depicted in Figure 3 are reported.

As it can be seen from Figure 9, not all of the partitions properly describe the (β − α) energy difference. In particular,

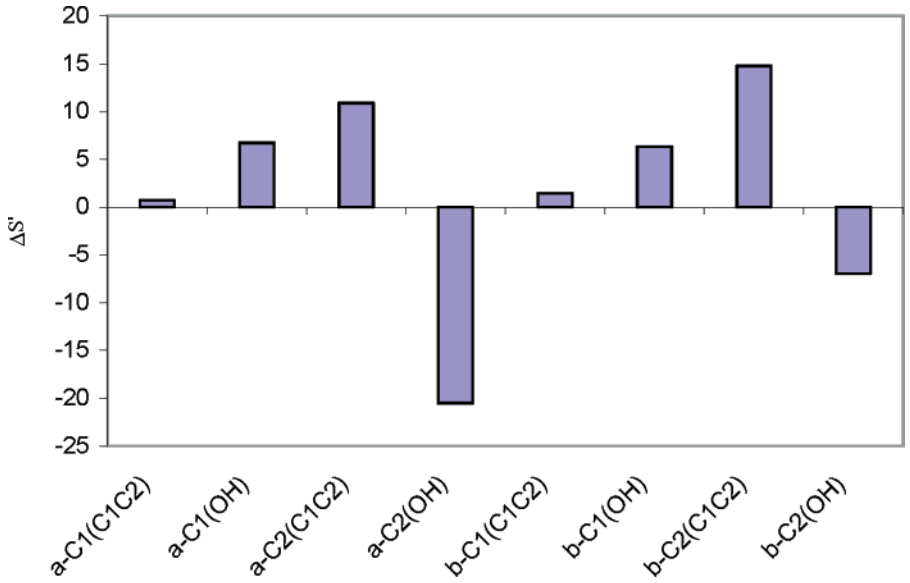


Figure 8. ΔS values {in degrees/[dm(g/cm^3)]} for the different ONIOM partitions of the α (left) and β (right) C1C2 bichiral systems.

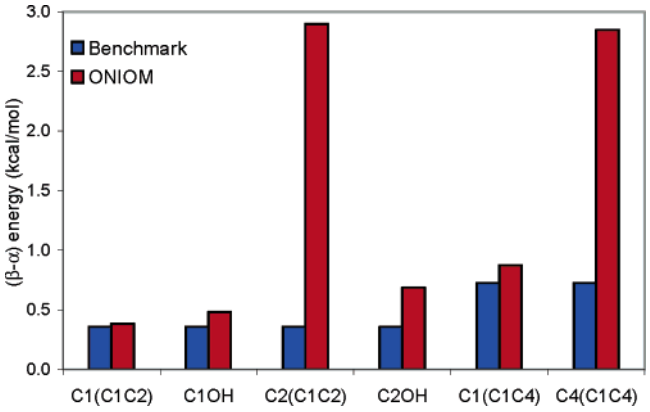


Figure 9. Benchmark and ONIOM values for the (β − α) energy difference in kcal/mol for the different ONIOM partitions.

the C2(C1C2) and C4(C1C4) give a completely wrong picture. To better understand this behavior, it is important to analyze the nature of these ONIOM partitions.

As can be seen from Figure 3, for the C4(C1C4) partition, the anomeric center (C1−OH) is not included in the HM; therefore, the anomeric (β − α) difference is described only at the low level (i.e., in the LR); this prevents a correct evaluation of the differential stability of the two anomers. By contrast, in the C2(C1C2) partition, the anomeric carbon is in the HM but not the corresponding hydroxyl group. As the anomeric effect represents the axial preference of such a OH, it is obvious that we cannot properly describe the differential stability of the two anomers without including the whole moiety in the HM. This is confirmed by the comparison of the C2(C1C2) and C2(OH) results; the latter partition in fact is exactly identical to the former with the only exception of the inclusion of the anomeric OH in the HM layer.

5.2. Carbon Nuclear Shieldings. We finally present the results for the "local" nuclear shieldings of the C1, C2, and C3 carbon atoms.

In Figure 10, we report the absolute ΔS values for the carbon nuclear shieldings of the α (left) and β (right) anomers corresponding to the different ONIOM partitions.

First of all, we note that for all the partitions the ONIOM error is always less than 5% (the absolute values of the nuclear shieldings for the various carbon atoms range from 90 to 160 ppm). This shows, once more, that ONIOM can be reliably used to compute nuclear shieldings also for cyclic systems.[28b]

Going into more detail, we can observe that for C1 the largest error is found for C2(C1C2), that is, where the hydroxyl group is not included in the HM. For C2, on the contrary, the less accurate result is found for C1(C1C2); this can be due to two different effects, one, as in the previous case, related to the OH being excluded from the HM and the other related to what we can call a "border effect", that is, the presence of the "cut" between low and high layers exactly at the C2 atom. By comparing with the C1(OH) result for the same C2 atom, we can observe that the OH effect is more intense than the "border" one, because in this partition the error in the ONIOM description diminishes with the inclusion of the OH group into the HM layer. For the C3 atom, we observe a regular behavior which can be easily explained in terms of border effect, the only active in this case.

From both the energy and the nuclear shielding analyses, it clearly comes out that the most important factor in these systems is the proper description of the anomeric moiety which can only be obtained by including all the related atoms in the HM.

In section 4.3, we have shown that this picture does not apply for the OR; there, however, we have implicitly assumed a "local" character of the property because the comparison was made using absolute values. It is thus worth checking if a different analysis, that is, assuming the OR as a global property, can help in rationalizing the ONIOM behavior for this property. To do that, we introduce here the difference of the OR values for α and β anomers using the

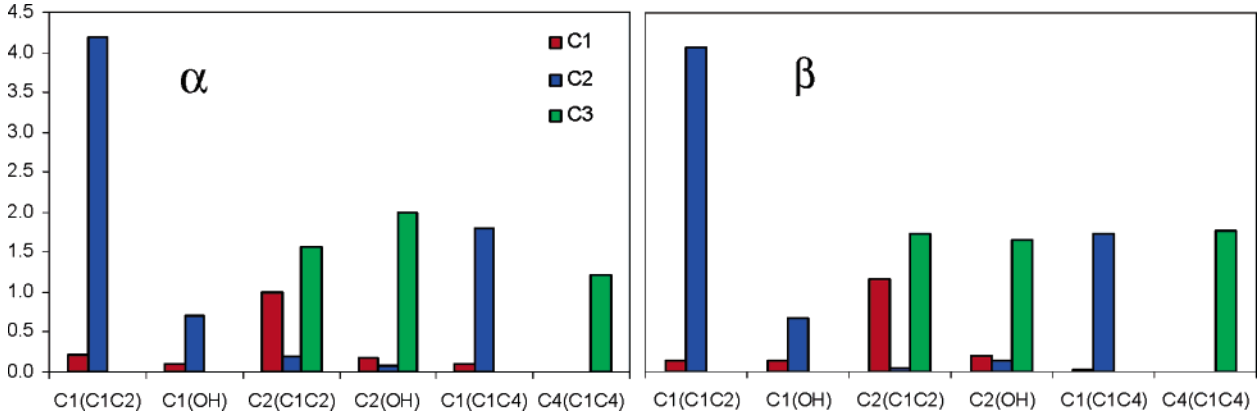


Figure 10. ΔS values (in ppm) for the C1, C2, and C3 carbon nuclear shieldings of the α and β anomers.

Table 4. ΔS Values {in degrees/[dm(g/cm^3)]} for the Differences in the [α] of the Two Anomers ($OR_\alpha - OR_\beta$)

	$\Delta S([\alpha]_\alpha - [\alpha]_\beta)$
C1(C1C2)	0.69
C2(C1C2)	−3.87
C1(C1C4)	−22.69
C4(C1C4)	7.04

different ONIOM partitions. Let us analyze the data reported in Table 4 by considering first the C1C2 bichiral system.

In section 4.1, we have suggested that two chiral centers do not equally contribute to the net chiral behavior, but instead the C1 center plays the major role in both α and β anomers. This is confirmed by these results; in fact, the partition which includes C1 in the HM gives a very low error. By contrast, when we consider the nonadjacent C1C4 system for which, in section 4.1, we were not able to identify a dominant chiral center (at least not for both anomers simultaneously), a far more irregular behavior is found. In fact, a larger error is found passing from the C4(C1C4) not including the anomeric moiety in the HM to the complementary C1(C1C4), that is, a completely opposite behavior with respect to what is found for the energy difference (see Figure 9).

6. Conclusions

In this work, we have presented for the first time a study of the local or global character of the optical rotation of mono- and bichiral prototypes of glucose using the ONIOM method. The results obtained for monochiral systems seem to suggest that OR can be considered as a local property at least in the ONIOM sense, and thus it can be studied by applying the same strategy used for other local properties (like NMR), namely, a proper definition of the low and high partition and a correct combination of the corresponding levels of calculation in terms of the *S* test.

By contrast, the bichiral systems studied present a much more complex behavior. In these systems, the property is not well behaved; that is, it cannot be classified as a purely local or purely global property. The limited dimensions of the prototypes chosen allow only an approximated analysis of ONIOM partitions including both chiral systems, and thus the reasons for the observed behavior remain largely unclear. Some guesses, however, can be given. First of all, it is immediate to accept that OR in chiral systems without asymmetric atom(s) (such as helicenes or similar systems) is necessarily nonlocal. By contrast, in the most common cases in which chirality is determined by one or more asymmetric atoms (like in the systems studied here), the local or nonlocal character strongly depends on the number of chiral centers and their proximity.[35] Further analyses on larger bichiral (cyclic and open) systems with different distances between chiral centers will surely lead to a more complete picture of the nature of OR and to verification of whether a layered method like ONIOM can be applied to the study of this property. These studies are in progress.

Acknowledgment. C.O.d.S. thanks the Brazilian agencies CNPq and FAPERJ. The authors are very grateful to Dr. Thom Vreven, Prof. Jacopo Tomasi, and Prof. Keiji Morokuma for the discussions and suggestions about this work.

References

(1) Brewster, J. H. *J. Am. Chem. Soc.* **1959**, *81*, 5475−5483; 5483−5493.

(2) Rees, D. A. *J. Chem. Soc. B* **1970**, 877−884.

(3) Hudson, C. S. *J. Am. Chem. Soc.* **1925**, *47*, 268−280.

(4) Whiffen, D. H. *Chem. Ind. (London)* **1956**, 964.

(5) Lemieux, R. U.; Brewer, J. T. *Adv. Chem. Ser.* **1973**, *117*, 121−143.

(6) (a) Kirkwood, J. G. *J. Chem. Phys.* **1937**, *5*, 479−491. (b) Kirkwood, J. G. *J. Chem. Phys.* **1939**, *7*, 139−145.

(7) (a) Applequist, J. R.; Carl, J. R.; Fung, K.-K. *J. Am. Chem. Soc.* **1972**, *94*, 2952−2960. (b) Applequist, J. R. *J. Am. Chem. Soc.* **1973**, *95*, 8258−8262.

(8) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341−350.

(9) Birge, R. R.; Schick, G. A.; Bocian, D. F. *J. Chem. Phys.* **1983**, *79*, 2256−2264.

(10) Stevens, E. S.; Sathyanarayana, B. K. *Carbohydr. Res.* **1987**, *166*, 181−193.

(11) Polavarapu, P. L.; Zhao, C. *J. Am. Chem. Soc.* **1999**, *121*, 246−247.

(12) (a) Cheeseman, J. R.; Frisch, M. J.; Devlin, F. J.; Stephens, P. J. *J. Phys. Chem. A* **2000**, *104*, 1039−1046. (b) Stephens, P. J.; Devlin, F. J.; Cheeseman, J. R.; Frisch, M. J. *J. Phys. Chem. A* **2001**, *105*, 5356−5371.

(13) Ruud, K.; Helgaker, T. *Chem. Phys. Lett.* **2002**, *352*, 533−539.

(14) Polavarapu, P. L. *Chirality* **2002**, *14*, 768−781.

(15) (a) Stephens, P. J.; McCann, D. M.; Cheeseman, J. R.; Frisch, M. J. *Chirality* **2005**, *17*, S52−S54. (b) Stephens, P. J.; Devlin, F. J.; Cheeseman, J. R.; Frisch, M. J.; Bortolini, O.; Besse, P. *Chirality* **2003**, *15*, S57−S64.

(16) Crawford, T. D. *Theor. Chem. Acc.* **2006**, "online first".

(17) (a) Grimme, S. *Chem. Phys. Lett.* **2001**, *339*, 380−388. (b) Grimme, S.; Furche, F.; Ahlrichs, R. *Chem. Phys. Lett.* **2002**, *361*, 321−328. (c) Grimme, S.; Bahlmann, H.; Haufe, G. *Chirality* **2002**, *14*, 793−797.

(18) (a) Autschbach, J.; Patchkovskii, S.; Ziegler, T.; van Gisbergen, S. J. A.; Baerends, E. J. *J. Chem. Phys.* **2002**, *117*, 581−592. (b) Jorge, F. E.; Autschbach, J.; Ziegler, T. *J. Am. Chem. Soc.* **2005**, *127*, 975−985. (c) Autschbach, J.; Jensen, L.; Schatz, G. C.; Electra Tse, Y. C.; Krykunov, M. *J. Phys. Chem. A* **2006**, *110*, 2461−2473.

(19) Pecul, M.; Ruud, K. *Int. J. Quantum Chem.* **2005**, *104*, 916−829.

(20) (a) Stephens, P. J.; Devlin, F. J.; Cheeseman, J. R.; Frisch, M. J.; Mennucci, B.; Tomasi, J. *Tetrahedron: Asymmetry* **2000**, *11*, 2443−2448. (b) Mennucci, B.; Tomasi, J.; Cammi, R.; Cheeseman, J. R.; Frisch, M. J.; Devlin, F. J.; Gabriel, S.; Stephens, P. J. *J. Phys. Chem. A* **2002**, *106*, 6102−6113.

(21) Marchesan, D.; Coriani, S.; Forzato, C.; Nitti, P.; Pitacco, G.; Ruud, K. *J. Phys. Chem. A* **2005**, *109*, 1449−1453.

(22) da Silva, C. O.; Mennucci, B.; Vreven, T. *J. Org. Chem.* **2004**, *69*, 8161−8164.

(23) (a) da Silva, C. O.; Nascimento, M. A. C. *Carbohydr. Res.* **2004**, *339*, 113−122. (b) da Silva, C. O. *Theor. Chem. Acc.* **2006**, *116*, 137−147.

(24) (a) Stortz, C. A.; Cerezo, A. S. *Carbohydr. Res.* **2003**, *338*, 1679−1689. (b) Dixon, A. M.; Venable, R.; Widmalm, G.; Bull, T. E.; Pastor, R. W. *Biopolymers* **2003**, *69*, 448−460.

(25) (a) Humbel, S.; Sieber, S.; Morokuma, K. *J. Chem. Phys.* **1996**, *16*, 1959−1967. (b) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357−19363. (c) Froese, R. D. J.; Morokuma, K. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Kollman, P. A., Clark, T., Schaefer, H. F., III, Gasteiger, J., Schreiner, P. R., Eds.; Wiley: Chichester, U. K., 1998; Vol. 2, p 1244. (d) Dapprich, S.; Komaromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *THEOCHEM* **1999**, *461−462*, 1−21. (e) Vreven, T.; Morokuma, K. *J. Comput. Chem.* **2000**, *21*, 1419−1432.

(26) Morokuma, K. *Bull. Korean Chem. Soc.* **2003**, *24*, 797−801. See also ref 22d and references therein.

(27) Vreven, T.; Mennucci, B.; da Silva, O.; Morokuma, K.; Tomasi, J. *J. Chem. Phys.* **2001**, *115*, 62−72.

(28) (a) Karadakov, P. B.; Morokuma, K. *Chem. Phys. Lett.* **2000**, *317*, 589−596. (b) Rickard, G. A.; Karadakov, P. B.; Webb, G. A.; Morokuma, K. *J. Phys. Chem. A* **2003**, *107*, 292−300.

(29) Rosenfeld L. *Z. Physik.* **1928**, *52*, 161−174.

(30) Amos, R. D. *Chem. Phys. Lett.* **1982**, *87*, 23−26.

(31) (a) London, F. J. *Phys. Radium* **1937**, *8*, 3974−3976. (b) Ditchfield, R. *Mol. Phys.* **1974**, *27*, 789−807.

(32) Helgaker, T.; Ruud, K.; Bak, K. L.; Jørgensen, P.; Olsen, J. *Faraday Discuss.* **1994**, *99*, 165−180.

(33) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision A.1; Gaussian, Inc.: Pittsburgh, PA, 2003.

(34) Wiberg, K. B.; Wang, Y.; Vaccaro, P. H.; Cheeseman, J. R.; Trucks, G.; Frisch, M. J. *J. Phys. Chem. A* **2004**, *108*, 32−38.

(35) Mislow, K.; Siegel, J. *J. Am. Chem. Soc.* **1984**, *106*, 3319−3328.

CT600250W

JCTC Journal of Chemical Theory and Computation

Local Møller−Plesset Perturbation Theory: A Massively Parallel Algorithm

Ida M. B. Nielsen* and Curtis L. Janssen

Sandia National Laboratories, P.O. Box 969, Livermore, California 94551

Received June 1, 2006

Abstract: A massively parallel algorithm is presented for computation of energies with local second-order Møller−Plesset (LMP2) perturbation theory. Both the storage requirement and the computational time scale linearly with the molecular size. The parallel algorithm is designed to be scalable, employing a distributed data scheme for the two-electron integrals, avoiding communication bottlenecks, and distributing tasks in all computationally significant steps. A sparse data representation and a set of generalized contraction routines have been developed to allow efficient massively parallel implementation using distributed sparse multidimensional arrays. High parallel efficiency of the algorithm is demonstrated for applications employing up to 100 processors.

1. Introduction

The high-degree polynomial scaling of conventional molecular orbital-based correlated electronic structure methods is an obstacle to their application to mainstream chemical problems. Even for one of the computationally least expensive correlated methods, second-order Møller−Plesset (MP2) perturbation theory, the computational cost formally scales as $O(n^5)$, where n is the size of the molecular system.

The computational scaling may be improved, however, by formulating the equations in a basis of localized orbitals instead of the canonical molecular orbitals, which are by nature delocalized. This idea has been utilized by Pulay and Saebø in their work on the so-called local correlation method.[1−3] Using localized orbitals, Pulay and Saebø proposed restricting the excitation space for a given occupied orbital to a domain consisting of virtual orbitals that are all spatially close to the occupied orbital out of which the excitation occurs. By reducing the number of configurations to be included in the description of the wave function, this restriction of the correlation space may be utilized to attain reduced computational scaling of the correlation procedure.

Many applications of the local correlation idea, using various choices for the localized orbitals, have been presented for Møller−Plesset perturbation theory. Pulay and co-workers have developed local formulations for second- through fourth-order Møller−Plesset perturbation theory[1,2,4] using an orbital-invariant formulation employing localized orbitals and requiring an iterative solution of the amplitude equations, and low-order scaling implementations of the local MP2 method have been achieved.[5,6] Scuseria and Ayala[7] used a Laplace transform method[8] to obtain an atomic orbital-based linear scaling MP2 algorithm, and Friesner and co-workers[9,10] have implemented a pseudospectral localized MP2 method using the local correlation idea of Pulay and Saebø. Maslen and Head-Gordon[11] have developed a local MP2 method using nonorthogonal orbitals and an atoms-in-molecules local ansatz, and this method has been further developed to include a diatomics- and a triatomics-in-molecules approach.[12,13] Notably, Schütz and co-workers[14,15] employed the Pulay−Saebø local correlation scheme, using localized molecular orbitals for the occupied space and atomic orbitals, projected against the occupied space, for the virtual space, and achieved an algorithm that formally scales linearly with molecular size for local second-order Møller−Plesset theory (LMP2).

The achievement of low-order or linearly scaling implementations of MP2 methods has widened the range of molecules to which they can be applied, but, nonetheless, the application of reduced-scaling correlation methods to large molecules places heavy demands on computing resources. The scope of reduced-scaling correlation methods may be further extended, however, by developing algorithms tailored to application to massively parallel computers. In

* Corresponding author e-mail: ibniels@ca.sandia.gov.

10.1021/ct600188k CCC: $37.00
Published on Web 11/22/2006

this work, we focus our attention on the local MP2 method LMP2, using the local correlation idea of Pulay and Saebø in the context of MP2 theory, to develop a linear scaling massively parallel LMP2 algorithm. We are not aware of any parallel implementations of the LMP2 method reported in the literature, although a massively parallel algorithm for doing conventional (O(n^5)) MP2 using the orbital-invariant formulation (on which the LMP2 method is based) has been reported.[16] Additionally, the pseudospectral localized MP2 method has been parallelized previously, and the parallel performance has been illustrated for runs employing up to 16 processors.[17]

2. Theory

We will employ the local correlation scheme developed by Pulay and Saebø,[1−3] using an occupied orbital space represented by molecular orbitals localized by the Pipek−Mezey procedure[18] and an unoccupied space consisting of atomic orbitals that have been projected against the occupied orbital space to ensure orthogonality of the occupied and virtual spaces. The localized occupied orbitals are expressed as linear combinations of atomic orbitals using the coefficient matrix **L**

$$|i\rangle = \sum_{\mu} |\mu\rangle L_{\mu i} \tag{1}$$

with indices i and μ representing a localized molecular orbital and an atomic orbital, respectively. The projected atomic orbitals are obtained from the original atomic orbitals using the projection matrix **P** with elements $P_{\mu r}$

$$|r\rangle = \sum_{\mu} |\mu\rangle P_{\mu r} \tag{2}$$

$$\mathbf{P} = \frac{1}{2}\mathbf{D}_{\mathrm{v}}\mathbf{S} \tag{3}$$

where r represents a projected atomic orbital, **S** is the overlap matrix in the unprojected atomic orbital basis, and $\mathbf{D}_{\mathrm{v}}$ denotes the closed-shell Hartree−Fock density matrix constructed from the virtual orbitals. By forming **P** this way, rather than using $\mathbf{P} = \mathbf{I} - (1/2)\mathbf{DS}$ as has been done previously,[6,14,19] we avoid difficulties that may arise if the occupied density matrix **D** is constructed from a truncated set of atomic orbitals (when some orbitals have been eliminated in the Hartree−Fock procedure to avoid near linear dependencies), whereas **I** represents the full AO space. We note that, while the unoccupied space is not orthogonal, the occupied and unoccupied spaces are mutually orthogonal, and the localized occupied orbitals are orthogonal among themselves.

In the following, the indices i, j, k denote localized occupied molecular orbitals, and the indices r, s represent projected atomic orbitals. For each localized occupied orbital i, a domain [i] is created, comprising all projected atomic orbitals associated with a certain subset of the atoms. This subset is determined by including the atoms whose atomic orbitals contribute the most to the Mulliken population of i, until a spanning criterion, measuring how well the included atomic orbitals span orbital i, is satisfied. The procedure is described in detail elsewhere.[19,20] Our implementation uses a default threshold of 0.02 for the residual error associated with the spanning criterion,[19,20] but a different (typically smaller) threshold can be specified in the input, if desired. Using an orbital-invariant formulation of MP2 theory, the MP2 correlation energy can be expressed as[1]

$$E_{\mathrm{MP2}}^{\mathrm{corr}} = \sum_{ij} \mathrm{Tr}[\mathbf{K}_{ij}(2\mathbf{T}_{ji} - \mathbf{T}_{ij})] = \sum_{ijrs} K_{ij}^{rs}(2T_{ij}^{rs} - T_{ij}^{sr}) \tag{4}$$

where $\mathbf{K}_{ij}$ is a matrix whose elements K_{ij}^{rs} are the two-electron integrals ($ri|sj$), and $\mathbf{T}_{ij}$ is the matrix of double substitution amplitudes T_{ij}^{rs}. The double substitution amplitudes are determined from the set of linear equations

$$\mathbf{R}_{ij} = \mathbf{K}_{ij} + \mathbf{F}\mathbf{T}_{ij}\mathbf{S} + \mathbf{S}\mathbf{T}_{ij}\mathbf{F} - \mathbf{S}\sum_{k}[F_{ik}\mathbf{T}_{kj} + F_{kj}\mathbf{T}_{ik}]\mathbf{S} = 0 \tag{5}$$

which must be solved iteratively. $\mathbf{R}_{ij}$ represents the residual matrix with elements R_{ij}^{rs}, **F** and **S** are the virtual−virtual blocks of the Fock matrix and the overlap matrix, respectively, in the projected atomic orbital basis, and F_{ik} is an element of the Fock matrix in the basis of localized occupied orbitals.

In the iterative solution of eq 5, the double substitution amplitudes are updated using a procedure previously described for the local coupled-cluster method.[19] This procedure involves transforming the residual matrix $\mathbf{R}_{ij}$ to an orthogonal basis, computing an update to the double substitution amplitudes from the transformed residual elements

$$\Delta T_{ij}^{rs} = \frac{-R_{ij}^{rs}}{\epsilon_r + \epsilon_s - F_{ii} - F_{jj}} \tag{6}$$

and transforming the computed amplitude update back to the original projected atomic orbital basis. The transformation to an orthogonal basis is performed for each ij pair individually using a transformation matrix obtained by diagonalization of the Fock matrix in the subspace spanned by the projected atomic orbitals in the [ij] pair domain, and ϵ_r, ϵ_s in eq 6 denote diagonal elements of the Fock matrix in this basis.

The local MP2 method uses the orbital-invariant expressions from above and reduces the computational scaling of the MP2 procedure as follows. The excitation space is restricted by allowing double excitations out of occupied orbitals into only the virtual orbitals in the corresponding orbital domains. Thus, for an occupied orbital pair ij, double excitations are allowed only into the pair domain [ij], which is formed by combining the individual orbital domains [i] and [j]. To further reduce the number of double substitution amplitudes, excitations are allowed out of only a subset of the ij pairs, namely those pairs for which the minimum distance R_{ij} between any two atoms in the domains [i] and [j] is less than a certain threshold value. Using these restrictions on the included double substitutions, the number of double substitution amplitudes and integrals K_{ij}^{rs} will grow only linearly with the size of the molecule, and the computation of the LMP2 correlation energy (from eq 4), thus, scales linearly with the molecular size. To achieve a linearly scaling LMP2 procedure, the integral transformation generating the K_{ij}^{rs} integrals and the iterative solution of eq

Localize occupied molecular orbitals
Create domains
Compute screening quantities required for integral transformation
Perform integral transformation generating $(ri|sj)$

Begin LMP2 iterations
  Compute LMP2 residual
  Compute $\Delta\mathbf{T}$
  DIIS extrapolation of $\mathbf{T}$
  Compute $E_{\mathrm{MP2}}^{\mathrm{corr}}$, $\Delta E_{\mathrm{MP2}}^{\mathrm{corr}}$
  Check for convergence on $\Delta E_{\mathrm{MP2}}^{\mathrm{corr}}$ and $\Delta\mathbf{T}$
End LMP2 iterations

Figure 1. Outline of the local MP2 procedure.

5 must also scale linearly, however, and this requires application of integral prescreening throughout the transformation and utilization of the sparsity of the overlap and Fock matrices. The employed screening procedures will be explained in more detail below.

3. A Scalar LMP2 Algorithm

In this section we describe the scalar implementation of the LMP2 method from which our parallel algorithm has been developed. An outline of the steps involved in the computation of the LMP2 energy is given in Figure 1. Initially, the occupied molecular orbitals are localized, the domains are created, and the various screening quantities that must be employed to achieve linear scaling are computed. The transformation matrices required in the iterative procedure to transform the residual to an orthogonal basis (cf. section 2) are also computed at this point and stored for the remainder of the calculation. The two-electron integral transformation is then performed to generate the transformed integrals required for the LMP2 procedure. The LMP2 integral transformation is similar to that of a conventional MP2 computation but uses different transformation matrices and a different screening protocol. The screening employed in the integral transformation is outlined in Appendix A. The effect of varying the various thresholds used in the screening required to achieve linear scaling in the local MP2 procedure has been investigated in some detail by Schütz et al.[14] and by Saebø and Pulay.[6] In the present work, we have used the thresholds given in Appendix A.

When the two-electron integral transformation is complete, the iterative procedure is entered. In each iteration, the residual is first computed from eq 5, and an update, $\Delta\mathbf{T}$, for the double substitution amplitudes is then computed from eq 6. Using $\Delta\mathbf{T}$ and DIIS extrapolation, if desired, the double substitution amplitudes $\mathbf{T}$ are then updated, and a new value for the energy $E_{\mathrm{corr}}^{\mathrm{MP2}}$ is computed. Finally, a check for convergence on both $\Delta\mathbf{T}$ and $\Delta E_{\mathrm{corr}}^{\mathrm{MP2}}$ is performed.

In our implementation, all four-index quantities, i.e., the two-electron integrals K_{ij}^{rs} (as well as the half- and three-quarter-transformed integrals), the double substitution amplitudes T_{ij}^{rs}, and the residual elements R_{ij}^{rs} are stored as multidimensional sparse arrays. In previous implementations, these quantities have been stored as matrices, e.g., storing matrices $\mathbf{K}_{ij}$ with elements K_{ij}^{rs}.[6,14] We have implemented a parallel sparse data representation to handle distributed sparse multidimensional arrays and also a set of generalized contraction routines to perform multiplications of such arrays. Using this data representation and the associated contraction routines greatly simplifies programming, eliminating the need for explicit index manipulation, allowing contractions to be performed in a single line of code, and making parallelization more straightforward. The parallel sparse multidimensional arrays and contraction routines are discussed in more detail in the next section and in Appendix B.

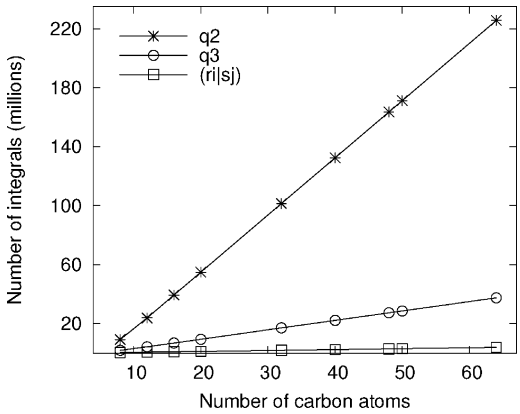


Figure 2. The size of the integral arrays required for the LMP2 procedure for linear alkanes C_nH_{2n+2} using the cc-pVDZ basis; q2 and q3 represent the half-transformed and three-quarter-transformed integral arrays, respectively.

For the last two terms in eq 5, involving left and right multiplication with the overlap matrix, the multiplication can be carried out either inside or outside of the summation over *k*. Schütz et al.[14] found that performing the multiplication inside of the summation was faster because the multiplication could be carried out with only the small local dimensions of the overlap matrix. Saebø and Pulay,[6] however, found multiplication outside of the summation to be the faster alternative in their local MP2 program. We have chosen to perform the multiplications with the overlap matrix outside of the summation because this makes possible a straightforward parallel implementation of this step using our parallel sparse data representation for **S**, **F**, and **T** (vide infra).

To illustrate the performance of the scalar LMP2 algorithm, computations were performed for a series of linear alkanes C_nH_{2n+2} using the cc-pVDZ basis set,[21] freezing the core orbitals, and using a cutoff of 8.0 bohr for the distance threshold R_{ij} (cf. section 2) for including *ij* pairs. For the computations reported here, convergence of the energy to 10^{-7} hartrees required 10 iterations. We note that a dynamic update scheme has been used previously,[5,6,14] updating the double substitution amplitudes as soon as the corresponding residual elements have been computed. This procedure speeds up convergence but would entail interprocessor communication (and result in nondeterministic convergence) when parallelized. We have elected to implement a parallel computation of the residual that does not require interprocessor communication (vide infra), updating all double substitution amplitudes at once and using the DIIS procedure[22] to accelerate convergence. The scaling behavior of the algorithm is shown in Figures 2 and 3. The number of transformed two-electron integrals, which, once generated, must be kept for the duration of the LMP2 computation, scales linearly with the molecular size. Likewise, the size of the intermediate arrays (partially transformed two-electron

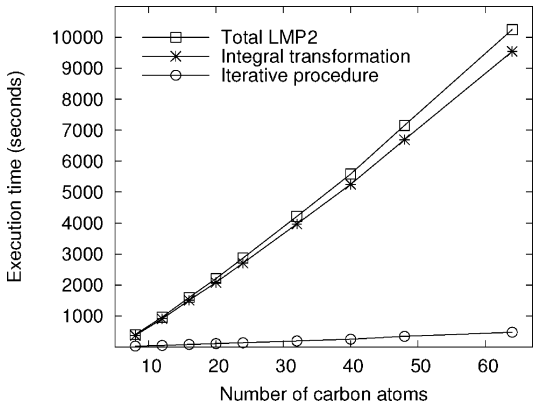


Figure 3. Computational times (wall times) for the LMP2 procedure obtained for linear alkanes C_nH_{2n+2} using the cc-pVDZ basis. The computations were performed on an AMD 2.4 GHz Opteron processor.

integrals), which are required during the integral transformation only, increases linearly with the molecular size. The number of double substitution amplitudes T_{ij}^{rs} is the same as the number of transformed integrals ($ri|sj$).

The computational time (wall time) required for the integral transformation, the iterative procedure, and the full LMP2 procedure is illustrated in Figure 3. Both the integral transformation and the iterative procedure scale linearly with the molecular size. The scaling obtained for the entire LMP2 procedure is slightly above linear because the timings include the time required to localize the orbitals, and this step, although fast, scales as n^3.

4. A Massively Parallel LMP2 Algorithm

Using the linear scaling LMP2 implementation described in the previous section, we have developed a massively parallel LMP2 algorithm. The parallel LMP2 program has been implemented as part of the massively parallel quantum chemistry program MPQC.[23] We have chosen to design a parallel LMP2 algorithm that uses only standard MPI collective communication to ensure portability of the code to architectures for which one-sided message passing is not available. To create a scalable algorithm, all large data arrays have been distributed, all computationally significant steps have been parallelized, and a communication scheme that does not create bottlenecks has been used. Both data and computational tasks are distributed over two occupied indices, allowing efficient utilization of a large number of processors.

The computationally dominant steps in the LMP2 procedure are the two-electron integral transformation and the computation of the LMP2 residual in the iterative procedure. The computation of the correlation energy (eq 4) and the update of the amplitudes (eq 6) are fast steps that are parallelized by letting each node process only the terms corresponding to the locally held *ij* pairs. We will describe the parallel implementation of the two-electron integral transformation and the computation of the residual in detail below.

4.1. Parallel Two-Electron Integral Transformation. Two-electron integrals of the type ($ri|sj$) with two occupied

Loop over all M, N shells $|M \geq N$
  If (my M, N pair):
  Loop over all R, S shells
    Allocate AO integral block $(\underline{M}R|\underline{N}S)$
  end R,S loop
  Compute $(\underline{M}R|\underline{N}S)$ for current M, N pair and all R, S
  Loop over R, S blocks of $(\underline{M}R|\underline{N}S)$
    Loop over j
      Allocate $(\underline{M}R|\underline{N}j)$ block
    end j loop
  end R, S loop
  $(\underline{M}R|\underline{N}j) = (\underline{M}R|\underline{N}S) * L(S, j)$
  Loop over R, j blocks of $(\underline{M}R|\underline{N}j)$
    Loop over $i|ij$ valid pair
      Allocate $(\underline{M}i|\underline{N}j)$ block
    end i loop
  end R, j loop
  $(\underline{M}i|\underline{N}j) = (\underline{M}R|\underline{N}j) * L(R, i)$
end $\underline{M}, \underline{N}$ loop

Redistribute half-transformed integrals:
$(\underline{M}i|\underline{N}j)|M \geq N \rightarrow (M\underline{i}|N\underline{j})|i \geq j$

Loop over blocks $M, N, \underline{i}, \underline{j}$ of $(M\underline{i}|N\underline{j})$
  Loop over $s|s \in [ij]$
    Allocate $(M\underline{i}|s\underline{j})$ block
  end s loop
end $M, N, \underline{i}, \underline{j}$
$(M\underline{i}|s\underline{j}) = (M\underline{i}|N\underline{j}) * P(N, s)$

Loop over blocks $M, \underline{i}, s, \underline{j}$ of $(M\underline{i}|s\underline{j})$
  Loop over $r|r \in [ij]$
    Allocate $(r\underline{i}|s\underline{j})$ block
  end r loop
end $M, \underline{i}, s, \underline{j}$ loop
$(r\underline{i}|s\underline{j}) = (M\underline{i}|s\underline{j}) * P(M, r)$

Figure 4. Outline of the parallel integral transformation. Indices *M*, *R*, *N*, *S* represent shells of atomic orbitals. The transformation matrices **L** and **P** are defined in the text. Distributed indices are underlined. The steps in which the integrals ($\underline{M}R|\underline{N}S$), ($\underline{M}R|\underline{N}j$), ($\underline{M}i|\underline{N}j$), ($M\underline{i}|s\underline{j}$), and ($r\underline{i}|s\underline{j}$) are allocated are preceded by integral screening as outlined in Appendix A.

and two unoccupied indices are required for the LMP2 computation. The fully transformed integrals ($ri|sj$) can be kept in aggregate memory without introducing serious memory bottlenecks, but some of the partially transformed integral arrays are too large to fit in memory, and these arrays are therefore generated in a direct fashion that requires storage of only a small subsection of the arrays at any given time. To facilitate evenly distributed loads, the integrals are distributed over two indices, and in the iterative procedure, each node will process only the locally held integrals. The parallelization scheme employed here is similar to that used by Wong et al.[24] in a canonical parallel integral transformation.

The parallel integral transformation is outlined in Figure 4. The screening steps are not shown but are performed as described in Appendix A. The transformation involves computation of the two-electron integrals in the atomic orbital (AO) basis and four subsequent quarter transformations, each transforming one index. The transformation matrices are the matrix **P** for the unoccupied indices and the matrix **L** for the occupied indices. The computation of the AO integrals and the first two quarter transformations are performed within a loop over pairs of shells *M* and *N* ($M \geq N$), and these steps are parallelized by distributing the *M*, *N* pairs across nodes. For each *M*, *N* shell pair, the AO integral array

(*MR*|*NS*) is computed for all contributing *R*, *S*, and this array is then transformed to generate the quarter-transformed integrals (*MR*|*Nj*) for the current *M*, *N* pair. The (*MR*|*Nj*) array, subsequently, is transformed to generate the half-transformed integrals (*Mi*|*Nj*). The number of AO integrals and quarter-transformed integrals can become quite large, even though the size of the arrays scales linearly with the molecular size. A memory bottleneck is avoided by only generating these arrays for one *M*, *N* shell pair at a time before they are transformed and discarded. The half-transformed integral array is much smaller, and this array, and, subsequently, the three-quarter and fully transformed integral arrays, can be kept in memory. The first half-transformation generates the integrals (*Mi*|*Nj*) distributed over *M*, *N*, and no interprocessor communication is required.

After completion of the first half-transformation, a redistribution of the half-transformed integrals is performed, replacing the distribution over *M*, *N* shell pairs with a distribution over *i*,*j* pairs ($i \geq j$). This is the only communication step required in the entire integral transformation, and after the redistribution of integrals, each node independently completes the integral transformation generating the fully transformed integrals (*ri*|*sj*). The work in the second half-transformation is distributed by letting each node process only the *i*,*j* pairs held locally, and upon completion of the integral transformation, each node holds the (*ri*|*sj*) array for a subset of the included *i*,*j* pairs.

We note that the parallel integral transformation exploits the $M \geq N$ symmetry in the computation of the AO integrals and in the first two quarter transformations. Thus, a 4-fold redundancy is required in the AO integral computation. In the last two quarter transformations, the $i \geq j$ symmetry is utilized, producing fully transformed integrals (*ri*|*sj*) for $i \geq j$, and the included *i*,*j* pairs are distributed across nodes.

4.2. Parallel Computation of the LMP2 Residual. The computation of the LMP2 residual accounts for the majority of the time spent in the iterative procedure. We have implemented two parallel schemes for computation of the residual, employing replicated and distributed double substitution amplitudes, respectively. By replicating the double substitution amplitudes, load balance can more easily be achieved in the iterative procedure, and the amplitudes can be replicated without introducing memory bottlenecks. Replication of the amplitudes, however, necessitates a global summation step involving all the elements of $\Delta\mathbf{T}$ in the iterative procedure, which will reduce parallel performance as the number of processors grows large. An alternative parallel implementation of the residual computation employing distributed double substitution amplitudes has therefore also been investigated.

In the parallel implementation employing replicated double substitution amplitudes, the residual is computed from eq 5 by letting each node compute the residual arrays $\mathbf{R}_{ij}$ for the subset of the *ij* pairs for which the (*ri*|*sj*) integrals are held locally. The first term on the right-hand side of eq 5 is a simple update of the residual with the locally held (*ri*|*sj*) integrals. The second and third terms are contractions of blocks of double substitution amplitudes with virtual−virtual blocks of the Fock and overlap matrices. These terms are computed by looping over the local *ij* pairs on each node and, for the current *ij* block of **T**, performing the contraction as two subsequent matrix multiplications. The last two terms in eq 5 involve multiplication of *k*, *j* and *i*, *k* blocks of the double substitution amplitudes by the occupied−occupied Fock matrix elements F_{ik} and F_{kj}, respectively, forming the term $\sum_k[F_{ik}\mathbf{T}_{kj} + F_{kj}\mathbf{T}_{ik}]$, followed by matrix multiplication with the overlap matrix on the left and right. Again, only the *i*, *j* values corresponding to the locally held *ij* pairs are processed on each node, and because the double substitution amplitudes are replicated, the required $\mathbf{T}_{kj}$ and $\mathbf{T}_{ik}$ elements are readily available. The computation of the residual requires no interprocessor communication, and the distribution of work over two indices makes it possible to achieve load balance even for a large number of processors, provided that the number of processors is significantly smaller than the number of included *ij* pairs. A global communication step, accumulating the individual contributions to $\Delta\mathbf{T}$ (eq 6) computed on each node, is required, however, and as demonstrated below, this step impacts the parallel scaling of the iterative procedure when using a large number of processors.

To avoid the global accumulation step for the double substitution amplitudes, we have also implemented a version of the iterative procedure in which the double substitution amplitudes are distributed. Two different distributions of **T** are then required, namely distribution over *ij* pairs and distribution over just one occupied index. The former distribution is needed for parallel efficiency in the computation of the correlation energy and for the $\mathbf{FT}_{ij}\mathbf{S}$ and $\mathbf{ST}_{ij}\mathbf{F}$ terms in eq 5, and the latter distribution is required for computation of the last two terms on the right-hand side of eq 5. The only communication step required is now a redistribution of **T** (from *ij* pairs to just one occupied index), but this step is scalable, and the overall communication time can therefore be reduced. In this scheme, however, load imbalance becomes a problem as the number of processors increases, because the last two terms in eq 5 are now parallelized by distribution of only one occupied index.

To avoid both communication bottlenecks and load imbalance in the iterative procedure, a one-sided asynchronous message passing scheme could be implemented, and we have explored the use of such schemes in our previous work on massively parallel computation of MP2 energies and gradients.[25,26] Using asynchronous message passing in the iterative LMP2 procedure would allow distribution of the double substitution amplitudes (eliminating the need for global communication) and distribution of work over two indices (achieving load balance) by letting each node request amplitudes from other nodes when needed.

5. Parallel Performance

Computations reported in this section were performed on a Linux cluster with dual-processor 3.6 GHz Xeon nodes connected via an InfiniBand network with a full fat-tree topology.

The parallel performance of the algorithm is illustrated in Figure 5. The speedup curves shown were obtained for a linear alkane, $C_{32}H_{66}$, using the correlation-consistent basis

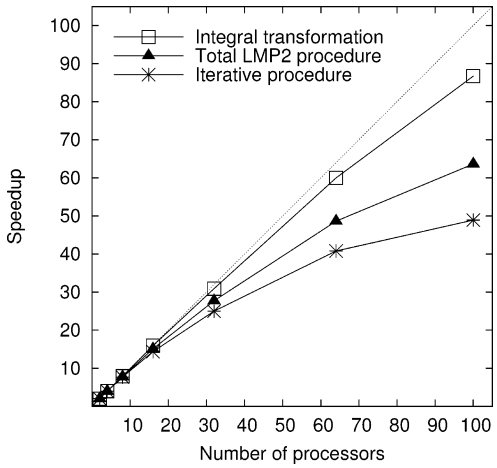


Figure 5. Parallel speedups for the LMP2 procedure obtained for the linear alkane $C_{32}H_{66}$ using the cc-pVDZ basis.

set cc-pVDZ[21] (778 basis functions). Computations were carried out using a number of processors ranging from 1 to 100, employing only one processor per node. The speedups shown were computed relative to the computational time required on one processor. Converging the energy to 10^{-7} hartrees required 10 iterations. As explained above, we have implemented two different versions of the iterative procedure, employing replicated and distributed double substitution amplitudes, respectively. For the chosen test case, load imbalance caused the algorithm using distributed amplitudes to be the slower of the two for all measured processor counts, and the results presented below were obtained using the algorithm employing replicated amplitudes.

The dominant computational steps are the two-electron integral transformation and the iterative procedure, which take up 92% and 7% of the computational time, respectively, when running on one node. As the number of nodes increases, these percentages change somewhat because a few remaining serial steps take up an increasing fraction of the computational time and because the speedups are higher for the integral transformation than for the iterative procedure.

As shown in Figure 5, the parallel efficiency obtained in the integral transformation remains high as the number of processors increases, and a speedup of 87 is obtained using 100 processors. In the integral transformation, work is distributed over two indices, viz., shell pairs in the first half-transformation and pairs of occupied orbitals in the second half-transformation, and provided that the number of processors is significantly smaller than the number of such pairs, high parallel efficiency can be obtained. The only communication step required in the integral transformation is the redistribution of the half-transformed integrals. This step is scalable because the total number of integrals to be communicated by each processor is inversely proportional to the number of processors, and it takes a negligible amount of time, less than 2% of the time required for the integral transformation in all cases.

For the iterative procedure, high parallel efficiency is obtained up to around 32 processors after which the performance starts to degrade, and the efficiency is about 50% when running on 100 processors. As mentioned, the iterative procedure is fast compared to the integral transformation for basis sets that would typically be used in an MP2 computation, so the somewhat lower parallel performance of the iterative procedure does not have a serious impact on the overall parallel performance (vide infra). The performance degradation observed for the iterative procedure is due mainly to a communication step required in each iteration to collect partial contributions to the updated double substitution amplitudes from all nodes.

The speedups obtained for the total LMP2 procedure are also illustrated in Figure 5. The efficiency is high up to about 64 processors (obtaining a speedup of 49 on 64 processors) and degrades as the number of processors increases. The performance degradation observed for the overall speedup is caused mainly by the presence of some small computational steps that have not yet been parallelized and to a lesser extent by the degrading performance in the iterative procedure. The steps that have not been parallelized include the localization of the occupied orbitals, the computation of the domains, and the computation of some of the screening quantities required to attain linear scaling with the size of the molecule.

To further illustrate applications of the LMP2 code, we have computed LMP2 and conventional MP2 energies for a couple of globular molecules, namely two isomers of the boron nitride $C_{14}H_{20}BNO_3$. The two isomers, illustrated in Figure 6, are both candidates for the global minimum for this molecule. The cc-pVDZ, cc-pVTZ, and cc-pVQZ basis set were employed,[21] and the results are shown in Table 1. In all cases, the relative energies computed with the LMP2 method are close to the conventional MP2 energies, with somewhat better agreement for the larger sets. The fraction of the MP2 correlation energy recovered at the LMP2 level increases with basis set improvement, ranging from about 97% at the cc-pVDZ level to 99% at the cc-pVQZ level.

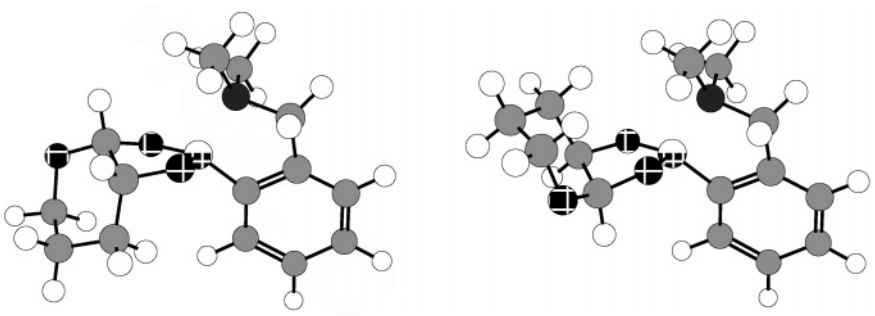


Figure 6. Two isomers of the boron nitride $C_{14}H_{20}BNO_3$ (isomer 1 on the left, isomer 2 on the right).

Table 1. LMP2 and Conventional MP2 Energies for Two Isomers of $C_{14}H_{20}BNO_3$[a]

	LMP2		MP2	
	isomer 1	isomer 2	isomer 1	isomer 2
		cc-pVDZ		
E_{corr}	−2.642280	−2.640938	−2.716272	−2.715842
E_{total}	−848.308810	−848.303178	−848.382802	−848.378083
ΔE	3.53		2.96	
		cc-pVTZ		
E_{corr}	−3.278991	−3.277799	−3.333433	−3.333024
E_{total}	−849.168225	−849.163091	−849.222666	−849.218315
ΔE	3.22		2.73	
		cc-pVQZ		
E_{corr}	−3.516106	−3.515809	−3.549913	−3.549537
E_{total}	−849.458891	−849.454743	−849.492698	−849.488470
ΔE	2.60		2.65	

[a] Relative energies ($\Delta E = E_{total}$[isomer 2] − E_{total}[isomer 1]) in kcal mol^{-1}, other energies in hartrees; employing a domain completeness threshold of 0.01 and an R_{ij} distance threshold of 15.0 bohr (cf. section 2) in the LMP2 procedure.

Finally, to show computational times required for some larger LMP2 computations, we have also computed LMP2 energies with larger basis sets for a couple of the linear alkanes, viz. $C_{24}H_{50}$ using the aug-cc-pVTZ basis[21,27] (2254 basis functions) and $C_{50}H_{102}$ employing the cc-pVTZ set (2928 basis functions). Using 100 nodes, the total wall times required for the LMP2 computation (excluding the Hartree−Fock part) for $C_{24}H_{50}$ ($E_{LMP2} = -942.508278$ hartrees) and $C_{50}H_{102}$ ($E_{LMP2} = -1962.104490$ hartrees) were 102 and 14 min, respectively, using a value of 8.0 bohr for the distance threshold R_{ij} for including ij pairs. The time required for the basis set with 2254 functions is significantly longer than for the set with 2928 functions because the former is an augmented set for which larger domains are required and for which the screening in the integral transformation is less effective.

6. Concluding Remarks

We have demonstrated the feasibility of highly efficient massively parallel computation of local correlation energies by developing and implementing a massively parallel LMP2 algorithm. Our implementation of the LMP2 method scales linearly with the molecular size in the number of two-electron integrals and double substitution amplitudes as well as in the computational time. Using a parallel sparse data representation and a set of generalized contraction routines developed to handle distributed sparse multidimensional arrays, a massively parallel implementation of the LMP2 algorithm has been achieved. The parallel implementation employs only standard MPI collective communication to increase portability of the code, and the communication scheme is designed to prevent serious bottlenecks as the number of processors increases. To avoid memory bottlenecks, the first half of the integral transformation is direct, and the second half of the transformation employs fully distributed data arrays. Load balance has been achieved by distributing all computationally significant tasks over two indices, allowing utilization of a large number of processors. The parallel performance of the algorithm has been illustrated, and high parallel efficiency has been demonstrated for computations performed using up to 100 processors.

Appendix A. Screening of Two-Electron Integrals

To achieve linear scaling in the integral transformation, it is necessary to employ integral screening for the computation of the AO integrals and in each of the four subsequent transformation steps. Each of the five steps in the integral transformation in which integrals are allocated (cf. Figure 4) is preceded by one or more screening steps, so that the integral will be allocated only if it is greater than a certain threshold. In our notation, M, R, N, and S are shells of atomic orbitals, ν and σ represent atomic orbitals, i and j denote localized occupied molecular orbitals, and r and s are projected atomic orbitals. The following quantities are employed in the screening procedure:

$$T_{NS}^{\text{Schwarz}} = \text{Max}_{\nu,\sigma}|(\nu\sigma|\nu\sigma)|^{0.5},\ \nu \in N, \sigma \in S \quad \text{(A-1)}$$

$$T_{N}^{\text{Schwarz}} = \text{Max}_{S}(T_{NS}^{\text{Schwarz}}),\ \text{all } S \quad \text{(A-2)}$$

$$P_{Ns}^{\text{max}} = \text{Max}_{\nu}(|P_{\nu s}|),\ \nu \in N \quad \text{(A-3)}$$

$$P_{N}^{\text{max}} = \text{Max}_{s}(P_{Ns}^{\text{max}}),\ \text{all } s \quad \text{(A-4)}$$

$$(PP)_{MN}^{\text{max}} = \text{Max}_{r,s}(P_{Mr}^{\text{max}}\ P_{Ns}^{\text{max}}),\ rs \in [ij]\ \text{(all valid } ij \text{ pairs)} \quad \text{(A-5)}$$

$$L_{Nj}^{\text{max}} = \text{Max}_{\nu}(|L_{\nu j}|),\ \nu \in N \quad \text{(A-6)}$$

$$L_{N}^{\text{max}}(j) = \text{Max}_{i}(L_{Ni}^{\text{max}}),\ \text{all } i \mid ij \text{ valid pair} \quad \text{(A-7)}$$

$$\text{L}_{N}^{\text{max}} = \text{Max}_{j}(L_{Nj}^{\text{max}}),\ \text{all } j \quad \text{(A-8)}$$

$$L^{\text{max}} = \text{Max}_{N}(L_{N}^{\text{max}}),\ \text{all } N \quad \text{(A-9)}$$

$$(LL)_{MN}^{\text{max}} = \text{Max}_{i,j}(\text{L}_{Mi}^{\text{max}}\ \text{L}_{Nj}^{\text{max}}),\ \text{all } i,j \mid ij \text{ valid pair} \quad \text{(A-10)}$$

$$(LL)_{M}^{\text{max}} = \text{Max}_{N}((LL)_{MN}^{\text{max}}),\ \text{all } N \quad \text{(A-11)}$$

$$(LL)^{\text{max}} = \text{Max}_{M}((LL)_{M}^{\text{max}}),\ \text{all } M \quad \text{(A-12)}$$

Using the loop structure detailed in Figure 4, the screening protocol is implemented as follows. Initially, after entering the M, N loops, a preliminary screening is performed

$$T_{M}^{\text{Schwarz}} * T_{N}^{\text{Schwarz}} * (PP)_{MN}^{\text{max}} * (LL)^{\text{max}} \geq \epsilon \quad \text{(A-13)}$$

A loop over shells R is then entered, using an ordering of the R shells determined by sorting T_{MR}^{Schwarz} (for the current M and all R) in descending order. The screening

$$T_{N}^{\text{Schwarz}} * T_{MR}^{\text{Schwarz}} * (PP)_{MN}^{\text{max}} * (LL)^{\text{max}} \geq \epsilon \quad \text{(A-14)}$$

is then carried out, and, if not passed, the algorithm breaks out of the R loop, skipping all remaining R values. Otherwise, for the current R, another screening is performed

$$T_N^{\text{Schwarz}} * T_{MR}^{\text{Schwarz}} * (PP)_{MN}^{\max} * (LL)_R^{\max} \geq \epsilon \quad \text{(A-15)}$$

and, if passed, the S loop is entered. The S shells are sorted according to descending size of T_{NS}^{Schwarz}. Inside the S loop, a screening

$$T_{NS}^{\text{Schwarz}} * T_{MR}^{\text{Schwarz}} * (PP)_{MN}^{\max} * (LL)_R^{\max} \geq \epsilon \quad \text{(A-16)}$$

is first done to test whether to break out of the S loop. If passed, another screening step

$$T_{NS}^{\text{Schwarz}} * T_{MR}^{\text{Schwarz}} * (PP)_{MN}^{\max} * (LL)_{RS}^{\max} \geq \epsilon \quad \text{(A-17)}$$

determines whether to allocate the AO integral block $(MR|NS)$.

Allocation of the $(MR|Nj)$ integrals entails a loop over j indices that have been ordered according to decreasing $L_{Sj}^{\max}$. The screening step

$$|(MR|NS)|^{\max} * (PP)_{MN}^{\max} * L_{Sj}^{\max} * L_R{}^{\max} \geq \epsilon \quad \text{(A-18)}$$

determines whether to break out of the j loop, and, if continuing within the j loop, another screening

$$|(MR|NS)|^{\max} * L_{Sj}^{\max} * (PP)_{MN}^{\max} * L_R^{\max}(j) \geq \epsilon \quad \text{(A-19)}$$

determines whether to allocate the $(MR|Nj)$ block. The integral bound $|(MR|NS)|^{\max}$ represents the maximum absolute value of the AO integral for all atomic orbitals in the shell quartet M, R, N, S. In the second, third, and fourth quarter transformations, the following screening steps are performed before allocating the $(Mi|Nj)$, $(Mi|sj)$, and $(ri|sj)$ integrals, respectively:

$$|(MR|Nj)|^{\max} * L_{Ri}^{\max} * (PP)_{MN}^{\max} \geq \epsilon \quad \text{(A-20)}$$

$$|(Mi|Nj)|^{\max} * P_{Ns}^{\max} * P_M^{\max} \geq \epsilon \quad \text{(A-21)}$$

$$|(Mi|sj)|^{\max} * P_{Mr}^{\max} \geq \epsilon \quad \text{(A-22)}$$

The bound $|(MR|Nj)|^{\max}$ represents the maximum absolute value of a quarter-transformed integral for a fixed occupied index j and for all atomic orbitals in the shells M, R, and N, and the remaining integral bounds are defined analogously. In the present work, a value of 10^{-8} was used for all ϵ thresholds in the screening procedure.

Appendix B. Parallel Sparse Data Representation

Efficient massively parallel implementation of local correlation methods requires a means of handling distributed sparse matrices and multidimensional arrays. The sparsity of matrices and multidimensional arrays, e.g., the overlap matrix, the double substitution amplitudes T_{ij}^{rs}, and the two-electron integrals $(ri|sj)$, must be exploited to achieve linear scaling, and at least some of the multidimensional arrays, including the partially transformed two-electron integrals, must be distributed across nodes to avoid memory and communication bottlenecks. We have therefore implemented a parallel sparse data representation and an associated set of generalized contraction routines to perform contractions of distributed sparse matrices and multidimensional arrays. The most important features of these routines are discussed here, and a more thorough discussion of the implementation can be found elsewhere.[28]

Our parallel sparse data representation has been designed to permit rapid prototyping, ease of use, and good performance. The primary type is the C++ generic class Array. The primary template parameter for the Array class is an integer specifying the dimensionality the array, and of the Array relies on a variety of auxiliary classes that describe blocks, distribution schemes, index ranges, etc. An Array stores the block descriptors for contributing blocks and the data for each block. The block sizes for virtual indices are the number of basis functions on the atom associated with the block, and for occupied indices the block size is one. This permits the majority of blocks to be large enough to make their manipulation efficient while also enabling sufficiently fine-grained parallel distribution of the work. An optimized dense matrix multiplication routine is used for multiplication of the individual blocks. The Array class contains functions for manipulating the blocks, including copying arrays, summation of arrays, performing a binary product (contraction) of arrays to produce a third array, and doing a contraction that results in a scalar. Parallelization is supported by member functions that convert replicated to distributed arrays, distributed to replicated arrays, and distributed arrays to a different distribution scheme. Helper classes are also employed to permit the use of the implicit summation convention in the C++ source code, greatly increasing the readability of the program.

Upon declaration of an Array, its dimensionality must be specified, and the array must be initialized to get the proper ranges for its indices in each dimension. Additionally, the initialization of the array may specify a bound such that only blocks containing elements greater than the bound will be stored. For instance, the four-dimensional array Ints holding the transformed integrals $(ri|sj)$ may be declared and initialized as

```
Array<4> Ints;

Ints.init(vir,occ,vir,occ,bound);
```

An array may be allocated by allocating each of its blocks individually or by allocating the entire array at one time. For the Ints array, a single block may be allocated as

```
Ints.allocate_block(blockinfo);
```

where blockinfo specifies which block is to be allocated. Alternatively, an entire array may be allocated in one setting, e.g., from another, already allocated, array using the |= allocation operator

```
Array<4> T;

T.init(vir,occ,vir,occ,bound);

T(r,i,s,j) |= Ints(r,i,s,j);
```

As an example of application of our contraction routines, consider computation of the fully transformed two-electron integrals $(ri|sj)$, which are kept in the Ints array. These

integrals are obtained in the fourth quarter transformation, and this transformation can be performed in parallel in a single line of code in the program

$$\text{Ints(r,i,s,j)} = \text{q3(mu,i,s,j)*P(mu,r);}$$

The q3 and Ints arrays are both distributed across nodes, and the contraction will be done locally on each node using only the local elements of the arrays. Thus, once the data distribution has been determined for the arrays involved, the distribution of work is automatically achieved. Finally, to illustrate a contraction of arrays producing a scalar, consider the computation of the correlation energy (cf. eq 4) from the integrals and the double substitution amplitudes, stored in the arrys Ints and T, respectively. The Ints array (and T, if desired) is distributed, and the parallelization is handled by the contraction routine, which, again, works only on the locally stored elements. The parallel computation of the correlation energy, then, is handled by the following three lines of code

$$e_{corr} = 2 * \text{Ints(r,i,s,j)} * \text{T(r,i,s,j);}$$

$$e_{corr} - = \text{Ints(r,i,s,j)} * \text{T(s,i,r,j);}$$

$$\text{msg} \rightarrow \text{sum}(e_{corr});$$

The summation after the two contractions is required because each node computes only a partial contribution to the correlation energy.

References

(1) Pulay, P.; Saebø, S. *Theor. Chim. Acta* **1986**, *69*, 357−368.

(2) Saebø, S.; Pulay, P. *J. Chem. Phys.* **1987**, *86*, 914−922.

(3) Saebø, S.; Pulay, P. *Ann. Rev. Phys. Chem.* **1993**, *44*, 213−236.

(4) Saebo, S.; Pulay, P. *J. Chem. Phys.* **1988**, *88*, 1884−1890.

(5) Rauhut, G.; Pulay, P.; Werner, H.-J. *J. Comput. Chem.* **1998**, *19*, 1241−1254.

(6) Saebø, S.; Pulay, P. *J. Chem. Phys.* **2001**, *115*, 3975−3983.

(7) Ayala, P. Y.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 3660−3671.

(8) Häser, M.; Almlöf, J. *J. Chem. Phys.* **1992**, *96*, 489−494.

(9) Murphy, R. B.; Beachy, M. D.; Friesner, R. A.; Ringnalda, M. N. *J. Chem. Phys.* **1995**, *103*, 1481−1490.

(10) Friesner, R. A.; Murphy, R. B.; Beachy, M. D.; Ringnalda, M. N.; Pollard, W. T.; Dunietz, B. D.; Cao, Y. *J. Phys. Chem. A* **1999**, *103*, 1913−1928.

(11) Maslen, P. E.; Head-Gordon, M. *Chem. Phys. Lett.* **1998**, *283*, 102−108.

(12) Maslen, P. E.; Head-Gordon, M. *J. Chem. Phys.* **1998**, *109*, 7093−7099.

(13) Lee, M. S.; Maslen, P. E.; Head-Gordon, M. *J. Chem. Phys.* **2000**, *112*, 3592−3601.

(14) Schütz, M.; Hetzer, G.; Werner, H.-J. *J. Chem. Phys.* **1999**, *111*, 5691−5705.

(15) Hetzer, G.; Schütz, M.; Stoll, H.; Werner, H.-J. *J. Chem. Phys.* **2000**, *113*, 9443−9455.

(16) Bernholdt, D. E.; Harrison, R. J. *J. Chem. Phys.* **1995**, *102*, 9582−9589.

(17) Beachy, M. D.; Chasman, D.; Friesner, R. A.; Murphy, R. B. *J. Comput. Chem.* **1998**, *19*, 1030−1038.

(18) Pipek, J.; Mezey, P. G. *J. Chem. Phys.* **1989**, *90*, 4916−4926.

(19) Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1996**, *104*, 6286−6297.

(20) Boughton, J. W.; Pulay, P. *J. Comput. Chem.* **1993**, *14*, 736−740.

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

(22) Pulay, P. *J. Comput. Chem.* **1982**, *3*, 556−560.

(23) Janssen, C. L.; Nielsen, I. B.; Leininger, M. L.; Valeev, E. F.; Seidl, E. T. *The Massively Parallel Quantum Chemistry Program (MPQC), Version 2.3.0-alpha*; Sandia National Laboratories: Livermore, CA, U.S.A., 2005. http://www.mpqc.org (accessed October 4, 2006).

(24) Wong, A. T.; Harrison, R. J.; Rendell, A. P. *Theor. Chim. Acta* **1996**, *93*, 317−331.

(25) Nielsen, I. M. B. *Chem. Phys. Lett.* **1996**, *255*, 210−216.

(26) Janssen, C. L.; Nielsen, I. M. B.; Colvin, M. E. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: Chichester, 1998; pp 1990−2000.

(27) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796−6806.

(28) Nielsen, I.; Janssen, C.; Leininger, M. *Scalable Fault Tolerant Algorithms for Linear-Scaling Coupled-Cluster Electronic Structure Methods*; Sandia Report, SAND2004-5462; Sandia National Laboratories: Livermore, CA, U.S.A., 2004.

CT600188K

JCTC Journal of Chemical Theory and Computation

Spin-Component Scaling Methods for Weak and Stacking Interactions

J. Grant Hill and James A. Platts*

School of Chemistry, Cardiff University, Park Place, Cardiff CF10 3AT, U.K.

Received August 21, 2006

Abstract: New scaling parameters are presented for use in the spin-component scaled (SCS) variant of density fitted local second-order Møller−Plesset perturbation theory (DF-LMP2) that have been optimized for use in evaluating the interaction energy between nucleic acid base pairs. The optimal set of parameters completely neglects the contribution from antiparallel-spin electron pairs to the MP2 energy while scaling the parallel contribution by 1.76. These spin-component scaled for nucleobases (SCSN) parameters are obtained by minimizing, with respect to SCS parameters, the rms interaction energy error relative to the best available literature values, over a set of ten stacked nucleic acid base pairs. The applicability of this scaling to a wide variety of noncovalent interactions is verified through evaluation of a larger set of model complexes, including those dominated by dispersion and electrostatics.

1. Introduction

Intermolecular interactions between nucleic acid base pairs play a crucial role in the structure of nucleic acids such as DNA and RNA. In particular, stacking interactions between base pairs is a major driving force in processes such as molecular recognition and the folding of nucleic acids, among many others.[1−21] Computational modeling of such interactions is an attractive prospect, both for an increased understanding of structure and folding processes themselves and for possible design of new drugs with nucleic acids as their target. However, carrying out such calculations is far from trivial: the size of biologically relevant sections of DNA rules out the use of ab initio methods for entire systems and limits such approaches to smaller model systems such as base pairs of specific interest. Hybrid QM/MM approaches are therefore attractive, in which only part of the overall system is treated with quantum chemical methods.

A further problem arises from the computational methods required to describe London dispersion forces that dominate stacking interactions. The prototypical model for π-stacking interactions is the benzene dimer, and as such it has been investigated exhaustively within the limits of currently available methodology and computer hardware (see refs 22−26 and references therein). This work makes it clear that correlated calculations with basis sets of at least triple-ζ quality are required to attain even a qualitatively accurate picture. Calculation of accurate interaction energies for stacked nucleic acid bases requires a similar theoretical treatment. Second-order Møller−Plesset perturbation theory (MP2) is the most commonly used and computationally tractable method of describing dynamic electron correlation, which is ultimately the source of dispersion forces. Unfortunately, this method is known to consistently overestimate binding energies in both the benzene dimer and stacked nucleic acid base pairs. Coupled cluster theory, particularly CCSD(T), is accepted as the standard for evaluating noncovalent interaction energies, providing highly accurate binding energies when used with basis sets of at least aug-cc-pVTZ quality. Such calculations rapidly become prohibitively computationally demanding, especially when the additional time necessary to compute a counterpoise correction[27] (CP) for basis set superposition error (BSSE) is included. Therefore, the current method of choice for stacking interactions extrapolates to the complete basis set limit from a series of MP2 interaction energies and then adds a CCSD(T) correction using a smaller basis set,[20] an approach termed "CBS(T)" by some authors.

Recently, Grimme[28] proposed the spin-component scaling (SCS) MP2 method, which scales the correlation energy due

* Corresponding author e-mail: platts@cf.ac.uk.

10.1021/ct6002737 CCC: $37.00
Published on Web 11/29/2006

to parallel- and antiparallel-spin electron pairs, in order to overcome some of the known problems with canonical MP2. Previous work with SCS-MP2 shows impressive performance in calculating energies and geometries of small model systems[28−31] as well as larger host−guest systems, heats of formations, and reaction barriers.[32−36] We recently tested this approach, coupled with density-fitted, local MP2 (DF-LMP2)[37] for the benzene dimer, obtaining highly accurate binding energies and optimized geometries.[26] Local correlation methods[38−40] are virtually free of BSSE by construction, meaning they do not require CP correction. They also offer the potential for linear scaling with system size, enabling the study of much larger systems than with canonical MP2. Density fitting approximations (also known as resolution of the identity, RI) of the electron repulsion integrals yield a further order of magnitude reduction in computation time. Our DF-SCS-LMP2 calculations reproduced the CCSD(T) potential energy curves[23] for three orientations of the benzene dimer to within 0.1−0.2 kcal mol^{-1}, in a small fraction of the time required for CCSD(T) calculation.

In an effort to simplify and further accelerate the SCS-MP2 method, Head-Gordon has proposed a scaled opposite-spin (SOS) method, in which the parallel-spin contribution to the MP2 energy is neglected and the antiparallel contribution is scaled by a factor of 1.3.[41] While this allows use of a faster SOS-MP2 algorithm, it has yet to be tested on all but the smallest of systems. The same group has also developed an operator to adjust scaling depending on the distance between electrons, termed modified opposite-spin (MOS), to improve upon the SOS-MP2 description of long-range interactions.[42]

In this paper, we calculate DF-LMP2, DF-SCS-LMP2, and DF-SOS-LMP2 interaction energies for a set of stacked nucleic acid base pairs and compare these energies with the best available estimates from the literature.[20] With a view to obtaining accurate binding energies for larger and larger models of DNA, we propose a new a set of SCS scaling parameters, which we term spin-component scaled for nucleobases (SCSN), that accurately reproduce literature values for the stacked nucleic acid base pairs investigated. As a test of this DF-SCSN-LMP2 method, we compare the results to those for standard DF-LMP2 as well as the original SCS and SOS methods for a larger set of molecules, namely that denoted "S22" by Hobza et al.[21] This set includes complexes bound solely by dispersion forces, by electrostatic forces alone, and by combinations of these. As such, it provides a stringent test of the SCSN parameters proposed.

2. Computational Procedure

All ab initio calculations were carried out with the MOLPRO package of programs,[43] locally modified to separately output parallel- and antiparallel-spin electron pair energy contributions. Geometries of ten stacked nucleic acid base pairs and the S22 set were taken from refs 20 and 21, respectively. These geometries were obtained using the rigid monomer approximation. DF-SCS-LMP2 results employed the default scaling factors of 6/5 for antiparallel spins and 1/3 for parallel spins, while for DF-SOS-LMP2 data the parallel-spin scaling factor is set to zero and that of the antiparallel spins to 1.3.

Table 1. Interaction Energies (kcal mol^{-1}) of Stacked Nucleic Acid Base Pairs[a]

complex	DF-LMP2	SCS	SOS	SCSN	best estimate[b]
A···A	−10.87	−7.69	−6.10	−8.83	−8.5
A···C	−11.63	−8.79	−7.38	−10.29	−10.2
A···U	−11.05	−8.23	−6.82	−9.71	−9.8
C···C	−10.47	−8.16	−7.00	−9.78	−10.0
C···U	−10.54	−8.25	−7.10	−10.12	−10.4
G···A	−13.44	−10.17	−8.53	−11.78	−11.4
G···C	−11.42	−8.87	−7.59	−10.23	−10.6
G···G	−14.12	−10.86	−9.23	−12.64	−12.7
G···U	−12.76	−9.97	−8.57	−11.92	−12.1
U···U	−8.07	−5.97	−4.92	−7.47	−7.5
RMSE	1.31	1.67	3.02	0.24	
MD	1.12	−1.62	−3.00	−0.04	
MAD	1.12	1.62	3.00	0.20	

[a] SCS, SOS, and SCSN empirical scalings were all applied to the DF-LMP2 energy contributions from parallel- and antiparallel-spin pairs of electrons. [b] CBS(T) values from ref 20.

Initial Hartree Fock calculations utilized the DF−HF method[44] (also referred to as RI−HF[45] by some authors), but neither this nor subsequent DF-LMP2 calculations used local fitting domains.[37,44] All calculations employed the aug-cc-pVTZ augmented correlation consistent basis set of Dunning et al.[46] as the AO basis, with the aug-cc-pVTZ MP2-fitting[47] and cc-pVTZ JK-fitting[45] auxiliary basis sets of Weigend et al. applied in the DF-LMP2 and DF−HF calculations, respectively.

Orbital localization was performed with the Pipek−Mezey method.[48] The two most diffuse basis functions of each angular momentum type were the cause of poor localization in some cases. Their contribution was therefore removed from the localization criterion by setting the corresponding rows and columns of the overlap matrix to zero, and localized orbitals were obtained using a Newton−Raphson algorithm.

LMP2 localizes the occupied orbitals and projects out the occupied space from the virtual orbitals to create a set of projected atomic orbitals (PAOs).[38−40] Excitations from the dominant part of the occupied orbitals are then restricted to subspaces (domains) of the spatially close PAOs. Domain selection was carried out using a completeness criterion of 0.985, with the procedure described by Boughton and Pulay.[49] To ensure that only domains resembling bonding interactions were generated the Löwdin charge for a hydrogen atom to be included in the domain was increased to 0.15. Invariance to unitary transformations of the π-orbitals was achieved through merging the domains of those orbitals. Domains were all determined at large intermolecular separation and kept fixed in the calculation of the interaction energy.

3. Results and Discussion

Interaction energies for stacked nucleic acid base pairs are shown in Table 1 as evaluated by DF-LMP2, DF-SCS-LMP2, and DF-SOS-LMP2 approaches, along with CBS(T) literature values. It is clear that DF-LMP2 itself consistently overestimates the interaction energy, leading to a root-mean-square error (RMSE) of 1.3 kcal mol^{-1}. This shortcoming of MP2 and related methods has been documented before

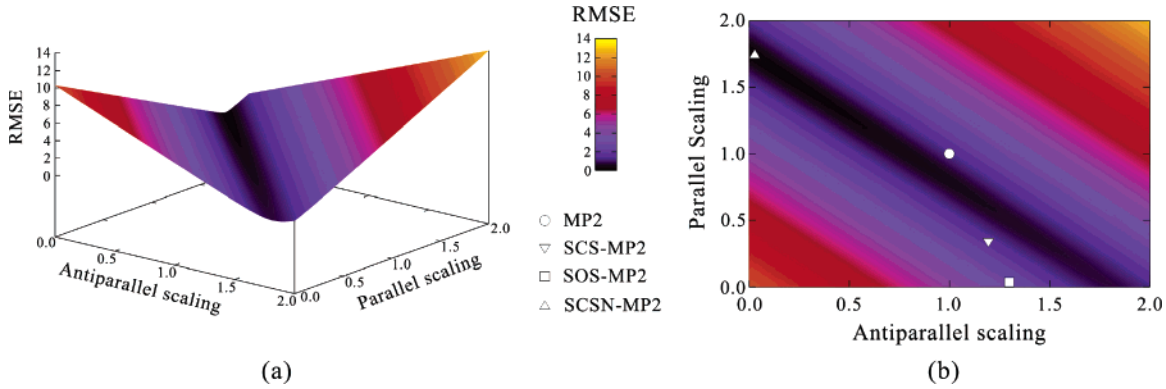


Figure 1. Plots of the root-mean-square error (RMSE) of the interaction energies of ten stacked nucleic acid base pairs compared to literature best estimates generated by separately scaling the parallel- and antiparallel-spin electron pairs contribution to the MP2 energy: (a) shows a three-dimensional surface of the RMSE values, while (b) is a two-dimensional projection of the same surface, with symbols indicating the position of different variants of MP2.

and therefore comes as no surprise. However, there is substantial variability in the discrepancy of DF-LMP2 from literature values, the largest being 2.4 kcal mol^{-1}, for A...A but the smallest just 0.1 kcal mol^{-1}, for C...U.

The default SCS parameters reduce binding energy in all cases and in doing so overcompensate for the deficiencies of DF-LMP2. This is most evident in those systems for which DF-LMP2 performs well, such as C...C and C...U, whereas DF-SCS-LMP2 yields reasonable interaction energies for A...A and G...A, for which DF-LMP2 performs poorly. Thus it seems that the default parameters reported by Grimme are not ideal for this set. Still worse are the results from SOS, which performs poorly compared with both DF-LMP2 and its SCS variant. The RMSE of this method is almost twice that of those approaches, and just as with SCS it overcompensates to produce interaction energies that are smaller (in absolute terms) than literature data. Since overcompensation occurs in both cases when the antiparallel contribution is increased and the parallel decreased, this may suggest that reducing the singlet scaling and increasing the triplet may be beneficial in this case. In any case, it is clear that neither SCS nor SOS scaling factors give an overall benefit over standard DF-LMP2 for these complexes.

To assess the effect of SCS parameters on the binding energy of stacked nucleic acid base pairs, the binding energy of the ten base pairs was calculated using a range of parallel and antiparallel scaling factors, and the RMSE was compared with the best available literature values evaluated. Figure 1 displays the RMSE for both parameters in the range of 0.00−2.00. This displays a valley of points with RMSE values of less than 1.0, such that any of the scalings corresponding to those points would be a reasonable choice for nucleic acid base pairs. However, the valley is rather flat, making it difficult to visually distinguish any local minima.

To obtain a more rigorous set of parameters, code was written to locate the optimum values of scaling parameters, using the analytical gradient of the RMSE expression and the BFGS algorithm, as described in ref. 50. This resulted in optimal values of −0.53 for antiparallel spins and 2.28 for parallel, which give a very low RMSE value of 0.17 kcal mol^{-1}. While this represents impressive accuracy, it is in effect a subtraction of correlation energy due to antiparallel spins from the binding energy. This does not seem to be physically realistic and hence may not be transferable to other systems outside this training set. We therefore constrained both parameters to be ≥0.0: this yielded a new set of 1.76 for parallel-spin electron pairs and 0.0 (i.e., no contribution) of the antiparallel component. The RMSE value for this pair of parameters was 0.24 kcal mol^{-1}, i.e. only marginally higher than the global optimum values and much lower than any of the methods tested above. Interaction energies of the ten base pairs are reported in Table 1 as DF-SCSN-LMP2. Encouragingly, unlike SCS or SOS parameters these new SCSN values reproduce accurate interaction energies for all ten base pairs, regardless of how close the original DF-LMP2 values were to the literature value.

To test the SCSN parameters, we have applied the same methodology to a larger and more varied set of complexes. To do this, we employ the set of 22 complexes denoted "S22" by Hobza et al.,[21] which includes stacked, hydrogen-bonded, and "mixed" complexes. Table 2 reports interaction energies for the S22 set obtained with DF-LMP2, DF-SCS-LMP2, DF-SOS-LMP2 and DF-SCSN-LMP2, along with CBS(T) literature values from ref 21. The first seven complexes (from the ammonia dimer to the Watson−Crick (WC) structure of the adenine·thymine complex) are all primarily bound by hydrogen bonding. The next eight systems (the methane dimer to the adenine·thymine stack) are all examples of London dispersion interaction, and finally the remaining seven systems have a mixture of both hydrogen bonding and dispersion binding them together.

The RMSE values again illustrate that on average DF-LMP2 produces binding energies that are on average 1 kcal mol^{-1} from literature values. Thus, it performs rather better than the default SCS or SOS scaling. In all cases SCS reduces binding energy, yet DF-LMP2 already underestimates binding energy in several complexes. Only in cases where canonical MP2 or DF-LMP2 substantially overestimates binding, such as the benzene and pyrazine dimers and the indole·benzene stacked complex, does SCS scaling produce more accurate results. In common with Table 1, SOS scaling produces binding energies that are even less negative than SCS and hence are even further from the best estimates. Although the SCSN parameters were optimized only for the

Table 2. Comparison of Interaction Energies (kcal mol^{-1}) for the S22 Set of Model Complexes[a]

complex	DF-LMP2	SCS	SOS	SCSN	best estimate[b]
ammonia dimer	−2.81	−2.40	−2.19	−2.84	−3.17
water dimer	−4.62	−4.16	−3.94	−5.01	−5.02
formic acid dimer	−17.30	−15.67	−14.86	−19.53	−18.61
formamide dimer	−14.81	−13.45	−12.77	−16.08	−15.96
uracil dimer (HBond)	−19.22	−17.40	−16.49	−21.33	−20.65
2-pyridoxine·2-aminopyridine	−15.72	−13.74	−12.75	−16.75	−16.71
adenine·thymine (HBond)	−15.35	−13.55	−12.65	−16.04	−16.37
methane dimer	−0.45	−0.27	−0.19	−0.32	−0.53
ethene dimer	−1.33	−0.81	−0.54	−1.14	−1.51
benzene·methane	−1.71	−1.08	−0.76	−1.35	−1.50
benzene dimer (PD)	−4.54	−2.51	−1.50	−2.91	−2.73
pyrazine dimer	−6.32	−4.15	−3.06	−4.64	−4.42
uracil dimer (stacked)	−10.26	−7.67	−6.38	−9.66	−10.12
indole·benzene (stacked)	−7.58	−4.52	−3.00	−5.30	−5.22
adenine·thymine (stacked)	−13.71	−9.78	−7.81	−11.90	−12.23
ethene·ethyne	−1.26	−0.88	−0.69	−1.60	−1.53
benzene·water	−3.33	−2.71	−2.41	−3.23	−3.28
benzene·ammonia	−2.48	−1.85	−1.54	−2.23	−2.35
benzene·hydrogen cyanide	−5.00	−4.11	−3.66	−5.16	−4.46
benzene dimer (T-shape)	−3.45	−2.33	−1.77	−2.89	−2.74
indole·benzene (T-shape)	−6.70	−5.08	−4.27	−6.01	−5.73
phenol dimer	−7.27	−5.94	−5.28	−6.90	−7.05
RMSE	1.05	1.65	2.39	0.36	
MD	0.15	−1.26	−1.97	0.04	
MAD	0.81	1.26	1.97	0.27	

[a] SCS, SOS, and SCSN empirical scalings were all applied to the DF-LMP2 energy contributions from parallel- and antiparallel-spin pairs of electrons. [b] CBS(T) values from ref 21.

stacked nucleic acid base pairs of Table 1, they also produce excellent results for the entire S22 set. Once again, SCSN performs well both in situations where standard DF-LMP2 overestimates the dispersion energy, such as the benzene dimer, as well as in cases where the unscaled approach produces reasonable values or even underestimates the binding energy. There are few cases where SCSN produces less accurate interaction energies than DF-LMP2: these include the methane dimer, the stacked uracil dimer, and the benzene·hydrogen cyanide complex. Even in these cases, SCSN results are still within 0.7 kcal mol^{-1} of CBS(T) data. As expected, the mean deviation (MD) indicates that both SCS and SOS consistently underestimate the binding energies in the S22 set, while the canonical and SCSN results are much more evenly spread on either side of the literature values. The mean absolute deviation (MAD) correlates with the trend seen in the RMSE values, and SCSN energies typically deviate little from the best available.

Further support for our choice of scaling factors comes from the observation that an RMSE of 0.802 kcal mol^{-1} results from the values obtained from gradient-based optimization (i.e., −0.53 for antiparallel and 2.28 for parallel: individual interaction energies are reported in the Supporting Information). While this is still an improvement over canonical DF-LMP2, it indicates that these parameters are indeed not transferable to the larger S22 set and therefore are not as applicable to a wide variety of systems as SCSN.

In our previous investigation of three isomers of the benzene dimer, DF-SCS-LMP2/aug-cc-pVTZ performed exceedingly well when compared with CCSD(T)/aug-cc-pVQZ results, with errors within 0.2 kcal mol^{-1} along entire potential energy scans of intermonomer separation. Applying SCSN scaling parameters to these potential energy curves (see the Supporting Information for plots of parallel-displaced and T-shaped benzene dimers) results in a slight reduction of accuracy when compared to the default SCS values. However, performance is still acceptable, with values within 0.3−0.4 kcal mol^{-1} of those from CCSD(T).

Default SCS scaling values only appear to work well when MP2 strongly overestimates the dispersion contribution to the interaction energy: this is difficult to predict a priori. Their usefulness for binding energies of truncated sections of DNA is therefore questionable, and this approach should be employed with caution, despite the encouraging results for some model systems.[26,28,31] SOS amplifies this problem, and, in light of the poor performance for both sets of data, it is hard to recommend its use in the evaluation of nucleic acid binding energies. However, Figure 1 suggests that increasing the antiparallel scaling to approximately 1.7 would considerably improve binding energies of stacked base pairs. DF-SCSN-LMP2 provides excellent interaction energies for systems bound by hydrogen bonds, dispersion interaction, and a mixture of the two. The combination of its performance for the benzene dimer and a wider range of noncovalent interactions displayed in Table 2 for the S22 set of model complexes suggests that DF-SCSN-LMP2 should be capable of providing high quality interaction energies in a wide range of systems. This is therefore a promising method for investigating the structure of and interactions involved in small sections of nucleic acids and proteins and their interactions with small molecules and drugs.

As shown in Figure 1, very many combinations of parallel and antiparallel contributions give rise to low errors for the ten stacked nucleobases and hence to a valley of acceptable solutions. As such, the specific scaling parameters we recommend should probably not be overinterpreted for physical meaning. However, it does seem evident that in all acceptable combinations the sum of parallel and antiparallel contributions is less than 2.0, i.e., one must reduce the binding energy due to electron correlation from that given by canonical MP2 to obtain RMSE < 1.0. As these SCSN scaling parameters have been specifically optimized for obtaining accurate binding energies for noncovalent interactions, and more specifically such interactions between nucleic acid base pairs, they have not been tested for general quantum chemical applicability. This is in contrast to both SCS- and SOS-MP2 which were designed to be more general methods. As such, we cannot currently recommend that SCSN scaling is utilized in applications other than the evaluation of noncovalent interactions.

The SCSN parameters were optimized based on parallel and antiparallel contributions evaluated with the aug-cc-pVTZ basis set. While the use of larger basis sets has not been extensively tested, it may be possible that doing so will produce binding energies further from the literature best estimates. Smaller basis sets should be employed with caution when the reference energy is obtained with DF−HF as the SCF BSSE is unlikely to be minimized in such situations. Existing basis set extrapolation schemes such as that of Helgaker et al.[51,52] can be generally expected to increase the binding energy of noncovalently bound systems, and hence at the CBS limit the SCS- and SOS-MP2 interaction energies should be improved while the typically overbinding canonical MP2 will be worse. As SCSN produces values very close to the literature values with the aug-cc-pVTZ basis set, an extrapolation will either improve or worsen a result depending on which side of the literature value the original value lies. It should also be noted that as SCSN completely neglects the contribution to the correlation energy from antiparallel-spin electron pairs the existing extrapolation schemes will be less optimal than in the case of canonical MP2.

As in the case of SOS-MP2, it may be possible to derive a more efficient algorithm for SCSN-MP2, since the contributions from antiparallel-spin electron pairs are neglected. However, as the SCSN scaling is readily applied to the already highly efficient DF-LMP2 results at no additional computational cost, this seems unnecessary at present. Additionally, DF-LMP2 provides qualitatively correct descriptions of noncovalent interactions, such that comparison of DF-SCSN-LMP2 with canonical DF-LMP2 binding energies may act as a form of "sanity check" highlighting potentially problematic systems.

4. Conclusions

The combination of the DF-LMP2 approach with empirical SCS scaling that appeared so promising for the benzene dimer has been shown to provide, on average, worse quality interaction energies for stacked nucleic acid base pairs and the S22 set of noncovalent interactions than canonical DF-LMP2. SOS scaling, which neglects the parallel-spin electron pair contribution, provides binding energies that are further still from the best estimate literature values. We hence propose a set of alternative SCS scaling parameters where the antiparallel contribution to the energy is eliminated and the parallel scaling is set to 1.76. These values were obtained by calculating the rms error from literature values for 10 stacked base pairs over a wide range of parameters. Gradient-based optimization indicated a physically unrealistic set of parameters, which despite giving a very small RMSE are not transferable to a more varied set of complexes. We term these new parameters "spin-component scaled for nucleobases" (SCSN) MP2 or DF-SCSN-LMP2.

Validation of this approach used Hobza's S22 set of model complexes as well as potential energy scans of the benzene dimer. This confirms that the SCSN method provides excellent quality results for a range of noncovalent interactions including hydrogen bonding and dispersion interactions. SCSN again outperforms the other scaling parameters on test, in addition to canonical DF-LMP2. These interactions are typical of those present between DNA bases, such that the SCSN scaling parameters are promising for study of such systems. The SCSN variant of DF-LMP2 seems especially promising, as the local nature of the electron correlation ensures that BSSE is minimized and potentially expensive counterpoise corrections can be avoided. The relatively low cost of the DF-LMP2 approach also means that reasonably large basis sets can be employed. When SCSN scaling is applied, the current evidence indicates that the result can be expected to be within 0.5 kcal mol^{-1} of that from much more expensive extrapolation and correction schemes.

Supporting Information Available: Potential energy curves for both the parallel-displaced and T-shaped configurations of the benzene dimer, along with interaction energies for stacked nucleic acid base pairs obtained with SCS parameters of −0.53 and 2.28. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Watson, J. D.; Crick, F. H. C. *Nature* **1953**, *171*, 737−738.

(2) Bugg, C. E.; Thomas, J. M.; Sundralingam, M.; Rao, S. T. *Biopolymers* **1971**, *10*, 175−219.

(3) Askew, B.; Ballester, P.; Buhr, C.; Jeong, K. S.; Jones, S.; Parris, K.; Williams, K.; Rebek, J. *J. Am. Chem. Soc.* **1989**, *111*, 1082−1090.

(4) Hunter, C. A.; Sanders, J. K. M. *J. Am. Chem. Soc.* **1990**, *112*, 5525−5534.

(5) Kim, J. L.; Nikolov, D. B.; Burley, S. K. *Nature* **1993**, *365*, 520−527.

(6) Hunter, C. A. *Chem. Soc. Rev.* **1994**, *23*, 101−109.

(7) Rebek, J. *Chem. Soc. Rev.* **1996**, *25*, 255−264.

(8) Claessens, C. G.; Stoddart, J. F. *J. Phys. Org. Chem.* **1997**, *10*, 254−272.

(9) McGaughey, G. B.; Gagne, M.; Rappe, A. K. *J. Biol. Chem.* **1998**, *273*, 15458−15463.

(10) Hobza, P.; Šponer, *J. Chem. Rev.* **1999**, *99*, 3247−3276.

(11) Mathews, D. H.; Sabina, J.; Zuker, M.; Turner, D. H. *J. Mol. Biol.* **1999**, *288*, 911−940.

(12) Meyer, E. A.; Castellano, R. K.; Diederich, F. *Angew. Chem., Int. Ed.* **2003**, *42*, 1210−1250.

(13) Jurečka, P.; Hobza, P. *J. Am. Chem. Soc.* **2003**, *125*, 15608−15613.

(14) Piacenza, M.; Grimme, S. *J. Comput. Chem.* **2003**, *25*, 83−99.

(15) Šponer, J.; Jurečka, P.; Hobza, P. *J. Am. Chem. Soc.* **2004**, *126*, 10142−10151.

(16) Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2005**, *7*, 1624−1626.

(17) Dąbkowska, I.; Gonzalez, H. V.; Jurečka, P.; Hobza, P. *J. Phys. Chem. A* **2005**, *109*, 1131−1136.

(18) Hayley, T. P.; Graybill, E. R.; Cybulski, S. M. *J. Chem. Phys.* **2006**, *124*, 204301.

(19) Waller, M. P.; Robertazzi, A.; Platts, J. A.; Hibbs, D. E.; Williams, P. A. *J. Comput. Chem.* **2006**, *27*, 491−504.

(20) Šponer, J.; Jurečka, P.; Marchan, I.; Luque, F. J.; Orozco, M.; Hobza, P. *Chem. Eur. J.* **2006**, *12*, 2854−2865.

(21) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985−1993.

(22) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Phys. Chem.* **1996**, *100*, 18790−18794.

(23) Sinnokrot, M. S.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200−10207.

(24) Hesselmann, A.; Jensen, G.; Schütz, M. *J. Chem. Phys.* **2005**, *122*, 014103.

(25) Park, Y. C.; Lee, J. S. *J. Phys. Chem. A* **2006**, *110*, 5091−5095.

(26) Hill, J. G.; Platts, J. A.; Werner, H.-J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4072−4078.

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

(28) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095−9102.

(29) Gerenkamp, M.; Grimme, S. *Chem. Phys. Lett.* **2004**, *392*, 229−235.

(30) Grimme, S. *Chem. Eur. J.* **2004**, *10*, 3423−3429.

(31) Piacenza, M.; Grimme, S. *Chem. Phys. Chem.* **2005**, *6*, 1554−1558.

(32) Goumans, T. P. M.; Ehlers, A. W.; Lammertsma, K.; Wurthwein, E. U.; Grimme, S. *Chem. Eur. J.* **2004**, *10*, 6468−6475.

(33) Hyla-Kryspin, I.; Grimme, S. *Organometallics* **2004**, *23*, 5581−5592.

(34) Parac, M.; Etinski, M.; Peric, M.; Grimme, S. *J. Chem. Theory Comput.* **2005**, *1*, 1110−1118.

(35) Grimme, S. *J. Phys. Chem. A* **2005**, *109*, 3067−3077.

(36) Grimme, S.; Muck-Lichtenfeld, C.; Wurthwein, E. U.; Ehlers, A. W.; Goumans, T. P. M.; Lammertsma, K. *J. Phys. Chem. A* **2006**, *110*, 2583−2586.

(37) Werner, H.-J.; Manby, F. R.; Knowles, P. J. *J. Chem. Phys.* **2003**, *118*, 8149−8160.

(38) Pulay, P. *Chem. Phys. Lett.* **1983**, *100*, 151−154.

(39) Saebø, S.; Pulay, P. *Ann. Rev. Phys. Chem.* **1993**, *44*, 213−236.

(40) Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1996**, *104*, 6286−6297.

(41) Jung, Y.; Lochan, R. C.; Dutoi, A. D.; Head-Gordon, M. *J. Chem. Phys.* **2004**, *121*, 9793−9802.

(42) Lochan, R. C.; Jung, Y.; Head-Gordon, M. *J. Phys. Chem. A* **2005**, *109*, 7598−7605.

(43) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Schütz, M.; Celani, P.; Korona, T.; Manby, F. R.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. *MOLPRO, version 2006.1*; a package of ab initio programs. See: http://www.molpro.net (accessed Oct 5, 2006).

(44) Polly, R.; Werner, H.-J.; Manby, F. R.; Knowles, P. *J. Mol. Phys.* **2004**, *102*, 2311−2321.

(45) Weigend, F. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285−4291.

(46) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796−6806.

(47) Weigend, F.; Köhn, A.; Hättig, C. *J. Chem. Phys.* **2002**, *116*, 3175−3183.

(48) Pipek, J.; Mezey, P. G. *J. Chem. Phys.* **1989**, *90*, 4916−4926.

(49) Boughton, J. W.; Pulay, P. J. *J. Comput. Chem.* **1993**, *14*, 736−740.

(50) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in FORTRAN 77: The Art of Scientific Computing*; Cambridge University Press: 1992.

(51) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243−252.

(52) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Olsen, *J. Chem. Phys. Lett.* **1999**, *302*, 437−446.

CT6002737

JCTC Journal of Chemical Theory and Computation

Density-Functional and Coupled-Cluster Singles-and-Doubles Calculations of the Nuclear Shielding and Indirect Nuclear Spin−Spin Coupling Constants of *o*-Benzyne

Trygve Helgaker*,†
Department of Chemistry, University of Durham, South Road, Durham, DH1 3LE, United Kingdom

Ola B. Lutnæs
Department of Chemistry, University of Oslo, P.O.B. 1033 Blindern, N-0315 Oslo, Norway

Michał Jaszuński
Institute of Organic Chemistry, Polish Academy of Sciences, 01-224 Warszawa, Kasprzaka 44, Poland

Received July 17, 2006

Abstract: Density-functional theory (DFT) and coupled-cluster singles-and-doubles (CCSD) theory are applied to compute the nuclear magnetic resonance (NMR) shielding and indirect nuclear spin−spin coupling constants of *o*-benzyne, whose biradical nature makes it difficult to study both experimentally and theoretically. Because of near-equilibrium triplet instabilities that follow from its biradical character, the calculated DFT NMR properties of *o*-benzyne are unusually sensitive to details of the exchange-correlation functional. However, this sensitivity is greatly reduced if these properties are calculated at the equilibrium of the chosen functional. A strong correlation is demonstrated between the quality of the calculated indirect spin−spin coupling constants and the quality of the calculated lowest triplet excitation energy in *o*-benzyne. Orbital-unrelaxed coupled-cluster theory should be less affected by such instabilities, and the CCSD NMR properties were only calculated at the experimental equilibrium geometry. For the shielding constants, the results in best agreement with experimental results are obtained with CCSD theory and with the Keal−Tozer KT1 and KT2 functionals. For the triply bonded carbon atoms, these models yield an isotropic shielding of 1.3, −3.3, and −1.2 ppm, respectively, compared with the experimentally observed shielding of 3.7 ppm for incarcerated *o*-benzyne. For the indirect spin−spin coupling constants, the CCSD model and the Perdew−Burke−Ernzerhof functional both yield reliable results; for the most interesting spin−spin coupling constant, ^1J (C≡C), we obtain 210 and 209 Hz with these two models, respectively, somewhat above the recently reported experimental value of 177.9 ± 0.7 Hz for *o*-benzyne inside a molecular container, suggesting large incarceration effects.

1. Introduction

The NMR shielding tensor of the triply bonded carbon atom of the 1,2-^{13}C-substituted *o*-benzyne was first measured by Orendt et al.[1] in 1996, who also presented quantum-chemical calculations using Hartree−Fock (HF) theory, second-order Møller−Plesset theory, and density-functional theory (DFT) with the Becke−Lee−Yang−Parr (BLYP) functional. These calculations indicated that there are large electron-correlation contributions to the shielding of the triply bonded carbon atom—in particular, for the individual tensor components. Moreover, only the DFT results were found to agree well

* Corresponding author phone: +47 22855428; fax: +47 22855441; email: trygve.helgaker@kjemi.uio.no.
† Permanent address: Department of Chemistry, University of Oslo, P.O.B. 1033 Blindern, N-0315 Oslo, Norway.

10.1021/ct600234n CCC: $37.00
Published on Web 11/22/2006

with experimental results, and the existence of a large antisymmetric component of the shielding tensor for an atom in a distorted triple bond was confirmed. Subsequently, in 1997, the shielding and indirect nuclear spin−spin coupling constants of *o*-benzyne were studied experimentally by Warmuth[2] and theoretically by Jiao et al.[3] In the experimental investigation of Warmuth, *o*-benzyne was trapped by guest incarceration inside a molecular container, allowing its nuclear magnetic resonance (NMR) spectrum to be recorded. In particular, the carbon−carbon triple-bond spin−spin coupling was measured at 177.9 ± 0.7 Hz—that is, within the expected range for such bonds. These findings were supported by DFT calculations,[3] which gave a value of 192 Hz.

Because *o*-benzyne has a weak biradical character, with a low-lying excited triplet state, it presents a challenge for single-reference methods—in particular, for spin-restricted methods, as needed for the study of NMR parameters. Moreover, because of experimental problems related to its instability, no complete NMR spectrum has yet been recorded of the isolated *o*-benzyne molecule. Thus, the first analysis of the shielding tensor of the triply bonded carbon atom was performed in an argon matrix,[1] while the more recent experimental spectrum, providing the remaining shielding constants and a number of spin−spin coupling constants, was recorded for *o*-benzyne inside a molecular container.[2] Clearly, there may be significant differences between the measured NMR parameters and the corresponding values of an isolated molecule, as indicated by incarceration changes of the NMR parameters for other molecules. Therefore, the NMR parameters of the isolated *o*-benzyne molecule are not yet known precisely.

In view of the unusual character of *o*-benzyne and the difficulties it presents to both theory and experiments, we have here undertaken an independent theoretical study of this compound, with an aim to establish its NMR parameters reliably. At the same time, we analyze the performance of various popular DFT exchange−correlation functionals with respect to the calculation of NMR parameters of this difficult system.

2. Computational Details

The ACES II package was used in the coupled-cluster singles-and-doubles (CCSD) calculations;[4] all the DFT results were obtained using Dalton.[5] The NMR parameters and excitation energies were calculated using coupled-cluster and DFT linear response theory, using a closed-shell singlet reference wave function. The CCSD spin−spin calculations were carried out with all electrons correlated, without orbital relaxation. In the CCSD and DFT shielding calculations, London atomic orbitals were used to ensure gauge-origin invariant results.[6]

A. DFT Functionals. Concerning the choice of DFT functionals, it has over the past few years become apparent that it is preferable to use different functionals for the nuclear shielding constants and for the indirect nuclear spin−spin coupling constants. Thus, for shielding constants, recent investigations by Keal et al.[7] have revealed that the Keal−Tozer KT1 and KT2 functionals[8] perform very well, outperforming the local-density approximation (LDA),[9] the BLYP functional,[10,11] and the Becke−3-parameter−Lee−Yang−Parr (B3LYP) hybrid functional.[11,12] For the indirect spin−spin coupling constants, early experience indicated that performance improves in the sequence LDA, BLYP, and B3LYP;[13] moreover, in a study of small rigid hydrocarbons,[14] it was observed that B3LYP provides couplings that rival those of wave-function methods such as multiconfigurational self-consistent field theory, the second-order polarization propagator approximation, and CCSD theory. More recently, however, it has emerged that other functionals give better coupling constants than does B3LYP. Thus, from the work of Maximoff et al.[15] and Keal et al.,[16] it has been established that superior overall performance (at least for light atoms) is provided by the B97-2[17] and B97-3[18] semiempirical functionals, while hydrocarbons (especially CH couplings) are best served by the Perdew−Burke−Ernzerhof (PBE) functional.[19]

In view of these findings, we have, in our study of *o*-benzyne, carried out calculations using the following generalized gradient approximation (GGA) and hybrid functionals: the semiempirical BLYP GGA functional (fitted to noble-gas results) and B3LYP hybrid functional (fitted also to thermochemistry); the nonempirical Perdew−Wang 91 (PW91)[20] and PBE GGA functionals, both based on the properties of the slowly varying electron gas; the nonempirical KT1 and semiempirical KT2 GGA functionals; and the 10-parameter semiempirical B97-1 (fitted to thermochemistry) and B97-2 (fitted also to accurate exchange−correlation potentials) hybrid modifications of Becke's 97 (B97) functional. We also examined the performance of the LDA, CAM-B3LYP,[21] and B97-3 functionals but found these to be unsuitable, giving either erroneous geometries or instabilities in the calculation of the NMR parameters.

B. Basis Sets. At all levels of theory, the calculation of NMR shielding and spin−spin coupling constants is highly sensitive to the quality of the one-electron basis set. The reliable calculation of the usually dominant Fermi-contact (FC) contribution to spin−spin coupling constants, in particular, requires a flexible inner-core *s* basis, containing functions with large exponents—see, for instance, refs 22 and 23. In our calculations, we have therefore used the correlation-consistent core−valence *X*-tuple cc-pCVXZ basis sets, which combine such flexibility with the flexibility needed in coupled-cluster theory.

Unless otherwise stated, we have in the DFT calculations used the cc-pCVTZ triple-ζ basis,[24] consisting of a [12s7p3d1f/6s5p3d1f] carbon basis and a [5s2p1d/3s2p1d] hydrogen basis, with a total of 314 contracted Gaussians for *o*-benzyne. In the more expensive CCSD calculations, the cc-pCVTZ basis set was used for the shielding constants and for the dominant FC contribution to the indirect spin−spin coupling constants. To reduce cost, the remaining spin−spin contributions—that is, the spin−dipole (SD), the diamagnetic spin−orbit (DSO), and the paramagnetic spin−orbit (PSO) contributions—were calculated in cc-pCVDZ double-ζ basis. Finally, to examine basis-set convergence in the inner-core region, selected additional spin−spin coupling calculations

Table 1. Nuclear Shielding Constants in *o*-Benzyne (ppm)

	experimental geometry[a]										
	BLYP	PW91	PBE	KT1	KT2	B3LYP	B97-1	B97-2	HF	CCSD	exp.[b]
C1	−34.1	−31.8	−30.6	−10.5	−12.2	−34.7	−30.4	−26.8	−42.1	1.3	3.7[c]
C3	46.6	48.3	49.2	63.3	61.9	48.3	51.3	53.7	58.4	67.6	59.5
C4	35.3	38.1	39.2	55.5	53.9	38.1	42.2	44.9	55.4	59.0	48.2
H3	24.8	24.6	24.5	24.6	24.6	24.7	24.7	24.6	24.4	24.6	25.2[d]
H4	24.0	23.8	23.7	23.9	23.9	24.0	24.0	24.0	24.6	24.2	25.8[d]
	optimized geometry										
	BLYP	PW91	PBE	KT1	KT2	B3LYP	B97−1	B97−2	HF		
C1	−24.8	−21.3	−20.5	−3.3	−1.2	−20.2	−16.4	−12.0	−9.6		
C3	46.0	48.9	49.4	61.4	63.8	50.3	52.4	56.2	62.5		
C4	31.5	35.6	36.4	51.4	53.2	36.3	39.8	43.7	53.6		
H3	24.8	24.6	24.5	24.6	24.9	24.9	24.8	24.9	25.0		
H4	24.0	23.8	23.7	23.8	24.1	24.2	24.1	24.1	24.9		

[a] From microwave measurements.[31] [b] Recomputed from the experimental data of ref 2, using for the absolute shielding of C in TMS 186.37 ppm and for H in TMS 32.775 ppm.[37] [c] This constant is −6.6 ppm in ref 1, see text and Table 2. [d] Use of incarceration shifts of 1.23 and 3.48 ppm from Table 1 of ref 2 gives 26.6 and 25.0 ppm for H3 and H4, respectively.

Table 2. Shielding Tensor of the Triply Bonded Carbon Atom in *o*-Benzyne (ppm)

component[a]	CCSD[b]	HF[b]	KT1[b]	exp.[c]
σ_n	−53.6	−58.3	−62.2	−53.6
$\sigma_\perp$	−84.7	−150.4	−63.9	−79.6
$\sigma_{\|\|}$	142.0	179.8	116.2	113.4
angle	12.9	18.4	14.1	±6.7
$\sigma_{antisym}$	−90.5	−143.6	−73.3	not available
σ_{aniso}	211.2	284.1	179.3	180.0
σ_{ave}	1.3	−9.6	−3.3	−6.6

[a] σ_n is the out-of-plane tensor component; $\sigma_\perp$ and $\sigma_{||}$ are the in-plane components, approximately (see the line "angle") perpendicular and parallel to the triple bond. [b] CCSD results at the experimental geometry,[31] HF and KT1 at the optimized geometries. [c] Recomputed from the experimental data of ref 1, using for the absolute carbon shielding in TMS 186.37 ppm.

were carried out in the larger cc-pCVQZ quadruple-ζ basis, in contracted and uncontracted forms.

For DFT calculations of indirect nuclear spin−spin coupling constants, the HIIIsu3 basis set[13] has been shown to yield good results and has been extensively used for hydrocarbons[25] and, in particular, for the calculation of vibrational corrections to spin−spin coupling constants.[26] This basis consists of Huzinaga's HIII basis set[27,28] with the polarization functions and contractions of van Wüllen[29] and Kutzelnigg et al.[30] To ensure flexibility near the nuclei, the *s* functions have been decontracted and three steep *s* functions added in an even-tempered manner. We have used the HIIIsu3 basis to check the convergence of the cc-pCVTZ and cc-pCVQZ calculations and also for the calculation of vibrational corrections to the spin−spin coupling constants.

C. Geometries. The choice of molecular geometry is often a critical one in theoretical studies of molecular properties. A good approach is often to use consistently the experimental geometry. For *o*-benzyne, an experimental gas-phase geometry is available from recent microwave measurements.[31] However, because of its biradical character, *o*-benzyne has a low-lying triplet excited state that is poorly described by single-reference spin-restricted methods. In particular, in the spin-restricted reference wave functions used here (the HF reference wave function of CCSD and the Kohn−Sham noninteracting reference wave function of DFT), triplet instabilities are found near the minimum of the potential-energy surface. In such situations, the best description of triplet perturbations is obtained at the minimum of the chosen electronic-structure model—for example, the minimum of a specific DFT functional. Even small deviations from this geometry (when the experimental geometry or a geometry optimized with a different functional is used) may seriously affect the quality of calculated triplet properties such as the FC contributions to spin−spin coupling constants and triplet excitation energies. For this reason, we have in our *o*-benzyne study carefully considered the effect of calculating triplet properties at both the experimental geometry and the optimized geometry.

For DFT, two sets of calculations were therefore carried out for each functional—one at the experimental geometry, another one at the geometry obtained by energy minimization, using the same functional as in the subsequent NMR calculation. This minimization shortened the triple bond in all cases, by up to 2.4 pm from the experimental value of 126.4 pm, whereas the C—C≡C bond angle increased by up to 1.1° from the experimental value of 126.4°. In HF theory, the triple bond of 121.5 pm is much too short, while the C—C≡C bond angle is 127.6°. The all-electron CCSD/cc-pCVTZ geometry minimization was prohibitively expensive. For the CCSD model, therefore, we report here only NMR parameters calculated at the experimental geometry.

Although the geometry changes upon DFT energy minimization are relatively small, they significantly improve the calculated shielding and spin−spin coupling constants—in particular, for the atoms involved in the triple bond (see below). However, because we have not uncovered any general correlations (valid for all functionals) between the geometry changes and the changes in the NMR parameters upon energy minimization, we do not discuss the geometry changes further here, noting only that a good agreement with the experimental geometry does not necessarily translate into good NMR parameters.

In the following, the carbon atoms are numbered clock-

Table 3. Indirect Nuclear Spin−Spin Coupling Constants in *o*-Benzyne, Calculated in the cc-pCVTZ Basis at the Experimental Geometry[a] (Hz)

	BLYP	PW91	PBE	KT1	KT2	B3LYP	B97-1	B97-2	CCSD
^1J(C1≡2)	236.4	664.9	225.9	213.0	434.1	577.3	448.5	566.4	210.2
^1J(C2−C3)	75.4	121.9	69.4	69.8	96.7	108.9	109.3	110.1	83.4
^1J(C3=C4)	59.7	68.8	54.1	55.8	63.9	68.8	74.3	69.4	63.3
^1J(C4−C5)	77.9	90.0	71.8	73.8	83.9	90.1	93.2	89.2	77.2
^2J(C1C3)	3.2	−45.3	1.5	1.9	−21.0	−24.4	−16.3	−24.9	2.0
^2J(C1C5)	−15.4	−90.4	−16.4	−15.2	−53.1	−73.2	−52.0	−68.7	−13.8
^2J(C3C5)	−0.8	−11.9	−1.4	−1.3	−7.4	−7.2	−6.4	−7.6	−3.3
^3J(C1C4)	14.2	83.7	14.4	13.6	48.2	69.3	49.7	66.2	13.5
^3J(C3C6)	2.1	9.8	2.2	1.8	5.8	6.6	6.4	7.4	4.8
^1J(C3H3)	166.5	174.2	152.4	159.4	180.0	170.4	166.6	157.9	155.8
^1J(C4H4)	147.5	146.3	134.6	141.0	155.4	144.6	143.9	134.7	137.4
^2J(C2H3)	−6.3	−51.7	−7.3	−4.0	−22.8	−36.9	−23.3	−31.1	−5.8
^2J(C3H4)	8.9	8.6	8.1	9.9	9.9	8.9	7.1	7.1	4.0
^2J(C4H3)	7.4	−1.8	6.6	8.4	4.8	1.3	3.0	1.7	4.5
^2J(C4H5)	11.3	12.8	10.5	12.4	13.7	13.1	10.5	10.9	6.5
^3J(C1H3)	11.3	57.5	11.6	9.6	28.4	44.5	28.8	36.8	8.6
^3J(C1H5)	15.5	−7.9	14.4	16.6	5.2	−6.8	3.1	−2.1	13.1
^3J(C3H5)	14.2	12.4	13.4	15.1	15.1	11.9	12.6	11.8	11.5
^3J(C4H6)	5.5	13.3	5.3	5.5	9.1	11.1	8.9	9.7	5.6
^4J(C1H4)	11.9	37.9	11.9	12.2	26.8	32.6	21.8	26.6	4.7
^4J(C3H6)	−2.6	−8.3	−2.6	−2.7	−5.4	−5.8	−5.0	−5.5	−3.3
^3J(H3H4)	5.1	3.3	4.6	4.5	4.2	3.1	3.9	3.4	4.6
^3J(H4H5)	3.7	5.5	3.3	2.9	4.3	5.4	4.4	4.4	4.0
^4J(H3H5)	0.5	2.4	0.5	0.5	1.3	2.2	1.1	1.3	−0.4
^5J(H3H6)	2.8	7.7	2.8	2.6	4.6	5.6	4.0	4.4	2.0

[a] From microwave measurements.[31]

wise, beginning with the triple bond C1≡C2. The hydrogen atoms are numbered likewise, with H*n* attached to C*n*.

3. Results

A. NMR Shielding Constants. A comparison of the experimental and calculated shielding constants in Table 1 indicates that the best results are obtained with the KT1 and KT2 functionals, in agreement with the conclusions of ref 8. For the carbon atoms, the KT1 and KT2 shielding constants (at the optimized geometries) are fairly close to both the CCSD values and the experimental values. However, because we have not calculated vibrational corrections for the shielding constants and because the experimental values of refs 1 and 2 differ noticeably, we cannot draw any definite conclusions regarding the quality of the calculated shielding constants.

For all DFT functionals, geometry relaxation increases the C1 shielding and reduces the C4 shielding, while the C3 changes are smaller and less systematic. For the triply bonded C1 atom, the shortening of the triple bond upon relaxation significantly improves agreement with experimental results, as also observed in ref 3. This effect is particularly pronounced in HF theory, where the C1 shielding changes by more than 30 ppm upon geometry relaxation.

The hydrogen shielding constants in Table 1 differ from the experimental values by more than is usual for such constants. Also, unlike the experimental values, all of our calculations (except the HF calculation at the experimental geometry) indicate that H3 is more shielded than H4. However, one should bear in mind that the experimental values quoted in Table 1 (25.2 and 25.8 ppm for H3 and H4, respectively) were obtained by assuming the same large incarceration shift of 2.7 ppm for the two protons in *o*-benzyne, equal to the incarceration shift in benzene.[2] For other molecules, the two incarceration shifts differ significantly—for example, using the shifts of 1.23 and 3.48 ppm in Table 1 of ref 2, we obtain instead 26.6 and 25.0 ppm, respectively, for the H3 and H4 shieldings in *o*-benzyne.

As seen from Table 2, the CCSD, HF, and KT1 shielding constants σ_{ave} calculated for the triply bonded carbon atom are in reasonable agreement both with experiment and with one another. However, for the HF model, this agreement arises from a fortuitous cancellation of the errors in the in-plane tensor components. Moreover, whereas the HF anisotropy is too large, the CCSD and KT1 anisotropies are closer to the experimental values. Consequently, it appears that the HF antisymmetric component is too large—the much lower (in absolute value) CCSD and KT1 results are in better agreement with each other. The KT1 shielding components of the triply bonded carbon atom agree also with the BLYP results of ref 1.

B. Indirect Nuclear Spin−Spin Coupling Constants. *1. Calculation of Spin−Spin Coupling Constants.* In Tables 3 and 4, we have listed the indirect nuclear spin−spin coupling constants in *o*-benzyne, calculated at the experimental geometry of ref 31 and at the optimized geometry, respectively. There is an unusually large variation among the results obtained with the different DFT functionals. A comparison with CCSD suggests that the most accurate DFT values are those obtained with the PBE functional, at its optimized

Table 4. Indirect Nuclear Spin−Spin Coupling Constants in *o*-Benzyne, Calculated in the cc-pCVTZ Basis at the Optimized Geometry[a] (Hz)

	BLYP	PW91	PBE[b]	KT1	KT2	B3LYP	B97-1	B97-2	vib.[c]	exp.
^1J (C1≡C2)	222.3	389.2	208.7	205.7	309.7	322.2	310.3	334.0	−1.0	177.9 ± 0.7
^1J (C2−C3)	80.6	103.7	74.7	74.0	92.4	96.8	104.4	99.0	−2.6	75.7 ± 0.9
^1J (C3=C4)	54.2	60.0	48.9	49.9	57.3	61.5	67.8	62.4	−1.1	50.9 ± 0.8
^1J (C4−C5)	76.0	84.1	71.3	72.9	81.6	82.6	88.7	83.6	−0.7	71.0 ± 0.8
^2J (C1C3)	5.5	−18.1	4.1	3.6	−8.3	−3.9	−3.8	−5.8	−3.2	
^2J (C1C5)	−12.5	−47.7	−13.0	−13.6	−33.3	−31.3	−29.3	−33.0	−3.0	
^2J (C3C5)	−1.0	−8.2	−1.5	−1.7	−5.8	−4.2	−4.9	−5.1	−0.8	
^3J (C1C4)	12.0	43.8	11.9	12.6	30.2	30.1	28.6	32.6	3.7	
^3J (C3C6)	2.2	7.2	2.2	1.9	4.8	4.9	5.5	5.9	0.3	
^1J (C3H3)	167.4	172.2	153.7	162.9	177.1	166.6	165.2	154.4	3.5	
^1J (C4H4)	149.1	148.6	135.3	142.8	153.7	145.6	144.5	134.1	1.8	
^2J (C2H3)	−1.3	−21.9	−2.3	0.8	−10.3	−11.6	−9.2	−11.6	−0.3	
^2J (C3H4)	9.1	7.3	8.1	10.1	9.1	7.6	6.6	6.1	0.0	
^2J (C4H3)	6.2	0.5	5.4	7.4	4.3	3.0	3.0	2.4	−0.4	
^2J (C4H5)	11.6	11.0	10.9	13.1	12.7	10.7	9.5	9.3	0.4	
^3J (C1H3)	9.0	28.0	9.1	8.1	17.4	19.1	16.1	17.9	3.4	
^3J (C1H5)	15.3	6.5	14.5	15.8	11.4	8.2	9.7	8.3	−0.6	
^3J (C3H5)	13.5	13.0	12.9	14.4	14.7	12.1	12.3	11.7	0.7	
^3J (C4H6)	4.7	8.7	4.6	4.7	7.0	6.9	6.6	6.6	0.5	
^4J (C1H4)	9.8	20.5	9.9	11.1	17.6	14.7	12.8	13.6	1.5	
^4J (C3H6)	−2.5	−5.8	−2.5	−2.6	−4.4	−4.0	−4.0	−4.2	−0.1	
^3J (H3H4)	4.8	4.6	4.4	3.9	4.6	4.4	4.3	4.2	0.4	
^3J (H4H5)	3.2	4.4	3.0	2.5	3.7	4.2	3.8	3.7	0.4	
^4J (H3H5)	0.2	0.8	0.2	0.3	0.5	0.6	0.3	0.3	0.2	
^5J (H3H6)	2.4	5.0	2.6	2.3	3.7	3.5	3.0	3.1	0.1	

[a] Experimental data of ref 2. [b] Very similar results are obtained in the cc-pCVQZ basis—for instance, ^1J (C1≡C2) = 205.4 Hz, ^1J (C2−C3) = 74.6 Hz, ^1J (C3=C4) = 49.2 Hz, and ^1J (C4−C5) = 71.4 Hz—and in the uncontracted cc-pCVQZ basis—see Table 7. [c] Zero-point vibrational corrections calculated at the PBE/cc-pCVTZ level of theory.

Table 5. Lowest Triplet Excitation Energy in *o*-Benzyne, Calculated at the Experimental and Optimized Geometry (eV)[a]

geometry	BLYP	PW91	PBE	KT1	KT2	B3LYP	B97-1	B97-2
experimental[b]	1.687	0.824	1.608	1.783	1.134	0.915	1.163	1.044
optimized	1.889	1.181	1.828	1.912	1.469	1.425	1.587	1.544

[a] The experimental[33] and EOM-CCSD excitation energies are 1.656 and 2.037 eV, respectively [b] Ref 31.

geometry. Except for the very small ^4J (H3H5) constant, all PBE coupling constants have the correct sign and are similar to the CCSD constants. Moreover, when PBE and CCSD differ, the PBE values are sometimes closer to the experimental values, indicating that the difference may arise from a CCSD error. Recently, in a comparison of 20 density functionals, the PBE functional was shown to yield the best ^1J (CH) constants.[15] Although, in a subsequent study, PBE did not show the same good performance for other coupling constants and other molecules,[16] it performs very well indeed for hydrocarbons.

The accuracy of the calculated spin−spin coupling constants, which is usually dictated by the accuracy of the FC contribution, correlates strongly with the quality of the lowest vertical triplet excitation energies listed in Table 5, computed with linear response theory using the same functional as in the energy optimization.[32] In all cases, geometry optimization increases the triplet excitation energy, indicating that the optimization moves the molecule away from the nearby triplet instability. Thus, whenever the lowest calculated triplet excitation energy at the experimental geometry is much too small (for PW91, KT2, B3LYP, B97-1, and B97-2), the corresponding ^1J (C1≡C2) coupling constant is too large by a factor of 2−3. At the optimized geometry, where the calculated excitation energy is much closer to the accurate (experimental and theoretical) value of about 1.65 eV,[33,34] the calculated coupling constant is smaller and closer to the experimental value. These correlations between the lowest triplet excitation energy and the spin−spin coupling constants are easily understood from Ramsey's theory,[35] where the dominant FC contribution is expressed as a sum over triplet states, with excitation energies in the denominator. Among the DFT functionals, the very poor performance of the PW91 functional is striking—in particular, in comparison with the good performance of the similarly constructed PBE functional. Apparently, the PW91 functional provides a poor description of triplet perturbations in this molecule, giving a FC contribution to the ^1J (C1≡C2) coupling constant that is 3 times larger than that obtained with the PBE functional.

To understand better the reasons for the large errors in the DFT spin−spin coupling constants of *o*-benzyne, we have calculated, at the same levels of theory, the coupling constants of acetylene, both at the $D_{\infty h}$ equilibrium geometry and at a bent C_{2v} geometry with an H−C≡C bond angle of 135.0°. At equilibrium, the DFT acetylene results are all consistent with one another—in the cc-pCVTZ basis, for

Table 6. ^1J (C1≡C2) Coupling Constant in Acetylene and *o*-Benzyne (Hz)[a]

	BLYP	PW91	PBE	KT1	KT2	B3LYP	B97-1	B97-2	CCSD	exp.
linear C_2H_2	222.1	221.0	209.8	211.3	220.1	225.2	235.8	224.6	202.5	185.0[b]
bent C_2H_2[a]	221.0	267.2	206.5	210.1	261.7	253.7	259.6	271.0	196.9	
o-benzyne	236.4	664.9	225.9	213.0	434.1	577.3	448.5	566.4	210.2	177.9 ± 0.7

[a] All calculations in the cc-pCVTZ basis. The H−C≡C angle is 135.0° in the bent C_2H_2 model, compared with 126.4° in *o*-benzyne. [b] See ref 38 for a discussion of this (derived) C_2H_2 equilibrium spin−spin coupling constant and ab initio results. The experimental gas-phase value of ^1J (C1≡C2) is 174.78 ± 0.02 Hz.[37]

Table 7. Carbon−Carbon Indirect Nuclear Spin−Spin Coupling Constants in *o*-Benzyne and Benzene (Hz)

	o-benzyne							benzene[a]
	PBE[b]	PBE[c]	PBE[d]	PW91[e]	PW91[f]	CCSD[g]	exp.[h]	PBE[b]
^1J (C1≡C2)	208.7	203.0	211.7	389.2	191.9	210.2	177.9 ± 0.7	50.6
^1J (C2−C3)	74.7	73.8	76.7	103.7	72.8	83.4	75.7 ± 0.9	
^1J (C3=C4)	48.9	48.5	50.4	60.0	49.8	63.3	50.9 ± 0.8	
^1J (C4−C5)	71.3	70.5	72.8	84.1	69.0	77.2	71.0 ± 0.8	
^2J (C1C3)	4.1	3.5	3.7	18.1	3.4[i]	2.0		−0.5
^2J (C1C5)	−13.0	−14.0	−14.5	−47.7	−10.9[i]	−13.8		
^2J (C3C5)	−1.5	−1.9	−1.9	−8.2	−1.3	−3.3		
^3J (C1C4)	11.9	13.2	13.7	43.8	10.8	13.5		9.2
^3J (C3C6)	2.2	2.9	2.9	7.2	2.1	4.8		

[a] Only the unique ^1J (CC), ^2J (CC), and ^3J (CC) coupling constants are tabulated. [b] Calculation in the cc-pCVTZ basis at optimized geometry. [c] Calculation in the uncontracted cc-pCVQZ basis at optimized geometry. [d] Calculation in the HIIIsu3 basis at optimized geometry. [e] Calculation in the cc-pCVTZ basis at PW91 optimized geometry. [f] Calculation by Jiao et al.[3] [g] FC term calculated in the cc-pCVTZ basis, the remaining terms in the cc-pCVDZ basis. [h] Experimental results from ref 2. [i] It appears from this compilation that in ref 3 the ^2J (C1C3) and ^2J (C1C5) constants have been interchanged.

example, all ^1J (C1≡C2) values are between 209 and 236 Hz, see Table 6. Moreover, the DFT couplings are noticeably closer to the CCSD value of 203 Hz than to the experimental value of 185 Hz. At this geometry, the error relative to the experiment arises from basis-set incompleteness—when the FC contribution is recalculated in the cc-pCVQZ basis, we obtain 189 Hz at the PBE level and 180 Hz at the CCSD level. By contrast, in bent acetylene, the differences among the DFT functionals become large, albeit not as large as in *o*-benzyne (the C−C≡C angle in *o*-benzyne is 126.4°).

In Table 7, we have listed the carbon−carbon spin−spin coupling constants of *o*-benzyne calculated at different levels of theory, together with the experimental values and the coupling constants in benzene. First, comparing the cc-pCVTZ, uncontracted-cc-pCVQZ, and HIIIsu3 coupling constants calculated with the PBE functional, we find good agreement overall, with the largest differences for the ^1J (C≡C) coupling. For this coupling, the three basis sets yield 208.7, 203.0, and 211.7 Hz, respectively. Noting that the cc-pCVTZ and HIIIsu3 results never differ by more than 3 Hz, we conclude that the cc-pCVTZ basis set has sufficient flexibility in the core region.

Next, we note that the PBE/cc-pCVTZ value of 209 Hz for ^1J (C≡C) differs dramatically from the PW91/cc-pCVTZ value of 389 Hz. By contrast, Jiao et al.[3] obtained 192 Hz for this coupling, also with the PW91 functional. The reason for their better agreement with the experimental value of 178 Hz is the use of a shorter triple-bond length—interested primarily in the bonding of *o*-benzyne, these authors deduced the geometry from a comparison of the calculated and experimental NMR parameters. For many functionals, including PW91, the one-bond spin−spin coupling constant decreases strongly with a shortening of the triple bond, explaining the discrepancy between our results and those of Jiao et al.[3]

2. Survey of Spin−Spin Coupling Constants. In the following, we shall survey the indirect nuclear spin−spin coupling constants of *o*-benzyne, basing our discussion on the CCSD results and on the PBE results obtained at the optimized geometry, taking these to be the most reliable complete sets of coupling constants for this molecule.

For the one-bond carbon−carbon coupling constants, we first note the strong coupling across the triple bond. Indeed, our PBE/cc-pCVTZ value of 209 Hz, close to the CCSD value of 210 Hz, suggests that ^1J (C1≡C2) in isolated *o*-benzyne at equilibrium is larger than both the DFT result of Jiao et al. (192 Hz) and the experimental result of Warmuth (178 Hz). Whereas the discrepancy with Jiao et al. arises from different geometries in the calculations, the discrepancy with the experiment arises because we have neglected the effects of incarceration inside a molecular container. Because incarceration is likely to affect the geometry of *o*-benzyne, it will also affect the geometry-sensitive coupling across the triple bond.

The remaining PBE one-bond carbon−carbon coupling constants are, like those of Jiao et al.,[3] in good agreement with the experimental measurements, differing by only 1−2 Hz. On the other hand, the agreement with the CCSD values, which are too high by 5−15 Hz, is now poorer. (We note that the rovibrational corrections discussed below make the agreement of all these constants with the experimental values slightly poorer.) Finally, in agreement with the experimental values, the PBE ^1J (C3=C4) value of 49 Hz is closest to the one-bond carbon−carbon coupling of 51 Hz in benzene, the

PBE values of ^1J (C2—C3) and ^1J (C4—C5) being 75 and 71 Hz, respectively.

Regarding the geminal and vicinal carbon−carbon spin−spin coupling constants in *o*-benzyne, we note that the PBE constants agree well with our CCSD results and almost as well with the PW91 results of Jiao et al.,[3] provided we interchange their ^2J (C1C3) and ^2J (C1C5) values. Interestingly, whereas the geminal coupling is only −0.5 Hz in benzene, it is much stronger in *o*-benzyne, varying from −13 to +4 Hz; furthermore, whereas the vicinal coupling is 9 Hz in benzene, the two distinct vicinal couplings are 2 and 12 Hz in *o*-benzyne. Clearly, all carbon−carbon couplings are strongly affected by the triple bond in *o*-benzyne, not just those close to the triple bond.

Turning our attention to the carbon−proton interactions, we note from Table 4 that the value of the one-bond coupling ^1J (C4H4) of 135 Hz differs little from that in benzene (136 Hz), whereas the coupling next to the triple bond ^1J (C3H3) is enhanced by about 20 Hz. By contrast, the geminal and vicinal carbon−proton couplings are more strongly affected, as can be rationalized from the more delocalized character of these couplings. Thus, the geminal carbon−proton couplings in *o*-benzyne vary between −2.3 and 10.9 Hz (compared with 2.5 Hz in benzene); similarly, the vicinal couplings of *o*-benzyne vary between 4.6 and 14.5 Hz (compared with 6.7 Hz in benzene). Finally, we note that the long-range ^4J (C1H4) coupling of 12 Hz differs significantly both from the ^4J (C3H6) coupling of −2.6 Hz and from the corresponding benzene coupling of −0.7 Hz. Regarding the proton−proton interactions, the differences between *o*-benzyne and benzene are smaller, the values of ^3J (HH), ^4J (HH), and ^5J (HH) in benzene being 6.0, 1.2, and 0.5 Hz, respectively.

The relative importance of the different Ramsey contributions—that is, the FC, SD, DSO, and PSO contributions—is essentially the same at all levels of theory. We base the following discussion on the geometry-optimized PBE results, because we were unable to carry out the all-electron CCSD calculations of all four contributions in the cc-pCVTZ basis. The FC term dominates most of the coupling constants—in particular, the large ones. The PSO contribution is relatively large for the one-bond carbon−carbon couplings and constitutes, for example, about 13% of ^1J (C═C). The PSO term is important also for the proton−proton coupling constants, although it is here canceled by the DSO term as happens also in many other molecules—see, for example, ref 36. Except when it cancels the PSO term, the DSO term is always small. The SD term is significantly smaller but relevant (in relative terms) for some of the smaller constants. For the coupling across the triple bond, the PSO and SD terms both contribute about 2% but cancel coincidentally.

We have estimated the zero-point vibrational corrections to the spin−spin coupling constants as described by Ruden et al.,[26] including anharmonic as well as harmonic contributions. In evaluating the vibrational corrections, the first and second derivatives of the spin−spin coupling constants are calculated numerically, while the force constants are obtained numerically from analytically calculated gradients.

In Table 4, we have listed the corrections calculated at the PBE/cc-pCVTZ level of theory. A comparison with the HIIIsu3 results reveals a large uncertainty in the vibrational correction to ^1J (C1≡C2), due to a near cancellation of the harmonic and anharmonic contributions. In particular, for this coupling constant, the cc-pCVTZ and HIIIsu3 vibrational corrections are −1.0 and 0.5 Hz, respectively. [The remaining corrections are much less basis-set-dependent, except for an uncertainty of about 1 Hz in the ^1J (C3H3) and ^1J (C4H4) corrections.] Thus, although the sign of the vibrational correction to ^1J (C1≡C2) is unknown, its magnitude is small, modifying the coupling constant by less than 1%. Clearly, the large difference between the calculated and measured ^1J (C1≡C2) values does not arise from vibrations.

From Table 4, we note that the vibrational corrections to the one-bond spin−spin coupling constants in *o*-benzyne are all small, never exceeding 3% of the total coupling. By contrast, the geminal and vicinal corrections are large—reducing, for example, ^2J (C1C3) from 4.1 to 0.7 Hz and increasing ^3J (C1H3) from 9.1 to 12.5 Hz. Interestingly, the vibrational corrections to the small geminal and vicinal ^2J (C1C3), ^3J (C1C4), and ^2J (C1C5) coupling constants are all larger in magnitude than the vibrational correction to the much larger one-bond ^1J (C1C2) coupling constant. Although these corrections may be unimportant for the comparison with incarcerated *o*-benzyne, they do illustrate that indirect nuclear spin−spin coupling constants can sometimes be dramatically affected by molecular vibrations.

4. Conclusions

We have carried out an extensive theoretical study of the NMR parameters in *o*-benzyne, using DFT and coupled-cluster theory. Because of its special structure and biradical character, the calculation of these parameters for *o*-benzyne is difficult, the results depending critically on the choice of exchange−correlation functional and on the reference geometry. In general, we find that the DFT results become significantly more reliable when calculated at the optimized geometry, using the same functional for the geometry optimization and the calculation of NMR parameters. At these optimized geometries, the electronic structure (as described by restricted theories) appears to be less strained and less affected by, for example, triplet instabilities, as seen by comparing the lowest triplet excitation energy in *o*-benzyne calculated at the experimental and optimized geometries. We note that unrestricted electronic-structure theory cannot be used in our calculations to describe the unperturbed reference state because the associated spin contamination would make the calculation of NMR parameters meaningless.

Comparing the performance of the different exchange−correlation functionals at the optimized geometries, we find that the best results (in comparison with CCSD theory) are obtained with the KT1 and KT2 functionals for the shielding constants and by the PBE functional for the indirect spin−spin coupling constants. These findings are reassuring in the sense that these conclusions, regarding the relative performance of different functionals for NMR parameters, agree

with recent benchmark studies on other molecules. We are therefore confident that the results obtained with these functionals, at the optimized geometries, are qualitatively correct and may be meaningfully compared with experimental and previous theoretical results. Indeed, with these functionals, the agreement with the available experimental results is good.

Numbering the atoms consecutively, beginning with the triply bonded carbon atoms, we obtain −3.3, 61, and 51 ppm for the C1, C3, and C4 shieldings, respectively, at the KT1/cc-pCVTZ level of theory, compared with the experimental values of 3.7, 60, and 48 ppm. Particularly noteworthy is the small shielding constant of the triply-bonded carbon atom, in agreement with the experimental results. The differences may arise either from an inadequacy of the exchange−correlation functional or from incarceration.

For the spin−spin coupling constants, we find the performance of the single-reference PBE method similar to the performance of CCSD. At the PBE/cc-pCVTZ level of theory, we obtain 205 Hz for ^1J across the triple bond in isolated *o*-benzyne at equilibrium, 49 Hz across the double bond, and 75 and 71 Hz across the single bonds adjacent and not adjacent to the triple bond, respectively; the corresponding experimental values for the incarcerated molecule are 178, 51, 76, and 71 Hz, respectively. Comparing also with CCSD theory, we believe that the discrepancy with the experimental triple-bond spin−spin coupling is again due to incarceration and that the true equilibrium carbon−carbon triple-bond spin−spin coupling in *o*-benzyne is about 200 Hz.

Acknowledgment. We acknowledge financial support from the European Marie Curie Research and Training Network NANOQUANT, contract number MCRTN-CT-2003-506842. This work has received support from the Norwegian Research Council through a Strategic University Program in Quantum Chemistry (Grant No. 154011/420) and through a grant of computer time from the Program for Supercomputing.

References

(1) Orendt, A. M.; Facelli, J. C.; Radziszewski, J. G.; Horton, W. J.; Grant, D. M.; Michl, J. *J. Am. Chem. Soc.* **1996**, *118*, 846.

(2) Warmuth, R. *Angew. Chem., Int. Ed. Engl.* **1997**, *36*, 1347.

(3) Jiao, H.; Schleyer, P. v. R.; Beno, B. R.; Houk, K. N.; Warmuth, R. *Angew. Chem., Int. Ed. Engl.* **1997**, *36*, 2761.

(4) Aces II Mainz-Austin-Budapest-Version 2005, a quantum chemical program package, written by J. F. Stanton, J. Gauss, J. D. Watts, P. G. Szalay, and R. J. Bartlett with contributions from A. A. Auer, D. B. Bernholdt, O. Christiansen, M. E. Harding, M. Heckert, O. Heun, C. Huber, D. Jonsson, J. Juselius, W. J. Lauderdale, T. Metzroth, K. Ruud and the integral packages MOLECULE (J. Almlöf and P. R. Taylor), Props (P. R. Taylor), and ABACUS (T. Helgaker, H. J. Aa. Jensen, P. Jørgensen, and J. Olsen). See also Stanton, J. F.; Gauss, J.; Watts, J. D.; Lauderdale, W. J.; Bartlett, R. J. *Int. J. Quantum Chem. Symp.* **1992**, *26*, 879, as well as http://www.aces2.de (accessed Nov 2006) for the current version.

(5) Dalton, a molecular electronic structure program, release 2.0 (2005), see http://www.kjemi.uio.no/software/dalton/dalton.html (accessed Nov 2006).

(6) London, F. *J. Phys. Radium* **1937**, *8*, 397.

(7) Keal, T. W.; Tozer, D. J.; Helgaker, T. *Chem. Phys. Lett.* **2004**, *391*, 374.

(8) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2003**, *119*, 3015.

(9) Vosko, S.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.

(10) Becke, A. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.

(11) Lee, C.; Yang, W.; Parr, R. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.

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

(13) Helgaker, T.; Watson, M.; Handy, N. C. *J. Chem. Phys.* **2000**, *113*, 9402.

(14) Ruden, T. A.; Helgaker, T.; Jaszuński, M. *Chem. Phys.* **2004**, *296*, 53.

(15) Maximoff, S. N.; Peralta, J. E.; Barone, V.; Scuseria, G. E. *J. Chem. Theory Comput.* **2005**, *1*, 541.

(16) Keal, T. W.; Helgaker, T.; Sałek, P.; Tozer, D. J. *Chem. Phys. Lett.* **2006**, *425*, 163.

(17) Wilson, P. J.; Bradley, T. J.; Tozer, D. J. *J. Chem. Phys.* **2001**, *115*, 9233.

(18) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2005**, *123*, 121103.

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

(20) Perdew, J.; Chevary, J. A.; Vosko, S.; Jackson, K.; Pederson, M.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *46*, 6671.

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

(22) Helgaker, T.; Jaszuński, M.; Ruud, K.; Górska, A. *Theor. Chem. Acc.* **1998**, *99*, 175.

(23) Peralta, J. E.; Scuseria, G. E.; Cheeseman, J. R.; Frisch, M. J. *Chem. Phys. Lett.* **2003**, *375*, 452.

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

(25) Lutnæs, O. B.; Ruden, T. A.; Helgaker, T. *Magn. Reson. Chem.* **2004**, *42*, S117.

(26) Ruden, T. A.; Lutnæs, O. B.; Helgaker, T.; Ruud, K. *J. Chem. Phys.* **2003**, *118*, 9572.

(27) Huzinaga, S. Approximate Atomic Functions Technical Report, University of Alberta, Edmonton, Canada, 1971.

(28) Huzinaga, S. *J. Chem. Phys.* **1965**, *42*, 1293.

(29) van Wüllen, C. Die Berechnung Magnetischer Eigenschaften unter Berücksichtigung der Elektronkorrelation: die Multikonfigurations-Verallgemeinerung der IGLO-Methode, Ph.D. thesis, Ruhr-Universität Bochum, Bochum, Germany, 1992.

(30) Kutzelnigg, W.; Fleischer, U.; Schindler, M. In *NMR Basic Principles and Progress*; Springer: Berlin, 1990; Vol. 23, page 165.

(31) Kukolich, S. G.; McCarthy, M. C.; Thaddeus, P. *J. Phys. Chem. A* **2004**, *108*, 2645.

(32) Olsen, J.; Jørgensen, P. *J. Chem. Phys.* **1985**, *82*, 3235.

(33) Slipchenko, L. V.; Krylov, A. I. *J. Chem. Phys.* **2002**, *117*, 4694.

(34) Shao, Y.; Head-Gordon, M.; Krylov, A. I. *J. Chem. Phys.* **2003**, *118*, 4807.

(35) Ramsey, N. F. *Phys. Rev.* **1953**, *91*, 303.

(36) Pecul, M.; Dodziuk, H.; Jaszuński, M.; Lukin, O.; Leszczyński, J. *Phys. Chem. Chem. Phys.* **2001**, *3*, 1986.

(37) Jackowski, K.; Wilczek, M.; Pecul, M.; Sadlej, J. *J. Phys. Chem. A* **2000**, *104*, 5955.

(38) Jaszuński, M.; Ruud, K. *Chem. Phys. Lett.* **2001**, *336*, 473.

CT600234N

JCTC Journal of Chemical Theory and Computation

Theoretical Investigations of the Nature of Interaction of ClF with HF, H_2O, and NH_3

Junyong Wu, Jingchang Zhang, Zhaoxu Wang, and Weiliang Cao*

Institute of Modern Catalysis, Beijing University of Chemical Technology, State Key Laboratory of Chemical Resource Engineering, Beijing 100029, China

Received July 8, 2006

Abstract: The interactions of the first-row hydrides (HF, H_2O, and NH_3) with ClF have been investigated by performing calculations at the second-order perturbation theory based on the Møller−Plesset partition of the Hamiltonian with the aug-cc-pVTZ basis set. The geometries and vibrational frequencies in the present study were obtained by carrying out explicit counterpoise-corrected optimization. In order to understand that the Cl−X-type (X = F, O, and N) structure is more stable than the corresponding hydrogen-bonded structure in these complexes, the electronic properties were also investigated. Furthermore, the symmetry-adapted perturbation theory calculations were performed to gain more insight into the nature of the hydrogen-bond and Cl−X-type interactions. The analysis of the interaction energy components indicates that, in contrast to the hydrogen-bonded complexes, the inductive and dispersive interaction is the most important term in the Cl−X-type complexes, as we progress from HF to NH_3.

1. Introduction

The number of individual crystal structures, in which weak interactions have been reported to be important, has grown rapidly in recent years.[1−3] Therefore, understanding the nature of these intermolecular interactions is a necessary step toward a full rationalization of the packing and also a key preliminary step to design new crystals. By consideration that crystal packing results from the sum of many different contributions of directional and nondirectional intermolecular interactions, it is important that different types of interactions should be considered jointly in structure analysis. Recently, although research has traditionally focused on the more well-known hydrogen-bonded interactions,[3−6] a growing system of experimental and theoretical evidence confirms that interactions such as −X−Y− (X = F, Cl, Br, or I; Y = N, O, S or π) similarly play important roles in crystal engineering.[7−32]

Although the natures of hydrogen bonds and interactions of the −X−Y− type are different, it is still significant to compare the hydrogen bond with the interactions of the −X−Y− type. So in this paper, we investigated the nature of hydrogen bond and −X−Cl− type interactions of ClF with the first hydrides (HF, H_2O, and NH_3). We focused on the lowest interaction potential complexes of the investigated hydrogen-bonded and −X−Cl− type complexes. To further investigate the relative importance of electrostatic, dispersion, induction, and exchange-repulsion energies of these hydrogen-bonded and −X−Cl− type (also referred to in this paper as X−Cl-type) complexes, we have decomposed the interaction energies into these components using symmetry-adapted perturbation theory (SAPT).[33,34]

2. Theoretical Methods

All the results presented in this work were produced by employing either conventional supermolecular (SM) variational or SAPT methods.[33,34] Though the SM method is conceptually and computationally simple, it does not provide a clear picture of the interaction forces which are responsible for the interaction. On the other hand, the SAPT method computes the interaction energy directly as a sum of electrostatic, exchange-repulsion, induction, and dispersion contributions so that a physical interpretation of the interactions between the complex monomers can be evinced. The

* Corresponding author phone: +86 10 64444919; fax: +86 10 64434898; e-mail: caowl@mail.buct.edu.cn.

10.1021/ct600229n CCC: $37.00
Published on Web 12/10/2006

details of the calculations are briefly elaborated to aid in the discussion of the results.

2.1. Supermolecular Calculations. The second-order Møller-Plesset theory (MP2)[35] has been shown to be effective and accurate in determining the equilibrium structure and binding energy for many hydrogen-bonded and other weakly bonded complexes.[36] The basis set applied here is Dunning's correlation consisted basis set aug-cc-pVTZ.[37] The basis-set superposition error (BSSE) was eliminated by the standard counterpoise (CP) correction method of Boys and Bernardi.[38] Again, some authors claimed that the normal recipe of counterpoise correction of carrying out a single-point correction without further optimization could not find the correctly optimized structures and frequencies. They advocated that geometrical parameters, vibrational frequencies, and energies should be determined using explicit BSSE corrections.[39−42] So in the present study, ab initio structures of complexes were determined using counterpoise-corrected gradient optimization at the MP2(full)/aug-cc-pVTZ level. No symmetries were constrained in the optimization. Frequency calculations were performed to verify that the structures were minima on the potential energy surface. The BSSE-corrected vibrational frequencies and hence the zero-point vibrational energy (ZPVE) corrections were also evaluated for all complexes at the same theory level. The analysis went further with those obtained by means of the natural bond orbital (NBO) theory of Weinhold and co-workers.[43] The NBO analysis will allow us to quantitatively evaluate the charge transfer (CT) involving the formation of a hydrogen bond or halogen bond. All ab initio calculations were carried out with the Gaussian 03 suite of programs.[44] NBO analysis was conducted on the MP2-optimized structures, the Hartree−Fock (HF) densities, and the built-in subroutines of the Gaussian 03 program.

The performance of the all-electron MP2 method with the aug-cc-pVTZ basis set is examined by studying four small monomer molecules. The optimal bond lengths of HF and ClF calculated at the MP2(full)/aug-cc-pVTZ level are 0.9202 and 1.6346 Å, respectively, which is in very good agreement with the experimental values of 0.9170 and 1.6281 Å.[45] For water, the MP2(full)/aug-cc-pVTZ bond length of 0.9590 Å and bond angle of 104.2° compare well with the experimental bond length of 0.9555 Å and bond angle of 104.5°.[45] For ammonia, the MP2(full)/aug-cc-pVTZ bond length of 1.0098 Å and bond angle of 106.8° also compare well with the experimental bond length of 1.0120 Å and bond angle of 106.7°.[45]

2.2. SAPT Calculations. In this study, the SAPT calculations reported here used the correlation level technically designated as SAPT2, and they were carried out using the aug-cc-pVTZ basis set at the MP2(full)/aug-cc-pVTZ geometry. The SAPT interaction energy can be represented as $E_{int} = E_{int}^{HF} + E_{int}^{CORR}$, where E_{int}^{HF} is the sum of all of the energy components evaluated at the Hartee−Fock level and E_{int}^{CORR} is the sum of all of the energy components evaluated at the corrected level. E_{int}^{HF} can be represented as

$$E_{int}^{HF} = E_{elst}^{(10)} + E_{exch}^{(10)} + E_{ind,resp}^{(20)} + E_{exch-ind,resp}^{(20)} + \delta E_{int,resp}^{HF}$$

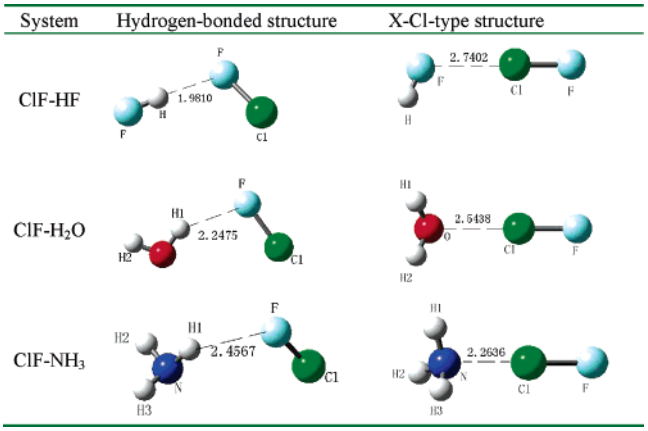


Figure 1. MP2(full)/aug-cc-pVTZ optimized low-energy structures (distances in Å). The dashed lines indicate noncovalent bond.

The superscript (*ab*) denotes orders in perturbation theory with respect to the intermolecular interaction operator and the intramolecular correlation operator, respectively. It can be seen from the above equation that the HF interaction energy includes first-order polarization and exchange and second-order induction and exchange-induction contributions. The subscript "resp" indicates that the induction and exchange-induction contributions include the coupled-perturbed HF response.[46] δE_{int}^{HF} contains the third- and higher-order HF induction and exchange induction contributions.

We have employed the SAPT2 approach, in which the correlative portion of the interaction energy E_{int}^{CORR} is nearly equivalent to the supermolecular MP2 correlation energy and can be represented as

$$E_{int}^{CORR} = E_{elst,resp}^{(12)} + E_{exch}^{(11)} + E_{exch}^{(12)} + {}^{t}E_{ind}^{(22)} + {}^{t}E_{exch-ind}^{(22)} + E_{disp}^{(20)} + E_{exch-disp}^{(20)}$$

Where ${}^{t}E_{ind}^{(22)}$ represents the part of $E_{ind}^{(22)}$ that is not included in $E_{ind,resp}^{(20)}$. A more detailed description of SAPT and some of its applications can be found in some recent references.[33,34,47−52] SAPT calculations were performed using the SAPT2002 program.[53]

3. Results and Discussion

3.1. Geometrical Parameters, Interaction Energies, and Vibrational Frequencies. The equilibrium geometries for the minimum energy hydrogen-bonded and X−Cl-type structures of ClF with the first hydrides (HF, H_2O, and NH_3) are displayed in Figure 1. Theoretical results for the interaction energies of these complexes are summarized in Table 1. Some selected geometrical parameters and vibrational frequencies are given in Tables 2 and 3.

Table 1 lists the interaction energies (ΔE), BSSE, interaction energies corrected for BSSE (ΔE^{CP}), ZPVE, interaction energies corrected for both BSSE and ZPVE ($\Delta E^{CP+ZPVE}$), and intermolecular distances (R_{int}) obtained at the MP2(full)/aug-cc-pVTZ level. The importance of the inclusion of electron correlation in the description of these complexes can be seen from the values of ΔE_{corr}. It is clear from Table

Table 1. Interaction Energies without (ΔE) and with (ΔE^{CP}) BSSE Correction and BSSE, ZPVE, and Interaction Energies for Both BSSE and ZPVE ($\Delta E^{CP+ZPVE}$)[a,b]

	ClF···HF	ClF···H_2O	ClF···H_3N	HF···ClF	H_2O···ClF	H_3N···ClF
ΔE	−3.066	−2.088	−1.627	**−2.844**	**−5.954**	**−12.623**
ΔE_{corr}	−1.733	−1.653	−1.471	**−1.537**	**−3.745**	**−11.184**
BSSE	1.055	0.792	0.689	**0.472**	**0.838**	**1.719**
ΔE^{CP}	−2.011	−1.296	−0.938	**−2.372**	**−5.116**	**−10.903**
ZPVE	1.077	0.763	0.493	**0.639**	**1.348**	**2.133**
$\Delta E^{CP+ZPVE}$	−0.934	−0.533	−0.445	**−1.733**	**−3.768**	**−8.770**
R_{int}	1.9810	2.2475	2.4567	**2.7402**	**2.5438**	**2.2636**

[a] All energies are in kcal/mol; ΔE_{corr} is determined from the difference between MP2 and HF binding energies (not corrected); R_{int} is the intermolecular distance. [b] Numbers in bold are values of the corresponding X−Cl-type complexes.

Table 2. Selected Geometrical Parameters (Å and deg), Frequencies (cm^{-1}), and IR Intensities (km/mol) of Hydrogen-Bonded Complexes Calculated at the MP2(full)/aug-cc-pVTZ Level[a]

			hydrogen-bonded complexes		
parameters	hydrogen-donor monomer	hydrogen-acceptor monomer	F−H···F−Cl	HO−H···F−Cl	H_2N−H···F−Cl
∠X−H···F			173.9	146.0	143.8
R(H···F)			1.9810	2.2475	2.4567
R(X−H)	0.9202 0.9590 1.0098		0.9229	0.9597	1.0099
ΔR(X−H)			0.0027	0.0007	0.0001
freq (X−H)	4135.3 (121.2) 845.7 (5.8) 532.9 (3.1)		4075.0 (349.6)	3841.6 (12.9)	3532.5 (5.9)
Δfreq (X−H)			60.3	4.10	0.40
R(Cl−F)		1.6346	1.6421	1.6392	1.6375
freq (Cl−F)		806.5 (27.8)	797.2 (33.8)	800.7 (31.0)	802.8 (30.6)

[a] X represents the F atom of HF, the O atom of H_2O, and the N atom of NH_3; values in parentheses are IR intensities.

Table 3. Selected Geometrical Parameters (Å and deg), Frequencies (cm^{-1}), and IR Intensities (in km/mol) of X−Cl-type Complexes Calculated at the MP2(full)/aug-cc-pVTZ Level[a]

		complexes with halogen acceptor		
parameters	halogen-donor monomer	FH	OH_2	NH_3
		FCl···XH_n		
∠F−Cl···X		178.1	179.8	180.0
R(Cl···X)		2.7402	2.5438	2.2636
R(Cl−F)	1.6346	1.6396	1.6507	1.7046
freq (Cl−F)	806.5 (27.8)	796.8 (42.3)	768.5 (84.2)	641.8 (227.5)
Δfreq (Cl−F)		9.7	38	164.7

[a] X represents the F atom of HF, the O atom of H_2O, and the N atom of NH_3; values in parentheses are IR intensities.

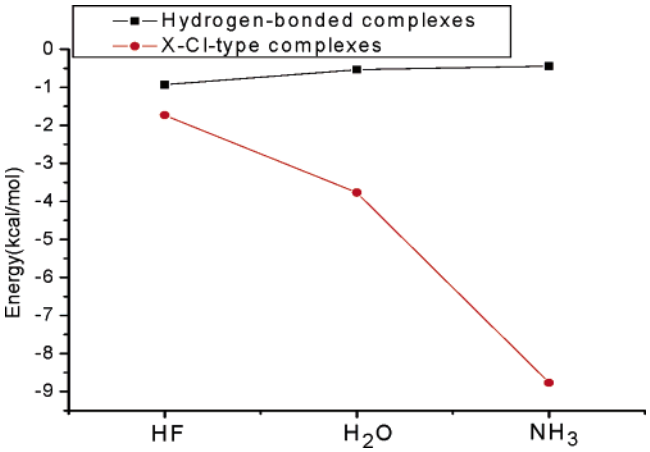


Figure 2. Comparison of the BSSE- and ZPVE-corrected interaction energies [MP2(full)/aug-cc-pVTZ] of all the hydrogen-bonded and X−Cl-type complexes.

1 that the interaction energies of these hydrogen-bonded minimum structures decrease in the order HF > H_2O > NH_3. This order is correlated to the electronegativity of the X atom of the hydride. On the contrary, the interaction energies of these corresponding X−Cl-type minimum structures increase in the order HF < H_2O < NH_3. This order is reasonable because the gas-phase basicity of the halogen atom acceptors is in the same order.

It is interesting to compare the hydrogen-bonded minimum structures (ClF−H_nX) with the corresponding X−Cl-type minimum structures (H_nX−ClF). The BSSE-uncorrected interaction energies (ΔE) of the hydrogen-bonded complex ClF−HF is larger than those for the corresponding X−Cl-type complex HF−ClF. However, the BSSE-corrected interaction energies (ΔE^{CP}) of the X−Cl-type structure HF−ClF are larger than those for the hydrogen-bonded structure ClF−HF by 0.361 kcal/mol. A plot of the ZPVE-corrected interaction energies of these complexes (Figure 2) evaluated at the MP2/(full)/aug-cc-pVTZ level reveals that all of the X−Cl-type complexes are more stable than the corresponding hydrogen-bonded complexes. The calculated interaction

Table 4. Natural Bond Orbital Analysis at the HF/aug-cc-pVTZ//MP2(full)/aug-cc-pVTZ Level (ΔE^2 in kcal/mol, $\Delta\epsilon$ in Hartree, Δq in Au)[a,b]

complexes	donor NBOs	δ	acceptor NBOs	δ	ΔE^2	$\Delta\epsilon$	Δq
ClF−H−F	F lone pair	1.992 (1.996)	H−F o˙ antibond	0.004 (0.000)	3.00	1.50	0.007
ClF−H−OH	F lone pair	1.994 (1.996)	H−O o˙ antibond	0.001 (0.000)	0.88	1.45	0.004
ClF−H−NH_2	F lone pair	1.995 (1.996)	H−N o˙ antibond	0.001 (0.000)	0.42	1.41	0.003
HF−Cl−F	F lone pair	1.993 (1.998)	Cl−F o˙ antibond	**0.005 (0.000)**	**2.57**	**1.07**	**0.004**
H_2O−Cl−F	O lone pair	1.973 (1.996)	Cl−F o˙ antibond	**0.022 (0.000)**	**9.17**	**0.94**	**0.020**
H_3N−Cl−F	N lone pair	1.857 (1.997)	Cl−F o˙ antibond	**0.119 (0.000)**	**48.32**	**0.75**	**0.130**

[a] Numbers in bold are values of the corresponding X−Cl-type complexes. [b] Data in the parentheses are the occupancies of the corresponding NBOs of the isolated molecules. Δq is the amount of charge transfer obtained by the natural population analysis (NPA).

energies of hydrogen-bonded complexes at the BSSE- and ZPVE-corrected levels lie in the −0.445 to −0.934 kcal/mol range, indicating a relatively weak hydrogen bond. In the case of X−Cl-type complexes, the interaction energies lie in the −1.733 to −8.770 kcal/mol range, indicating that the interactions are stronger. The greatest corrected intermolecular interaction in the dimers is −8.770 kcal/mol, belonging to the X−Cl-type complex, H_3N−ClF.

Table 2 shows that there is an elongation of the X−H and the Cl−F bonds upon hydrogen-bonded complex formation. The corresponding harmonic vibrational frequencies are also shown in Table 2. The frequency analysis shows the red-shifting character of the X−H−F interaction. In agreement with the computed X−H bond elongation, the X−H stretching frequencies are lower by 60.3−0.4 cm^{-1} in the complexes than the corresponding frequencies in the monomers. The individual red shift can be correlated directly to the magnitude of X−H bond elongations. In accordance with the interaction energies, the extent of the red shifts is shown to decrease in sequence from HF to NH_3.

Similarly, in the case of X−Cl-type complexes (FCl−XH_n), Table 3 shows that there is an elongation of the Cl−F bond upon complex formation. The corresponding harmonic vibrational frequencies and IR intensities are also shown in Table 3. The frequency analysis reveals the red-shifting character of the FCl−X interactions in the dimers. In agreement with the computed Cl−F bond elongation, the Cl−F stretching frequencies are lowered by 9.7−164.7 cm^{-1} in the complexes than the corresponding frequencies in the monomers. The individual red shift can be correlated directly to the magnitude of Cl−F bond elongations. In accordance with the interaction energies, the extent of the red shifts is shown to increase in sequence from HF to NH_3. As expected, the IR intensities behave in the same way.

3.2. NBO Analysis. For a better understanding of the interaction, a NBO analysis has been carried out at the HF/aug-cc-pVTZ level of theory using MP2(full)/aug-cc-pVTZ geometry. The occupancy (δ) of frontier molecular orbitals involving the CT between subsystems, the second-order perturbation energy lowering (ΔE^2) due to the interaction of donor and acceptor orbitals, and the difference ($\Delta\epsilon$) of energies between acceptor and donor NBOs, provided by NBO analysis, are collected in Table 4.

Table 5. Mulliken Charge Transfer from H_nX to ClF at the MP2/aug-cc-pVTZ and HF/aug-cc-pVTZ Level

complexes	Q_{CT} (au) MP2/aug-cc-PVTZ	Q_{CT} (au) HF/aug-cc-pVTZ
ClF···H−F	0.021	0.021
ClF···H−OH	0.014	0.014
ClF···H−NH_2	0.004	0.004
HF···Cl−F	0.009	0.009
H_2O···Cl−F	0.010	0.011
H_3N···Cl−F	0.112	0.115

Let us first repeat that the formation of a hydrogen-bonded complex involves CT from the proton acceptor to the proton donor. This results in the increase of electron density in the H−X antibonding orbitals of the proton donor. For the X−Cl-type, the case is a little similar.[42] The CT from the lone pairs of the electron donor in the halogen atom acceptor is mainly directed to the Cl−F antibonding orbitals of the halogen atom donor too. Because the CT accompanies the formation of hydrogen bonds or X−Cl bonds and plays a major role in it, ΔE^2 can be taken as an index to judge the strength of hydrogen bonds or X−Cl bonds. As can been see from Table 4, in the case of hydrogen-bonded complexes, the CT stabilization energies are computed to be 3.00, 0.88, and 0.42 kcal/mol for n(F)−o˙*(H−F), n(F)−o˙*(H−O), and n(F)−o˙*(H−N) interactions, respectively, which are comparable in magnitude to their interaction energies. For the X−Cl-type complexes, the charge-transfer stabilization energies are computed to be 2.57, 9.17, and 48.32 kcal/mol for n(F)−o˙*(Cl−F), n(O)−o˙*(Cl−F), and n(N)−o˙*(Cl−F) interactions, respectively, which are also comparable in magnitude to their interaction energies. The larger CT stabilization energies for the N−Cl-type complex H_3N−ClF confirm that it is a strong and partly covalent complex.

The result of the Mulliken charge transfer (Q_{CT}) is presented in Table 5. It is worth mentioning that the Mulliken charge transfer from H_nX to ClF takes place for all of the complexes. The amount of the Mulliken transferred charge is approximately equal at the MP2/aug-cc-pvTZ and HF/aug-cc-pVTZ levels.

3.3. SAPT Studies. To gain more insight the nature of the interaction, we further performed SAPT to analyze the interaction energy in terms of physically meaningful com-

Table 6. Partition of the Energies (kcal/mol) Derived from SAPT Calculations in Electrostatic (E_{elst}), Exchange (E_{exch}), Inductive (E_{ind}), and Dispersive (E_{disp}) Energies, as Defined by eqs 2−5[a,b]

	hydrogen-bond complexes			X−Cl-type complexes		
	ClF···HF	ClF···H_2O	ClF···H_3N	FCl···FH	FCl···OH_2	FCl···NH_3
E_{elst}	−2.456	−0.903	−0.926	**−3.404**	**−10.371**	**−36.666**
E_{exch}	2.539	0.963	1.257	**3.245**	**11.104**	**45.387**
E_{ind}	−1.072	−0.211	−0.169	**−0.602**	**−2.142**	**−12.199**
E_{disp}	−1.058	−1.120	−1.136	**−1.644**	**−3.740**	**−9.393**
E_{int}^{SAPT2}	−2.051	−1.271	−0.971	**−2.406**	**−5.148**	**−12.870**
$E_{int}^{(MP2)a}$	−2.011	−1.296	−0.938	**−2.372**	**−5.116**	**−10.903**

[a] MP2(full)/aug-cc-pVTZ counterpoise-corrected interaction energies. [b] Numbers in bold are values of the corresponding X−Cl-type complexes. All energies are in kcal/mol.

ponents such as electrostatic, induction, dispersion, and exchange energies.

The electrostatic component of the interaction energy is represented here by the sum of $E_{elst}^{(10)}$ and $E_{elst,resp}^{(12)}$. The induction contribution to the interaction energy is mainly contained in $E_{ind,resp}^{(20)}$. This is a second-order energy correction that results from the distortion of the charge distribution of one monomer by the electrostatic charge distribution of another monomer, and vice versa. This mutual polarization of the monomer by the static electric field of the other is the polarizabilities of the monomers. The leading intramonomer correlation contribution is concluded in ${}^tE_{ind}^{(22)}$ and accounts for only 2% of the induction energy. The attractive part of the induction energy is substantially quenched by the repulsive exchange-induction energy (represented by $E_{exch-ind,resp}^{(20)}$ and ${}^tE_{exch-ind}^{(22)}$). As noted by Jeziorski and co-workers,[33] any quantitatively accurate calculation of the induction energy cannot neglect the exchange-induction contribution. To simplify the analysis, for the present purposes, we have designated the exchange-dispersion and exchange-induction terms as dispersion and induction, respectively. So, the induction energy terms calculated here are $E_{ind,resp}^{(20)}$, ${}^tE_{ind}^{(22)}$, $E_{exch-ind,resp}^{(20)}$, and ${}^tE_{exch-ind}^{(22)}$. The dispersion energy is represented here by the sum of $E_{disp}^{(20)}$ and $E_{exch-disp}^{(20)}$, where $E_{disp}^{(20)}$ is the second-order dispersion energy, $E_{exch-disp}^{(20)}$ standing for the second-order correction for a coupling between the exchange repulsion and the dispersion energy. The exchange energy terms calculated here are $E_{exch}^{(10)}$, $E_{exch}^{(11)}$, $E_{exch}^{(12)}$, and $\delta E_{int,resp}^{HF}$. $E_{exch}^{(10)}$ accounts for the repulsion due to the Pauli exclusion principle and arises from the antisymmetry requirement of the wave function; $E_{exch}^{(11)}$ and $E_{exch}^{(12)}$ account for the effects of intramonomer correlation on the exchange repulsion. Therefore, The SAPT interaction energy, E_{int}, is given by

$$E_{int} = E_{elst} + E_{ind} + E_{disp} + E_{exch} \quad (1)$$

where

$$E_{elst} = E_{elst}^{(10)} + E_{elst,resp}^{(12)} \quad (2)$$

The SAPT-derived components of the interaction energy

$$E_{exch} = E_{exch}^{(10)} + E_{exch}^{(11)} + E_{exch}^{(12)} + \delta E_{int,resp}^{HF} \quad (3)$$

$$E_{ind} = E_{ind,resp}^{(20)} + {}^tE_{ind}^{(22)} + E_{exch-ind,resp}^{(20)} + {}^tE_{exch-ind}^{(22)} \quad (4)$$

$$E_{disp} = E_{exch-disp}^{(20)} + E_{disp}^{(20)} \quad (5)$$

are summarized in Table 6. It can be seen that the results of SAPT2 (E_{int}^{SAPT2}) are in good agreement with the results obtained at the MP2(full)/aug-cc-PVTZ level, suggesting that SAPT2/aug-cc-PVTZ is a proper method to study the intermolecular interactions in these studied complexes.

To elucidate the role of each term in the total interaction energy, we have plotted the magnitude of the individual interaction energy component (see Figure 3). Table 6 and Figure 3 show that the exchange energy is larger than the absolute value of the electrostatic energy for the three weak hydrogen-bonded complexes. The induction energy term and the dispersion energy term are nearly equal for the complex ClF−HF. For the ClF−H_2O and ClF−H_3N complexes, the induction energy is rather small, and the dispersion energy dominates. All of this is opposite of the decomposition for normal hydrogen bonds, which are known to be electrostatic in nature. Therefore, this type of interaction cannot be called conventional hydrogen bonding. The interaction may be considered as unconventional hydrogen bonding.

On the other hand, in the case of the X−Cl-type complexes, we find the electrostatic energies dramatically increase as we progress from HF to NH_3. The electrostatic interaction pulls the two monomers close to each other, which results in large values of all other energy components. For H_2O−ClF and H_3N−ClF, the exchange energy becomes even larger than the absolute value of the electrostatic energy, but the large attractive induction and dispersion energies lead to a very strong interaction. In the strongest X−Cl-type complex, H_3N−ClF, the induction energy is the most important attractive term, followed by the exchange and electrostatic energy, and is responsible for stabilizing the complex. This domination of the induction and exchange terms is the main feature of the strong and partly covalent bonds.

Additionally, the induction energy can be described to result from the interaction between the highest occupied molecular orbital and the lowest unoccupied molecular orbital. The inductive type of molecular orbital interaction can also be correlated to the extent of CT. It would be interesting to examine if the extent of charge transfer can be correlated to the total induction energy. We have carried out an analysis, using the charge transfer obtained by the natural population analysis (NPA) (Δq; Table 4). It can be seen from Figures 4 and 5 that, in the weak hydrogen-bonded

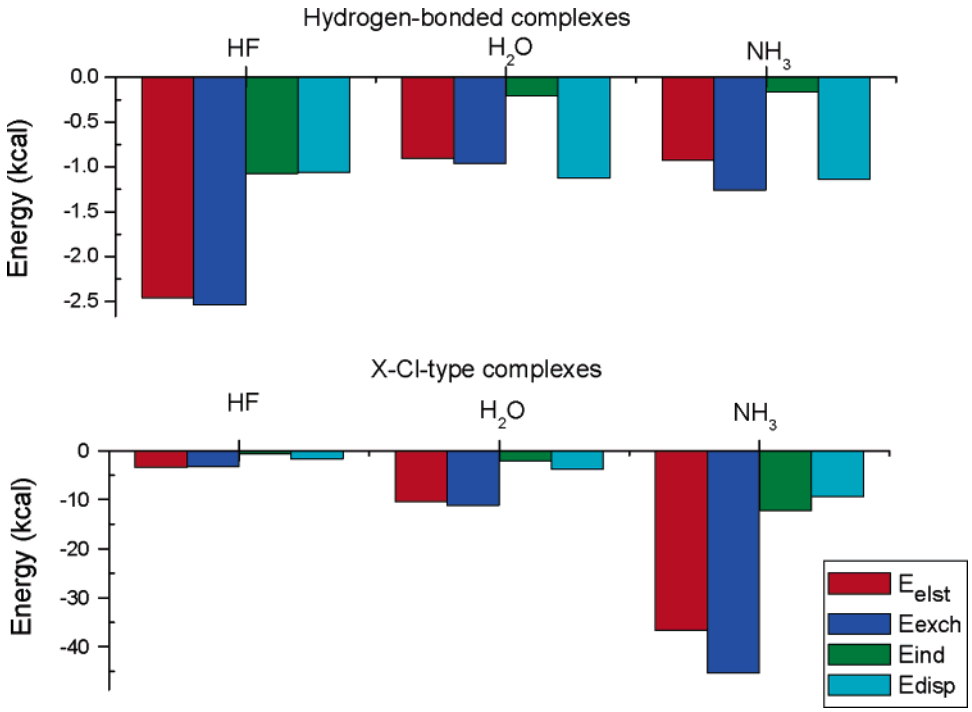


Figure 3. Bar plots of SAPT interaction energy components (E_{elst}, E_{ind}, E_{disp}, and E_{exch}) for the hydrogen-bonded complexes and X−Cl-type complexes according to eqs 1−5.

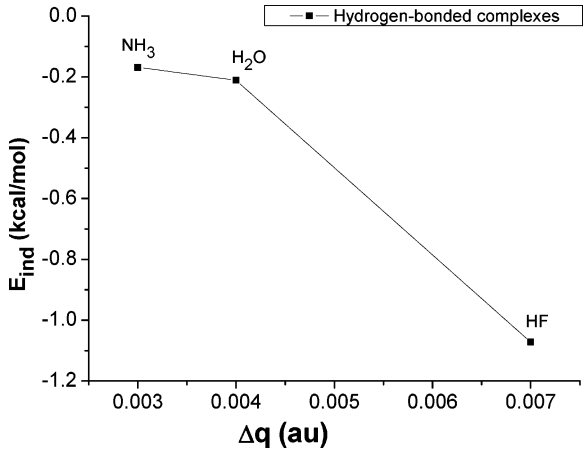


Figure 4. Correlation of the electronic charge transferred from ClF to the hydride with the total induction energy in the hydrogen-bonded complexes.

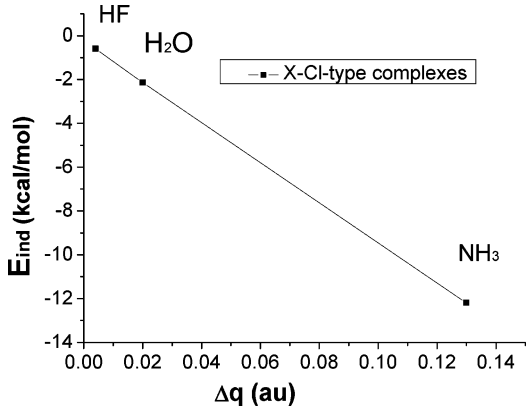


Figure 5. Correlation of the electronic charge transferred from the hydride to the ClF with the total induction energy in the X−Cl-type complexes.

complexes, a decrease in the magnitude of the charge transfer (Δq) from HF to NH_3 leads to a less induction energies. In the case of the X−Cl-type complexes, the charge-transfer correlates well with the total induction energies.

4. Conclusions

In summary, we have systematically investigated the interaction of the first-row hydrides with ClF. We also focused on the decomposition of the interaction energy into physically meaningful terms and on the basis of the parameters of molecular properties characterized and described the intermolecular interaction. Our results can be summarized as follows:

1. The total interaction energies of the X−Cl-type complexes are more stable than the corresponding weak hydrogen-bonded complexes. The charge-transfer analysis discloses the different physical properties of the weak hydrogen bond and X−Cl bond.

2. For all the investigated weak hydrogen-bonded complexes, the exchange energy outweighs the electrostatic energy, which is opposite of the decomposition for normal hydrogen bonds. This type of interaction may be considered as unconventional hydrogen bonding. The induction energy term and the dispersion energy term are nearly equal for the complex ClF−HF. For the ClF−H_2O and ClF−H_3N complexes, the induction energy is rather small, and the dispersion energy dominates.

3. For the X−Cl-type complex HF−ClF, the main interaction energy comes from the electrostatic energy, which slightly outweighs the exchange term. For the two strongest X−Cl-type complexes, H_2O−ClF and H_3N−ClF, the exchange energy outweighs the electrostatic energy but the large attractive induction and dispersion energies lead to a

very strong interaction. In the strongest N−Cl-type complex, H_3N−ClF, the induction energy is the most important attractive term, followed by the exchange and electrostatic energy, and is responsible for stabilizing the complex. This domination of the induction and exchange terms is the main feature of the strong and partly covalent bonds. This is also consistent with significant charge redistribution.

Acknowledgment. We are grateful to the Special Research Fund for the Doctoral Program of Higher Education (20050010014) for financial support. Authors would also like to thank Professor Krzysztof Szalewicz for providing the SAPT2002 program.

References

(1) Lehn, J. M. *Supramolecular Chemistry. Concepts and Perspectives*; VCH: Weinheim, Germany, 1995.

(2) Desiraju, G. R. *Crystal Engineering: The Design of Organic Solids*; Elsevier: Amsterdam, The Netherlands, 1989.

(3) Desiraju, G. R.; Steiner, T. *The Weak Hydrogen Bond In Structural Chemistry and Biology*; Oxford University Press: New York, 1997.

(4) Nishio, M.; Hirota, M.; Umezawa, Y. *The CH-Interaction*; Wiley-VCH: New York, 1998.

(5) Scheiner, S. *Molecular Interactions*: *From Van der Waals to Strong Bound Complexes*; Wiley: Chichester, U. K., 1997.

(6) Jeffrey, J. A.; Saenger, W. *Hydrogen Bonding in Biological Structures*; Springer-Verlag: Berlin, 1991.

(7) Umeyama, H.; Morokuma, K.; Yamabe, S. *J. Am. Chem. Soc.* **1977**, *99*, 330−343.

(8) Kollman, P.; Dearling, A.; Kochanski, E. *J. Phys. Chem.* **1982**, *86*, 1607−1614.

(9) Røeggen, I.; Dahl, T. *J. Am. Chem. Soc.* **1992**, *114*, 511−516.

(10) Price, S. L.; Stone, A. J.; Lucas, J.; Rowland, R. S.; Thornley, A. E. *J. Am. Chem. Soc.* **1994**, *116*, 4910−4918.

(11) (a) Legon, A. C.; Lister, D. G.; Thorn, J. C. *J. Chem. Soc. Chem. Commun.* **1994**, 757−758. (b) Legon, A. C.; Lister, D. G.; Thorn, J. C. *J. Chem. Soc., Faraday Trans.* **1994**, *90*, 3205−3212. (c) Bloemink, H. I.; Legon, A. C.; Thorn, J. C. *J. Chem. Soc., Faraday Trans.* **1994**, *90*, 781−787. (d) Legon, A. C. *Chem.—Eur. J.* **1998**, *4*, 1890−1897.

(12) Desiraju, G. R. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2311−2327.

(13) Latajka, Z.; Berski, S. *THEOCHEM* **1996**, *371*, 11−16.

(14) Ruiz, E.; Salahub, D. R.; Vela, A. *J. Phys. Chem.* **1996**, *100*, 12265−12276.

(15) Lommerse, J. P. M.; Stone, A. J.; Taylor, R.; Allen, F. H. *J. Am. Chem. Soc.* **1996**, *118*, 3108−3116.

(16) Bürger, H. *Angew. Chem., Int. Ed. Engl.* **1997**, *36*, 718−721.

(17) Zhang, Y.; Zhao, C.-Y.; You, X.-Z. *J. Phys. Chem. A.* **1997**, *101*, 2879−2885.

(18) Alkorta, I.; Rozas, I.; Elguero, J. *J. Phys. Chem. A.* **1998**, *102*, 9278−9285.

(19) Amico, V.; Meille, S. V.; Corradi, E.; Messina, M. T.; Resnati, G. *J. Am. Chem. Soc.* **1998**, *120*, 8261−8262.

(20) Farina, A.; Meille, S. V.; Messina, M. T.; Metrangolo, P.; Resnati, G.; Vecchio, G. *Angew. Chem., Int. Ed.* **1999**, *38*, 2433−2436.

(21) Legon, A. C., *Angew. Chem., Int. Ed.* **1999**, *38*, 2686−2714.

(22) Corradi, E.; Meille, S. V.; Messina, M. T.; Metrangolo, P.; Resnati, G. *Angew. Chem., Int. Ed.* **2000**, *39*, 1782−1786.

(23) Valerio, G.; Raos, G.; Meille, S. V.; Metrangolo, P.; Resnati, G. *J. Phys. Chem. A* **2000**, *104*, 1617−1620.

(24) Karpfen, A. *J. Phys. Chem. A* **2000**, *104*, 6871−6879.

(25) Walsh, R. B.; Clifford, W.; Padgett, C. W.; Metrangolo, P.; Resnati, G.; Hanks, T. W.; Pennington, W. T. *Cryst. Growth Des.* **2001**, *1*, 165−175.

(26) Metrangolo, P.; Resnati, G. *Chem.—Eur. J.* **2001**, *7*, 2511−2519.

(27) Romaniello, P.; Lelj, F. J. *J. Phys. Chem. A.* **2002**, *106*, 9114−9119.

(28) Nangia, A. *CrystEngComm.* **2002**, *17*, 1−6.

(29) Burton, D. D.; Fontana, F.; Metrangolo, P.; Pilatid, T.; Resnatic, G. *Tetrahedron Lett.* **2003**, *44*, 645−648.

(30) Wang, W. Z.; Wong, N. B.; Zheng, W. X.; Tian, A. M. *J. Phys. Chem. A* **2004**, *108*, 1799−1805.

(31) Reed, A. E.; Weinbold, F.; Curtiss, L. A.; Pochatko, D. J. *J. Chem. Phys.* **1986**, *84*, 5687−5705.

(32) (a) Gu, Y.; Kar, T.; Scheiner, S. *J. Am. Chem. Soc.* **1999**, *121*, 9411−9422. (b) Hobza, P.; Havlas, Z. *Chem. Rev.* **2000**, *100*, 4253−4264. (c) Hobza, P.; Havlas, Z. *Chem. Phys. Lett.* **1999**, *303*, 447−452. (d) Hermansson, K. *J. Phys. Chem. A* **2002**, *106*, 4695−4702. (e) Qian, W.; Krimm, S. *J. Phys. Chem. A* **2002**, *106*, 6628−6636. (f) Alabugin, I. V.; Manoharan, M.; Peabody, S.; Weinhold, F. *J. Am. Chem. Soc.* **2003**, *125*, 5973−5987.

(33) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887−1930.

(34) Jeziorski, B.; Moszynski, R.; Ratkiewicz, A.; Rybak, S.; Szalewicz, K.; Williams, H. L. In *Methods and Techniques in Computational Chemistry*: *METECC-94*; Clementi, E., Ed.; STEF: Cagliari, Italy, 1993; Vol. B, Medium Sized Systems, pp 79−129.

(35) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618−622.

(36) Hobza, P.; S̈yponer, *J. Chem. Rev.* **1999**, *99*, 3247−3276.

(37) (a) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007−1023. (b) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358−1364. It is well-known that the aug-cc-pVXZ series of basis sets was designed for frozen-core MP2 calculations. Here, we employed the aug-cc-pVTZ basis set for fully correlated calculations because our test calculations clearly showed that the values of the counterpoise-corrected binding energies obtained using fully correlated MP2 calculations are more close to their CBS limit values than those obtained using frozen-core MP2 calculations.

(38) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553−556.

(39) Simon, S.; Duran, M.; Dannenberg, J. J. *J. Chem. Phys.* **1996**, *105*, 11024−11031.

(40) Hobza, P.; Havlas, Z. *Theor. Chem. Acc.* **1998**, *99*, 372−377.

(41) Simon, S.; Duran, M.; Dannenberg, J. J. *J. Phys. Chem. A.* **1999**, *103*, 1640−1643.

(42) Karpfen, A. *J. Phys. Chem. A* **2000**, *104*, 6871−6879.

(43) (a) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735−746. (b) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 1736−1740. (c) Reed, A. E.; Curitss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 889−926.

(44) 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.; Pople, J. A. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(45) Lide, D. R. CRC *Handbook of Chemistry and Physics*, 73rd ed.; CRC Press: Boca Raton, FL, 1992.

(46) Williams, H. L.; Szalewicz, K.; Jeziorski, B.; Moszynski, R.; Rybak, S. *J. Chem. Phys.* **1994**, *98*, 1279−1291.

(47) (a) Moszynski, R.; Jeziorski, B.; Ratkiewicz, A.; Rybak, S. *J. Chem. Phys.* **1993**, *99*, 8856−8869. (b) Moszynski, R.; Jeziorski, B.; Rybak, S. *J. Chem. Phys.* **1994**, *100*, 1312−1325. (c) Moszynski, R.; Jeziorski, B.; Rybak, S.; Szalewicz, K.; Williams, H. L. *J. Chem. Phys.* **1994**, *100*, 5080−5092. (d) Rybak, S.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **1995**, *95*, 6576−6610. (e) Moszynski, R.; Cybulski, S. M.; Chalasinski, G. *J. Chem. Phys.* **1994**, *100*, 4998−5010. (f) Chalasinski, G.; Jeziorski, B. *Theor. Chim. Acta* **1977**, *46*, 277−290. (g) Moszynski, R.; Korona, T.; Wormer, P. E. S.; van der Avoird, A. *J. Phys. Chem. A* **1997**, *101*, 4690−4698.

(48) Szalewicz, K.; Cole, S. J.; Kolos, W.; Bartlett, R. J. *J. Chem. Phys.* **1988**, *89*, 3662−3673.

(49) Bukowski, R.; Szalewicz, K.; Chabalowski, C. *J. Phys. Chem. A* **1999**, *103*, 7322−7340.

(50) Milet, A.; Moszynski, R.; Wormer, P. E. S.; van der Avoird, A. *J. Phys. Chem. A* **1999**, *103*, 6811−6819.

(51) Wormer, P. E. S.; van der Avoird, A. *Chem. Rev.* **2000**, *100*, 4109−4143.

(52) Kim, K. S.; Tarakeshwar, P.; Lee, J. Y. *Chem. Rev.* **2000**, *100*, 4145−4185.

(53) Bukowski, R.; Cencek, W.; Jankowski, P.; Jeziorski, B.; Jeziorska, M.; Kucharski, S. A.; Misquitta, A. J.; Moszynski, R.; Patkowski, K.; Rybak, S.; Szalewicz, K.; Williams, H. L.; Wormer, P. E. S. *SAPT2002: An Ab Initio Program for Many-Body Symmetry-Adapted Perturbation Theory Calculations of Intermolecular Interaction Energies. Sequential and Parallel Versions, 2003*.

CT600229N

JCTC Journal of Chemical Theory and Computation

Quantum Calculations on Hydrogen Bonds in Certain Water Clusters Show Cooperative Effects

Vasiliy S. Znamenskiy and Michael E. Green*

Department of Chemistry, City College of the City University of New York, 138th St. and Convent Ave, New York, New York 10031

Received April 13, 2006

Abstract: Water molecules in clefts and small clusters are in a significantly different environment than those in bulk water. We have carried out ab initio calculations that demonstrate this in a series of clusters, showing that cooperative effects must be taken into account in the treatment of hydrogen bonds and water clusters in such bounded systems. Hydrogen bonds between water molecules in simulations are treated most frequently by using point-charge water potentials, such as TIP3P or SPC, sometimes with a polarizable extension. These produce excellent results in bulk water, for which they are calibrated. Clefts are different from bulk; it is necessary to look at smaller systems and investigate the effect of limited numbers of neighbors. We start with a study of isolated clusters of water with varying numbers of neighbors of a hydrogen-bonded pair of water molecules. The cluster as a whole is in a vacuum. The clusters are defined so as to provide the possible arrangements of nearest neighbors of a central hydrogen-bonded pair of water molecules. We then scan the length and angles of the central hydrogen bond of the clusters, using density functional theory, for each possible arrangement of donor and acceptor hydrogen bonds on the central hydrogen-bonding pair; the potential of interaction of two water molecules varies with the number of donor and acceptor neighbors. This also involves changes in charge on the water molecules as a function of bond length and changes in energy and length as a function of the number of neighboring donor and acceptor molecules. The energy varies by approximately 6 $k_B T$ near room temperature from the highest to the lowest energy when bond length alone is varied, enough to seriously affect simulations.

Introduction

The simulation of systems containing water, whether by molecular dynamics (MD) or Monte Carlo (MC), requires an interaction potential for the water molecules. Of the several successful point-charge water models in the literature, it appears that TIP3P (plus TIP4P and TIP5P[1−5] and more recently TIP4P-Ew) and SPC and SPC/E (and some of their descendents)[6,7] are most popular. There are some polarizable versions of these (e.g., polarizable SPC, or PSPC[8]), but these are more complex and difficult to use in simulations. For water clusters, excellent results can be obtained with the TTM2-F potential,[9−11] which is, however, also difficult to use in an MD simulation. The simpler versions allow much faster simulations that include explicit water. These models are built from point charges on or near the oxygen and hydrogens, plus van der Waals terms. They are insensitive to their surroundings; the same charges are used in all cases (Gaussian 03 now allows some variation, but this has not been much used yet). A new potential, based on quantum calculations with fitting to water, has very recently been proposed,[12] and it appears that it may be better than the point-charge models, although it has yet to be applied. Relatively early studies pointing out types of cooperativity include a study of water trimers, in which nonadditivity was calculated.[13] Several other studies demonstrated the nonadditive properties of specific, sometimes optimized, small rings or linear clusters. There are a number of studies of larger

* Corresponding author phone: (212) 650-6034; fax: (212) 650-6107; e-mail: green@sci.ccny.cuny.edu.

10.1021/ct600139d CCC: $37.00
Published on Web 12/08/2006

polyhedral water clusters as well: 20-mers were studied by Kuo et al.[14] Hexamers were evaluated by Tissandier et al.,[15] while Tsai and Jordan considered cubic clusters of size 8, 12, 16, and 20.[16] Anick studied more general polyhedral water clusters, finding cooperative effects in energy and bond length for equilibrium clusters that extended to next nearest neighbors.[17,18] Hodges et al.[19] compared ab initio results to two potentials (ASP-W2 and ASP-W4, which do not appear to be widely used at present) for some small clusters. A related study of trimers, tetramers, and pentamers estimated that omitting cooperative effects would lead to errors exceeding 20% for clusters larger than pentamers (although these were not explicitly calculated).[20] Ojamae and Hermansson[21] calculated frequencies and binding energies for water chains as well as one ring structure and a pentamer. Chen and Gordon broke down the interactions into several energy components (electrostatic, exchange, polarization, and charge transfer) in a paper largely concerned with techniques of calculation.[22] In general, these papers use optimized clusters, without studying the dependence of energy on bond lengths and angles. They do, however, find the dependence of optimized bond length and energy on geometry and bond topology. There is also experimental evidence for cooperativity in hydrogen bonding from IR spectra. Going beyond Zundel's work, Luck et al.[23] have shown shifts in IR spectra for hydrogen bonding to bases, principally demonstrating anti-cooperativity. New potentials,[24] or steps toward new potentials,[25] continue to appear. Typically these are based on quantum calculations, whether Empirical Valence Band or, more commonly, density functional, calculations. Some compare several density functional or alternate methods, principally for accuracy in describing water.[26−28]

Point-charge models do a very satisfactory job of reproducing the properties of bulk water, for which the parameters of the model are calibrated.[4,5,8,29,30] It is less certain that they do a satisfactory job of reproducing the properties of water in cases in which the water-molecule interactions cannot be averaged as in bulk. Polarizable models are not useful for small clusters as they are calibrated for the average electric field in bulk water, which can be very different in a small cluster. Other difficulties often arise when new properties, such as phase changes, are being simulated. We are particularly concerned with clefts in proteins in this context. The importance of water in such clefts is suggested by recent work by Anishkin and Sukharev[31] and by Beckstein and Sansom,[32] in which water in an ion channel cleft shows behavior that is perhaps surprising, as it leaves a "vapor phase" with a water surface inside a hydrophobic part of a protein. Here, a small number of water−water interactions with a near-vacuum surface might be a reasonably correct representation of the system. An accurate description of such systems requires the inclusion of cooperative interactions among the water molecules of the cluster, which in turn needs explicit representation of the forms of hydrogen bonding among the molecules. Point-charge models, by their definition, do not change when the surroundings change. Only with Car−Parrinello quantum MD (CPMD) is it possible to take such effects into account or to allow a proton to transfer to another molecule. However, CPMD requires such large computer resources that it cannot simulate systems over about 200 atoms. Proton transfer is required for a wide range of important biological processes. A procedure equivalent to Car−Parrinello has been used for the species $H_5O_2^+$ and $H_3O_2^-$.[33] The study of cooperative interactions in water clusters is of sufficient interest that work continues to appear on this subject; the eventual result will be a potential that is simple enough to use in a simulation but still includes the cooperative effects.

Calibration of a hydrogen-bond potential will be aided by spectroscopy. In order to reach this goal, it will be necessary to study not only clusters, whether optimized or not, but to move to studies as a function of distance and angle. Our procedure is a step toward a usable hydrogen-bond model. We also show how the energy of several types of donor and acceptor arrangements varies with bond-distance and -angle.

Surfaces are another type of system in which these effects are important, and surfaces bounding aqueous solutions in turn are important for most of biology. Hydrogen bonds in high electric fields require still more detailed treatment,[34] but even at low fields, calculations of surfaces give results that are not always in agreement with experimental results. Experimental results may show that the H-bond network at a surface fails to survive intact. In fact, there is a quite large amount of literature on water at various types of surfaces, especially metal surfaces, showing water that has increased (usually) density, and other rearrangements, sometimes in the first layer only, sometimes for as many as three layers. In some cases, as with brushite, there are two separate arrangements of water at the surface.[35] Even simulations using traditional potentials show differences in composition of surface layers, compared to bulk.[36] Our earlier work on MC simulations of water in pores, which included high electric fields and polarizable potentials, also showed changes in ordering and density [37−39.] Somewhat similar results were obtained by Guidoni et al.[40] Other groups have shown various effects at surfaces, including biologically interesting surfaces. Experimentally, for example, α-hemolysin, a large bacterial channel, changes its permeability with pH, apparently because, when the channel is charged, water is strongly hydrogen-bonded to the channel wall.[41] Experimental evidence exists for unusual behavior of water near surfaces, and standard simulations do not always adequately reproduce the experimental behavior. The importance of the solvent environment for the stabilization and conformation of peptide function is emphasized in a recent simulation study by Johnston et al. of three different environments of channel-forming peptides.[42] In this study, classical potentials are used, but the structure of the peptides depends strongly on the solvation. Changing the potential is equivalent to changing the solvent, as the solvent is included in the simulation in the form of its potential. The changes we discuss here may be more subtle than those of that simulation but still make a significant difference. Earlier work by Scheiner also considered many aspects of hydrogen bonding, including proton transfer along water wires (not as simulation). Much of this was reviewed by Scheiner in the context of kinetic isotope effects.[43] He has also considered cationic oligomers of water.[44] Most relevant is a study primarily directed at

methyl substitution, in which the electron density was calculated, and the Mulliken electrical charges on atoms obtained and related to the hydrogen bonds.[45] Much of this work was aimed at answering somewhat similar questions to those we address here, although the use of Mulliken charges suggests that the quantitative results may be of limited accuracy. We are concerned with the relation of neighboring water molecule hydrogen-bond donor and acceptor molecules to the energy, charge, and other properties of the hydrogen bonds. The problem does not appear to have been previously treated from this point of view. Cooperativity or nonpairwise additivity in clusters has been shown to be important for other properties, including simulations of the liquid state,[46] and infrared spectra.[47] It is of interest to note that Kirchner[46] has demonstrated that B3LYP does as well as MP2, provided a good basis set is used, and does better (or less poorly), if a smaller basis set is used.

In recent work, we found that the hydrogen bonds of a system of four acetic acid molecules and six waters had rather simple and surprising bond-length versus electron density, and bond length versus energy, relations.[48] This time, we seek to understand more than the charge density at the bond critical point (bcp) that is defined in the atoms in molecules (AIM) method of Bader.[49] We previously focused primarily on the bond properties that could be described in terms of the electron density and did not examine the variation in the bond energy, our first property of interest.

We have tested clusters consisting of water molecule pairs with zero to six neighbors, using density functional calculations. The central hydrogen bond in each pair has a certain set of properties (energy, charge, bond length, and electron density in the bond). Furthermore, the charge distribution on the atoms of the water molecules, and the dipole moments, are of interest in themselves. These properties of the bond and of the molecules also change as additional water molecules hydrogen-bond to the original pair.

We have studied all 36 possible sets of neighbors as a function of the oxygen−oxygen distance of the two central molecules to determine these properties with reasonable accuracy (when two donors, or two acceptors, appear on a single central molecule, they are counted as part of a single set, because they are equivalent by symmetry; otherwise, there would be 2^6, or 64, sets; the correct count of 36 is given in the Methods section). There are significant differences in the equilibrium bond distance, and even larger differences in the energy of the hydrogen bond. The energy differences would matter greatly in MD or MC simulations. Short hydrogen bonds are stronger and have some covalent character; then, the hydrogen-bonded pair shows a slightly polar character, in that a small charge is transferred between donor and acceptor molecules. The fact that shorter bonds are stronger is not a new finding; there is a substantial amount of literature referring to short, strong hydrogen bonds.[50−52] They have also been studied experimentally. Here, we have studied H-bond properties using several tools: AIM (Bader[49]), Natural Bond Orbitals (NBO[53]), and the energy and geometry of the bonds. In addition to the nonadditivity of energy, intermolecular charge transfer (something not normally considered for neutral clusters), bond length, and intramolecular charge distribution depend on the same factors.

Methods

1. Clusters. The clusters were set up with a central pair, donor and acceptor, of water molecules, with a hydrogen atom from one along the line connecting the two oxygen atoms. The $O_a-O_d-H_d$ (sic) (a = acceptor, d = donor) angle is approximately 3.4°, with the H_d being the hydrogen in the hydrogen bond from the acceptor. This pair was optimized under the following constraints: (i) all four OH covalent bonds were the same length, (ii) the two intramolecular HOH angles were equal, and (iii) the dihedral angle defined by the acceptor water oxygen and the three atoms of the donor molecule was set equal to zero.

This pair, optimized with constraints i−iii, formed the template for the neighbor pairs, which were added to form the other clusters. The same OH bond distance and water molecule angles were used; there could be a maximum of six "neighbors", two donors and two acceptors each, with the central pair forming each other's first donor or acceptor. Each "neighbor" was aligned with respect to one of the original molecules at a distance and angle that was appropriate to an isolated pair, using the constrained optimized results.

2. Cluster Definition. The clusters were defined by the number of donor and acceptor hydrogen bonds formed by each of the two central water molecules. The donor of the central hydrogen bond can donate one more hydrogen bond and have up to two acceptor bonds. The acceptor of the central bond can be the donor for zero, one, or two bonds, and the acceptor for one or two (the first bond from the other water in the pair counts as one, so zero is not possible). Thus, an isolated pair can be designated 1001, where the first 1 indicates the number of donor hydrogen bonds of the donor (there being only the one donor hydrogen bond in the water pair). There is a 0 in the second place, as there are no acceptor hydrogen bonds on that oxygen (no other water molecules are available to be the donors). Similarly, the acceptor of the pair is not a donor to any acceptor, so the third digit is 0. The fourth digit is 1, as the acceptor of the pair accepts the one central hydrogen bond, and only that one. Added "neighbor" molecules can be donors or acceptors, and the four-digit index is adjusted accordingly. For example, the four-digit index 1221 means that the donor water oxygen donates only the single hydrogen bond to the other central water (first digit), but it accepts two from other water molecules (second digit); the acceptor molecule donates two to other water molecules (third digit) but has only the one accepted hydrogen bond (fourth digit).

This notation is another way to see that there are a total of 36 possible sets, as mentioned previously (the first and last digits can be 1 or 2 and the two middle digits 0, 1, or 2, for a total of $2 \times 3 \times 3 \times 2 = 36$ possibilities). A similar notation has been introduced by Ohno et al.[47]

Clusters of two types, constrained and unconstrained, are used in the calculation. We define the two types of clusters in Table 1, Figure 1, and a later paragraph. In the end, we

Table 1. **Z** Matrix Showing Bond Lengths, Angles, and Dihedrals for the 2222 Cluster; Atom Labels Are as Shown in Figure 1[a]

water		**Z** matrix						
A: Initial Pair **Z** Matrix								
1	acceptor of central hydrogen bond	O_1						
		H_2	O_1	R_1				
		H_3	O_1	R_1	H_2	a_1		
5	donor of central hydrogen bond	O_{13}	O_1	oo	H_3	a_3	H_2	d_3
		H_{14}	O_{13}	R_1	O_1	a_2	H_3	d_2
		H_{15}	O_{13}	R_1	H_{14}	a_1	O_1	d_4
B: Other Water Clusters[b]								
1	acceptor of central hydrogen bond	O_1						
		H_2	O_1	r_1				
		H_3	O_1	r_1	H_2	a_1		
2	neighbor: acceptor for first water	O_4	O_1	oo	H_2	a_2	H_3	d_4
		H_5	O_4	r_1	O_1	a_3	H_2	$-d_2$
		H_6	O_4	r_1	H_5	a_1	O_1	d_3
3	neighbor: acceptor for first water	O_7	O_1	oo	H_3	a_2	H_2	d_4
		H_8	O_7	r_1	O_1	a_3	H_3	$-d_2$
		H_9	O_7	r_1	H_8	a_1	O_1	d_3
4	neighbor: donor for first water	O_{10}	O_1	oo	H_2	a_3	H_3	d_3
		H_{11}	O_{10}	r_1	O_1	a_2	H_2	d_2
		H_{12}	O_{10}	r_1	H_{11}	a_1	O_1	d_4
5	donor of central hydrogen bond[1]	O_{13}	O_1	oo*	H_3	a_3*	H_2	d_3*
		H_{14}	O_{13}	r_1	O_1	a_2*	H_3	d_2*
		H_{15}	O_{13}	r_1	H_{14}	a_1	O_1	d_4*
6	neighbor: acceptor for fifth water	O_{16}	O_{13}	oo	H_{15}	a2	H_{14}	d_4
		H_{17}	O_{16}	r_1	O_{13}	a_3	H_{15}	$-d_2$
		H_{18}	O_{16}	r_1	H_{17}	a_1	O_{13}	d_3
7	neighbor: donor for fifth water	O_{19}	O_{13}	oo	H_{14}	a_3	H_{15}	d_3
		H_{20}	O_{19}	r_1	O_{13}	a_2	H_{14}	d_2
		H_{21}	O_{19}	r_1	H_{20}	a_1	O_{13}	d_4
8	neighbor: donor for fifth water	O_{22}	O_{13}	oo	H_{15}	a_3	H_{14}	d_3
		H_{23}	O_{22}	r_1	O_{13}	a_2	H_{15}	d_2
		H_{24}	O_{22}	r_1	O_{23}	a_1	O_{13}	d_4

[a] Results of constrained optimization: r_1, a_1 − water internal parameters; oo, a_2, a_3, d_2, d_3, d_4 − hydrogen-bond geometry parameters. All values: $r_1 = 0.96\,413\,002$, $a_1 = 105.40\,080\,101$, oo = 2.90 045 756, $a_3 = 114.12\,081\,154$, $d_3 = 125.98\,640\,241$, $a_2 = 108.80\,066\,074$, $d_2 = 119.34\,653\,902$, $d_4 = 0$. [b] Use the results of Table 1A to create the remaining clusters. For the maximum cluster, with each central molecule having two donors and two acceptors, the parameters are given by the **Z** matrix.

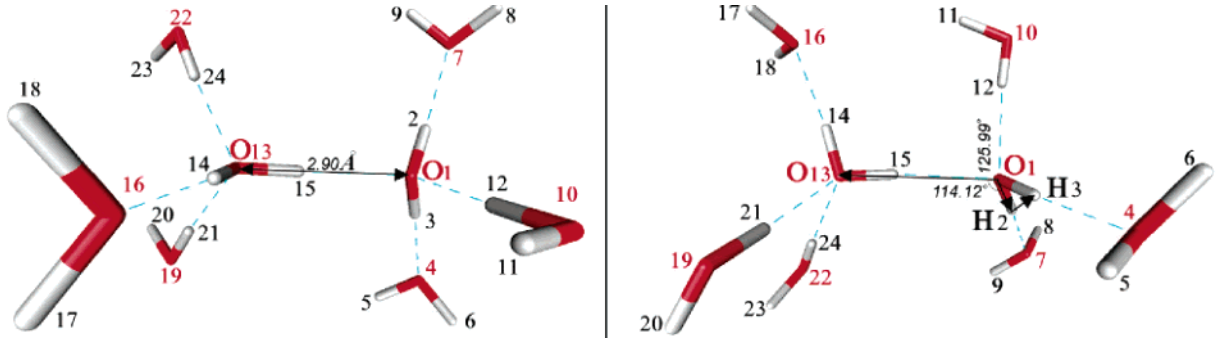


Figure 1. A 2222 cluster (maximum number of neighbors). The O−O distance is shown as 2.90 Å, and the HO−O (the HO is from the acceptor molecule, and the latter O is the central donor oxygen) angle as 114.12°. The dihedral angle formed by the central acceptor water plus the central donor oxygen is 125.99°. The remainder of the geometry is given in Table 1.

found little difference in the two types of cluster, but it was necessary to prove this by explicit calculation.

3. Definition of a Z Matrix. The **Z** matrix gives the positions, angles, and dihedrals of the atoms in the system in terms of the positions of atoms defined on lines above that for the atom being defined. The first column gives the atom being defined; the second, another atom; the third, the distance between them; the fourth, another atom defined earlier; the fifth, the angle formed by those three atoms; the sixth, another atom defined earlier; the seventh and last, the dihedral formed by those four atoms. The first three rows are incomplete, as insufficient atoms have been previously

defined. The first three atoms are the "anchor" that sets the positions of all others.

Figure 1 shows a 2222 cluster with the constrained values for certain angles and distances. The internal coordinates in the cluster are prepared by placing them in a ***Z*** *matrix*, as shown in Table 1. The **Z** matrix contains the results of the operations described above in section 1, and in this section. These values were used as the input to the calculation.

1. Variables That Were Scanned Are Marked by Asterisks. Figure 1 illustrates the 2222 cluster, and it and Table 1 give the initial values of the principal distances and angles. The constrained cluster was very close to the unconstrained cluster in all parameters: unconstrained bond lengths ranged from 0.9611 to 0.9699 Å, with a constrained value of 0.9641 Å for all four O−H covalent bonds. The "unconstrained pair" was defined by taking the initial central pair, freezing all oxygens (two to eight oxygens), and optimizing the hydrogens only. These conditions produced the very limited ranges of distance and angle that are defined as unconstrained here. Bond angles were 105.21−105.61° (unconstrained) and 105.40° (constrained); the dihedral defined as d_4 above was frozen at zero for the unconstrained as well as the constrained dimer but is in any case < 1°. Variables were scanned with and without the constraints. This gives an O−O distance of 2.90 056 Å versus 2.90 046 Å (constrained). The angle defined by the two donors and the H covalently bonded on the acceptor (a_3 above) was 114.08° versus 114.12° (constrained). The dihedral given by the two oxygens and the two hydrogens on the acceptor molecule (d_3 above) was 126.1° versus 126.0° on the constrained pair. Other results were also very close: the energy was −152.96 350 H versus −152.96 306 H (constrained)—we use energy units of hartrees (H), or millihartrees (10^{-3} H, written mH) throughout this paper; 1 mH ≈ k_BT, where k_B = Boltzmann's constant and T = temperature (K)—at approximately 300 K, where most simulations of biological interest are done. Electron density at the bond critical point is 0.02 464 versus 0.02 430 (constrained). Thus, the constraints are unlikely to have a serious effect on any conclusions, to better than 1%.

Although the unconstrained clusters were computed, the constrained values offer a reproducible set of conditions among all clusters, allowing more valid comparison. A wider range of values was found when the oxygens were not frozen, including cases in which some of the clusters were not completely stable. In the majority of the cases, the clusters changed little. However, that is not the point of this work, and we do not consider the fully optimized clusters further.

2. Clusters Other Than Minimal (1001) and Maximal (2222). The remaining 34 clusters were generated by starting with the 2222 cluster and deleting molecules, so that the remaining molecules had the same coordinates as in the 2222 cluster. The deleted molecules were replaced by ghost atoms, keeping their orbitals in the calculation. This is further discussed under basis set superposition error (BSSE) below.

3. Scans. The variables that were scanned were those marked by asterisks in Table 1B, in the central donor rows: O−O distance, two angles, and three dihedrals. In all scans, only one variable at a time was scanned, with all others held constant at their optimized (for the isolated pair) values. The results showed that only the O−O distance made a major difference. Cross terms were not investigated. The central oxygen−oxygen distance was scanned at 0.1 Å intervals, and the energy and other properties were redetermined as a single-point calculation at each position.

The O−O distance at the energy minimum was further defined by an additional scan of 40 points at 0.01 Å intervals. The minimum of the 0.1 Å scan was included in the range, although the range was not always centered exactly on the minimum. These energy values were also used for the computation of forces (see below). Scans were similarly performed on the O−H−O angle and on certain values of three dihedral angles, as shown in Figure 1, which shows the 2222 cluster. Angles and dihedrals gave relatively small effects (<0.8 mH, exceeding 0.5 mH only near the end of the range) as a function of neighbors and angles, as long as the angles did not grow beyond the level that continued to be hydrogen-bonded. The angle and dihedral effects were small, over the range that was investigated, ±10°. For a_3, the effect was at maximum less than 0.00 035 H, around $^1/_3k_BT$ (T = 300 K), and no regularity was found. However, changes in a_2 sometimes produced larger energy differences, sometimes double that of a_3, making it marginally large enough to be of importance in simulations at room temperature; results will be given below. Similarly, each of the three dihedral angles was tested at five points, the optimized value, and that value ±5° and ± 10°. Only d_4, of the dihedrals defined here, exceeded the 0.0005 H threshold and is included specifically in the results. The other two dihedrals did not quite reach 0.0003 H at 10° and are omitted. In this manner, the entire plot of the O−O distance versus the energy was generated (with the best H-bond angles, omitting second-order corrections), as well as the other properties, as a function of distance and of angles and dihedral angles.

4. Calculation Method. Density functional calculations were done using B3LYP/6-311++G**. The Gaussian 03 package[54] was used for all calculations. The charges on the hydrogens and oxygens were determined using NBO from Gaussian 03;[53] AIM calculations were carried out for certain clusters, as a check on the other methods of determining charge.

The hydrogen bond energy is calculated as follows:

$$E_{WW}(R) = E(\text{bond}) - E_{3.9} = E(\text{cluster}) - [\Sigma V_c(\text{pairs}) + \Sigma V_{vdW}(\text{pairs})] - E_{3.9} \quad (1)$$

V_c is defined as the coulomb energy between nonbonded pairs of molecules, with hydrogen bonds counted as bonds; V_{vdW} is the corresponding van der Waals energy. $E_{3.9}$ is the energy of the pair at an O−O distance of 3.9 Å, at which distance the hydrogen bond is considered to have disappeared, making the hydrogen-bond energy equal to zero. Therefore, the $E_{WW}(R)$ defined by eq 1 is the difference in water−water interaction energy of the hydrogen-bonded central pair of molecules at an oxygen−oxygen distance R, from the energy at 3.9 Å, and V_{vdW} is the total van der Waals energy of all nonbonded water pairs, calculated using TIP3P-model potential parameters. Because the contribution of van der Waals energy is so small (<2%), it makes little difference if the

values are somewhat approximate for this term. Although the distances between pairs of molecules hydrogen-bonded to either one of the two central molecules is constant, the charges change slightly with the distance between central molecules, so the energy pair terms, even between molecules at a constant distance and orientation with respect to each other, will change as well. Equation 1 also makes these corrections. The corrections for the coulomb energy of pairs hydrogen-bonded to different water molecules, as the distance changes between the central water molecules, amounts to several millihartrees. The corrections accompanying charge changes on water molecules both bonded to the same central water molecule, thus, at a constant distance from each other, are much smaller, less than 0.5 mH.

The van der Waals terms, as estimated from the TIP3P parameters, follow a similar pattern and are very small. Molecules that are bonded to the same central water, and which therefore remain at constant distance as the central water−water distance changes, have constant van der Waals interactions with each other as the central pair separates; these interactions will be removed by subtracting $E_{3,9}$. The van der Waals correction between molecules on opposite sides of the cluster is already small at the closest approach and diminishes from there. However, these corrections are included in the computation; the TIP3P accuracy is adequate for the small correction terms. The net energy from eq 1 subtracts the energy of all interactions with pairs other than the central pair. *The energy shown as the hydrogen-bond energy is thus the total interaction energy in the cluster less the energy of the pairwise interactions not involving the pair of molecules forming the hydrogen bond of interest.* Any self-energy terms for water molecules are independent of the distance for a given class of molecules, so subtracting $E_{3,9}$ removes this contribution for each cluster. As a result, the differences in energy with the distance between the central oxygens are meaningful for the central hydrogen bond.

Wiberg bond indices found using NBO are the sum of squares of off-diagonal density matrix elements between atoms. They function as a nominal bond order. The Wiberg indices give a measure of the bond interaction between two atoms; for our purposes, it shows the extent of bond overlap associated with the central hydrogen bond—the bond is not purely electrostatic.

5. Basis Set Superposition Error (BSSE) and Ghost Atoms. Because the hydrogen-bond energy is determined as a small difference between two large energies, the BSSE is still >10% of the hydrogen bond itself (both constrained and unconstrained for a single pair). Therefore, in calculating with neighbors, all calculations start from the 2222 configuration, and the preparation of smaller clusters is accomplished by deletion of the neighbor molecules, leaving ghost atoms; hence, all orbitals remain. Therefore, the BSSE is also subtracted, and no further correction is needed for this, making all results comparable. To estimate the actual value, the following calculation can be used: The energy of the 1001 cluster is computed, with all of the remainder of the 2222 cluster replaced by ghost atoms. The sum of the energies of the cluster with each of the 1001 water molecules separately replaced by ghost atoms is subtracted. The difference is −0.0 077 417 H. This is compared with the same calculation on 1001 without the ghost atoms of the remainder of 2222—that is, just two water molecules, in the same geometry. The water molecules that are removed from 1001 then have exactly the same energy as that of a single water molecule. Subtracting the two independent water molecules from the isolated 1001 pair gives −0.0 092 697 H. The difference, 0.0 014 720 H, would be the BSSE, were it not removed by using the ghost atoms. In all cases, the constrained values were used for water geometry, so that the results are consistent.

6. Atoms in Molecules (AIM)[49,55] *and Natural Bond Orbitals (NBO*[53]*).* AIM has been used to obtain the properties of the bond (principally, the bcp and electron density at the bcp). Other properties, such as the Laplacian, were not as informative and are not discussed further. The software used for this purpose was obtained from the Bureau for Innovative Software.[56]

7. Force. The force along the hydrogen bond at the energy minimum for each case is determined by fitting a parabola to the 11 energy values consisting of the energy minimum, and five points at 0.01 Å intervals on either side, for a total distance of 0.05 Å on each side of the minimum. The derivatives could then be determined analytically from the parabolas, with the second derivative giving the Hooke's Law constant.

8. Dipole Moments and Charges. The dipole moments of the central water pair were calculated for each of the 64 clusters. (Although 24 of them are almost redundant by symmetry, slight shifts in angle of a nearest neighbor break the symmetry. This may make a small contribution, in principle, so all 64 were calculated—the shift is not large enough for further discussion, the maximum value being 0.05*D*.) Charges on the central water atoms, from Gaussian 03, were used in the calculations. The dipoles are computed from the integrated wave function from Gaussian 03, and the charges are found using NBO.

Results and Discussion

The results are best expressed in figures showing the energy and other properties that were calculated. Figure 2 has the key finding, concerning energy, determined from eq 1.

The slope of the curves in Figure 2 differs; therefore, the forces on the molecules differ. This is one key to understanding why, and how, these would behave differently in a MD simulation. Note that the correlation is, to first order, with the index K_N, not individual structures. We group the curves by K_N value, as the total range is approximately 6 mH, while no single K_N spreads energy at the minimum over 1 mH (although it is close). Taking the energy for a given K_N as the average, we will always be within 0.5 mH. We shall see below that charge transfer is similarly correlated. This is therefore a step toward a potential for hydrogen bonds as a function of neighbor bonds, in which K_N is the key index. The TIP3P result is good on average, in bulk, but would not reproduce the particular behavior—energy and forces—in a system with a limited number of nearest neighbors. It is necessarily the case that the same would be true of any point-

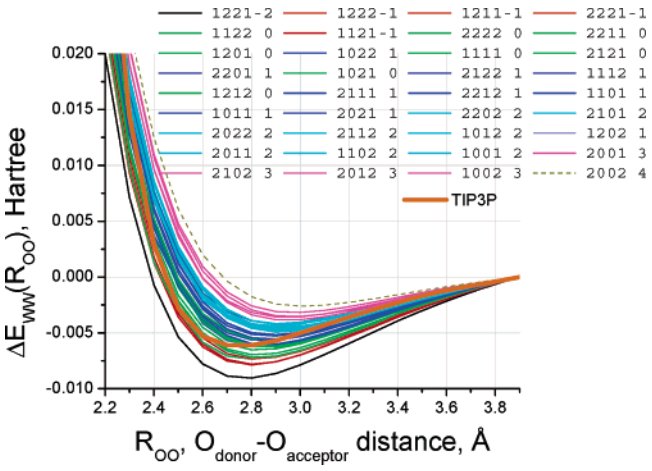


Figure 2. All 36 energy (O−O) distance curves for the total set of neighbor types (four-digit number—see text). Colors represent neighbor indices: The final single digit numbers = $K_1 - K_2 - K_3 + K_4 = K_N$, where K_i ($i = 1-4$) is the number of neighbors in the corresponding four-digit index; call K_N the net index. Curves with the same K_N values group together (same color means same K_N). The energy values increase monotonically with K_N; $K_N = -2$ (black) is lowest energy and +4 (dotted) weakest, with each color change at an increase of 1 in K_N. The heavy orange curve shows the TIP3P model result, with no dependence on neighbors, and a somewhat different shape of the potential-distance curve from the quantum calculations for any set of neighbors. $E_{WW}(R)$ for all clusters is set to zero at 3.9 Å, as defined in eq 1.

charge model. A polarizable model would do little better, in that it depends on the electric field, which in turn varies drastically (by a factor of close to three) in a local cluster. Thus, it differs from the bulk average for which such potentials are parametrized, making it useless to compare the results from such a potential. Therefore, in simulations involving limited-size clusters, we see strong evidence that the relation of energy as well as force to the number of neighbors of each type should be included in the water potential. In this study, only water is considered as a neighboring species, but in proteins, there are carbonyls, carboxylates, amines, and other groups as well. While these have been calibrated for TIP3P water with the OPLS potentials, cooperative effects would still not be represented. The variation of hydrogen bonds in differing conditions might be as great as that of water with different numbers of water neighbors.

The energies at the minima in Figure 2 cover a range equivalent to approximately 6 k_BT near room temperature. In MD, the force as well as the energy matters. For the TIP3P curve, in the range of 2.8∼3.2 Å, where the bonds have their greatest effect (counting the number in that range as well as the energy), force is about 0.0067 H Å^{-1}, while for the others it ranges from approximately 0.004 to 0.008 H Å^{-1} (to 0.010 for the 1221 cluster), meaning that the water molecules could move more or less rapidly than the average. The angle dependencies are much weaker, as noted in the methods section. However, one angle, a_2, and one dihedral, d_4, produce energies differing by slightly over ±0.5 mH, making them just large enough to take into consideration (while it might seem arbitrary to set a threshold of 0.5 mH, it is necessary to set some threshold at a reasonable value). If the other angle and dihedrals were included, they would produce similar results, but with a little more scatter and a little less significance. We characterize the angles by an equation of the form:

$$E(D) = C_1 D + C_2 D^2 \tag{2}$$

where in this case D can be either an angle or a dihedral. The equation was fitted to the five points ($D = -0°, -5°, 0, +5°, +10°$, with 0° being the value at the minimum), for each of the clusters. The plots in Figure 3 are C_1 versus C_2 (normalized by the average value of the plot). Again, we see that K_N is a valuable parameter in classifying the points, but the energy differences here are much smaller (roughly by an order of magnitude).

The properties of the hydrogen bonds, whether determined as bond order, electron density at the bond critical point, energy, or the index K_N, all have a similar relation to the O−O distance, as shown in Figure 4. There are several ways to show how the bond distance, energy, electron density, and bond length are related; the key point is that the relation is dependent on the arrangement of neighbors of the hydrogen-bonded pair of water molecules.

Equation 3 gives six such relations. All are fitted parabolas. Define $R_d = R_{OO} - 2.9$ Å (2.9 Å is the equilibrium distance for the 1001 cluster) where R_{OO} is the oxygen−oxygen distance of the central pair, for each of the single-point calculations in the scan. Then, we can express (1) the hydrogen-bond energy at the minimum, relative to the energy at 3.9 Å, ΔE_{WW} (3.9 Å = full separation, as shown in Figure 2); (2) the electron density at the bcp ρ (from AIM); (3) w, the sum of the Wiberg indices for the six atoms of the two central molecules; (4) the Wiberg index for the central hydrogen bond alone, w1; (5) the net index K_N; and (6) the Hooke's Law force constant F at the minimum distance determined within 0.5 Å of the energy minimum:

$$\Delta E_{WW} = -0.00\,474 + 0.2418 R_d - 0.06\,967 R_d^2 \text{ (H)} \tag{3a}$$

$$\rho = 0.02\,444 - 0.0585 R_d + 0.0977 R_d^2 \tag{3b}$$

$$w = 0.03\,502 - 0.17\,074 R_d + 0.55\,466 R_d^2 \tag{3c}$$

$$\text{w1} = 0.0238 - 0.13\,417 R_d + 0.45\,388 R_d^2 \tag{3d}$$

$$K_N = 1.97\,106 + 23.69\,411 R_d - 38.72\,767 R_d^2 \tag{3e}$$

$$F = 0.04\,271 - 0.16\,497 R_d + 0.11\,094 R_d^2 \text{ (H/Å}^2\text{)} \tag{3f}$$

Because there is little interaction between atoms that are neighbors but are not covalently bonded, the index values are small; however, they are appreciably different from zero, and the w1 index suggests some significant bond overlap for the hydrogen bond. The other quantities are given in Figure 4. R_d is at most ±0.13 Å, but the curvature is nevertheless appreciable. The Wiberg bond index for the hydrogen bond is roughly $^2/_3$ of the sum of these indices for

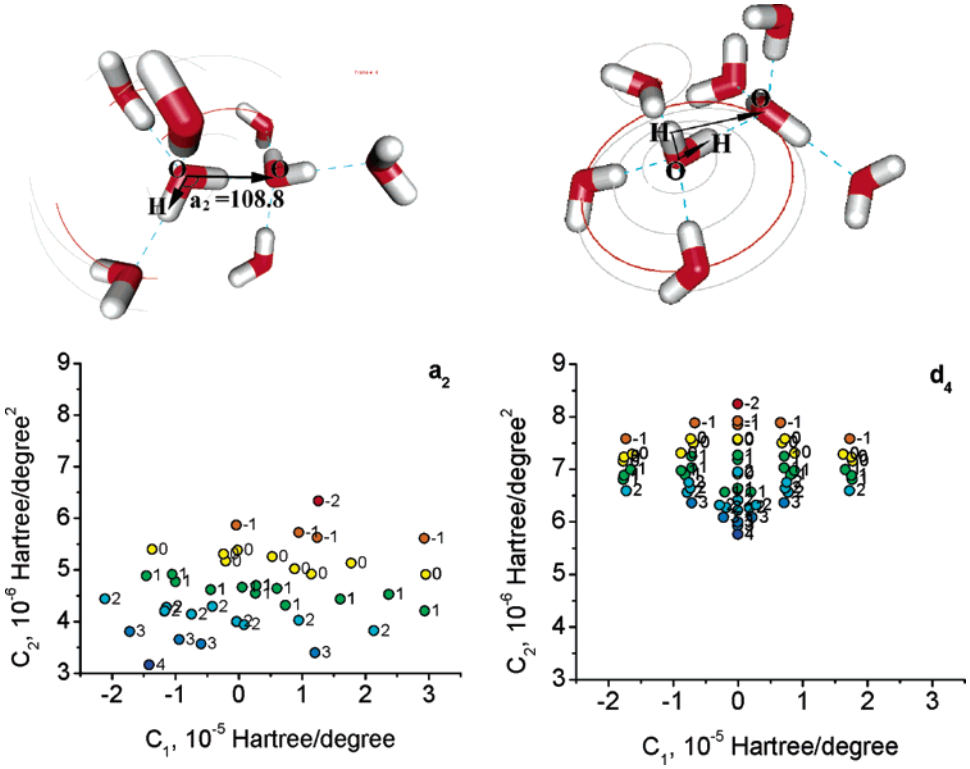


Figure 3. Angle a_2 and dihedral d_4. The values of angle a_2 and d_4 are shown as C_1 vs C_2 of eq 2. The top panels show a_2 and d_4 pictorially; the red arcs show how the oxygen atoms move, gray arcs the corresponding hydrogen atoms, when the angle or dihedral is scanned. The atoms that move do so as a rigid body. Those that do not move do not have arcs. The initial (zero) value of the $O_a-O_d-H_d$ (a = acceptor, d = donor) angle D is 108.8°, and that of the dihedral is 0°. The points in the lower plots are shown in color according to K_N and labeled with the K_N value. The energy scale of the two figures has been set up identically so that it is clear that the C_2 value, which provides the larger contribution to the energy, is larger for the dihedral (to get energy, multiply by the angle^2, hence, at 10°, 100). We do not have an explanation for the fairly even spacing of the dihedral C_1 values. The most obvious regularity shows that the smaller the K_N value is, the larger the C_2 value.

the nine pairs of atoms at all separations. Figure 4 suggests also the limited level of scatter in the relations.

Charge transfer between the central water molecules also occurs. Figure 5 shows the charge on the central donor molecule as the central pair and its associated neighbors are separated, for each group of possible neighbors. In this plot, the clusters are shown according to their first two (donor) indices. This is a result that cannot be reproduced by any of the standard point-charge models, for which the charges are fixed, with a net charge of zero. The charge includes transfer to the neighbors on the same side of the cluster, which is responsible for the nonzero value of the charge at large separations. The sum of the charges on the donor side, and on the acceptor side, does go to zero as required at large separations.

Forces are calculated as Hooke's Law constants from the parabolic fit near the energy minimum and are also well-ordered by the index K_N; the strongest bond, and the shortest, has $K_N = -2$ (i.e., the 1221 cluster), with $K_N = -1$ grouped close behind, and a reasonably well-ordered progression on to +4 (2002 cluster). The ordering is not perfect, with three clusters, with indices 22*xx*, having slightly larger constants than the equation gives.

What is surprising is the range of the Hooke's Law constants, from 0.0068 H Å^{-2} (1221) for the strongest down to 0.0028 H Å^{-2} for the weakest (2002). As the force constants control motion in a simulation, this is a matter of some importance.

Finally, we have the dipole moments, shown in Table 2. The charges on the central donor, in Figure 5, obtained from NBO calculations, show a clear relation to the donor half of the cluster and a much weaker relation to the acceptor half; only at short distances is the $Y-Z$ dependence sufficient that there is appreciable difference in charge. The charge on the donor water of the central pair in a cluster can be described, at a central oxygen−oxygen distance of 2.9 Å, by eq 4a:

$$Q = \Sigma i\, \delta Q_i \qquad (4a)$$

with

$$\delta Q = -0.0127 + 0.0013(W - 1) - 0.0020X - 0.0024Y + 0.0008(Z - 1) \quad (4b)$$

These quantities are defined as follows: (with W, X, Y, and Z replacing K_1, K_2, K_3, and K_4 in the equation and in Figure 5).

1. δQ_i and δQ. The δQ_i of eq 4a is determined from the charge transfer from the donor member of the central dimer to the (up to) four neighbors of that molecule. The resulting charge on the donor water is Q and is the quantity shown in Figure 5a.

The δQ of eq 4b is the charge transfer between the two members of the central dimer. This is the same as the total

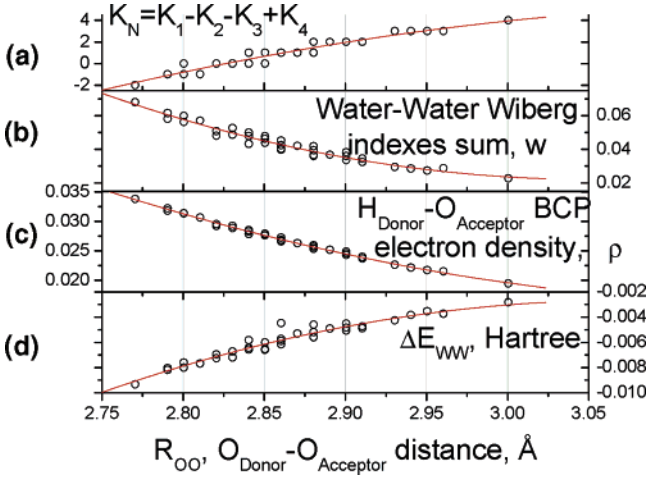


Figure 4. Properties of the clusters as a function of the O−O distance of the central pair at the energy minimum; the minimum covers 0.23 Å, from the shortest to longest bonds, while the energy varies by approximately 6 mH, or about 6 k_BT near room temperature: (a) $K_N = K_1 - K_2 - K_3 + K_4$, the net index; the other properties appear to correlate with this index more closely than with the individual indices. (b) the sum of Wiberg bond indices for all pairs of the central pair of atoms (sum of $3 \times 3 = 9$ indices, with three atoms on each molecule). (c) The electron density at the bcp for the hydrogen bond (calculated from AIM). (d) The energy of the system relative to that at a central O−O separation of 3.9 Å. For the parabolic fits to these results, see eqs 3a−3c and 3e, with two other quantities represented by eqs 3d and 3f. Note the correlation with K_N in particular carries through to all the properties; the K_N values can be read from the figure for the other three properties shown (extend a vertical line down from a). There are four cases in which there is an overlap of K_N values at the same bond distance, and the other properties tend to overlap at these distances also.

charge transfer between the donor half of the cluster and the acceptor half. This is also the charge-transfer Q shown in Figure 5b.

2. Dipoles. Dipoles are computed by integrating the wave function, hence, by a method nominally independent of the computation of the charge (of course, the two are related, but one is not found directly from another). We are primarily concerned here not with the total dipole moment but with the difference in the dipole due to the cooperative effects we are studying. Therefore, define

$$\delta\mu = \mu_{\text{cluster}} - (\mu_{\text{dimer}} + \Sigma_n \mu_n) \tag{5}$$

In other words, $\delta\mu$ is the vector difference between what the cluster would have for a dipole moment if there were no cooperative interactions and what it actually has. The values are calculated as vectors and then averaged over all 64 possible clusters. The isolated molecule contributions are subtracted in $\delta\mu$.

Table 2 shows the interesting result that the magnitude of the vector contribution of an added water molecule is roughly the same for all the possible water molecules that can be added to the central pair, on average. The computation has been done for the following cases: (1) Add water "physically" (i.e., when the water is not present, the wave functions are not left behind) using constrained clusters. *2*) Use ghost atoms when the neighbor water is not present, again using constrained clusters. This removes BSSE. (3) Relax the constraint to the extent of reoptimizing the OH distance in the central hydrogen bond, and use ghost atoms again. (4) Use TIP3P, which should make the cooperative effect zero. This is not shown in the table; the actual values were about 10^{-6}, too small to show in the table. That value gives the error, including round-off error, in the suite of programs used to get these values and served as a check. (5) Modify TIP3P, with the central OH bond optimized. These results were on the order of 10^{-3}, and are shown.

The average is computed by determining $\delta\mu$ for each cluster:

$$\delta\mu = \sum_i^6 k_i x_i \tag{6}$$

where the k_i is 0 or 1, depending on whether the molecule to which it refers is present or absent and the x_i represents the vector contributions of each water molecule. The result is fitted to allow the total cluster to have the correct value of $\delta\mu$, giving the magnitudes of the x_i. The result is averaged over all 64 clusters (each k_i has two values; hence, there are 2^6 total values). It is this average that is shown in Table 2.

Clearly, the dipoles are somewhat affected by BSSE, so the correction should be included. Also, the symmetry of the donors on the acceptor, and that of the two acceptors on the donor, is as expected (doing them separately both takes account of the possibility that the slight asymmetry of the angles may matter—evidently, it does not—and acts as a check on the calculation). Were any of the differences in these cases large, it would have shown an error. The differences are on the order of 10^{-5}, which shows that the angle asymmetry can be neglected, and that no error in the program can be found from this part of the calculation. What is especially important in this set of results is the fact that all the molecules make a similar contribution, suggesting that the nonlinear effects, although much too large to ignore, will not be too difficult to include in a fairly simple set of parameters in a new potential.

Consequences for Molecular Dynamics Simulations. For MD simulations, these relations suggest a major readjustment in treatment of water in limited clusters. The extent to which such clusters are a good description of water molecules bounded as, for example, in proteins remains to be demonstrated, as the clusters we have studied exist in a vacuum. However, it would be surprising if water in small bounded cavities behaved like bulk water; it is possible that certain groups, especially carboxyl or hydroxyl, could partially replace water. Other groups, such as amines, or charged groups, might produce results even more different from bulk water than a vacuum does. The OPLS[3] potentials are designed for proteins with TIP3P water, but again, cooperative effects are not included in any point-charge model. The differences will include deciding whether the water is likely to remain bound or will change positions rapidly, and whether the hydrogen bonds form a barrier to solutes or permit passage. Temperature dependence will also

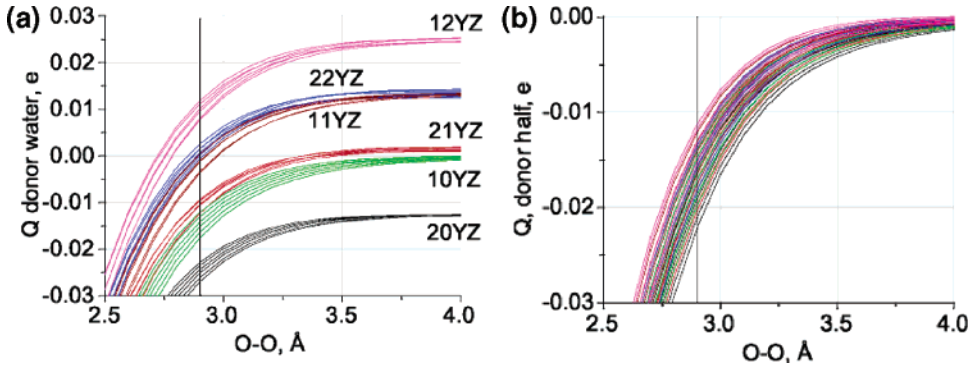


Figure 5. (a) Charge on the central donor molecule (for a definition of charge, see eq 4, with the text below) for all 36 clusters, grouped by donor molecule indices. Acceptor molecule indices for each curve are not shown, because they group tightly enough that they are not important. Note that the charge on the molecule when the two half clusters have separated at a large distance is approximately 0.013−0.014 charges for an increment of one in *X*, and about −0.010 to −0.012 for an increment of 1 in *W*. The charge on the central molecule does not go to zero at infinite separation (except, fortuitously, for the 21*YZ* clusters), because of charge transfer within the half cluster. The charge transferred for short bonds is much greater, which is not a surprise. Even at a large distance, the difference between having the central molecule being a donor twice and never an acceptor (20*YZ*), and being an acceptor twice and donor only the one time required by definition (12*YZ*), is > 0.03 electronic charges, which is appreciable. (b) Charge on the half clusters. This time, the charge must go to zero at large distances. The curves appear to cross, because of the conflicting effects of the donor and acceptors. However, the lowest curve (largest negative charge) again goes with the 20*YZ* (double donor, no acceptor) cluster, the least negative is the 12*YZ* cluster. The difference in charge at 2.9 Å is >0.01 *e*, large enough to be important.

Table 2. Contribution of Each Molecule to the Dipole Moment of the Cluster, Above that of the Isolated Molecule, for Each Type of Calculation; First and Second Rows, Donor to the Acceptor, Third Row, Acceptor to the Acceptor, Fourth Row, Donors to the Donor, Fifth and Sixth Rows, Acceptors to the Donor[a]

molecule as in caption	no ghost atoms	ghost atoms	ghost atoms OH reoptimized	modified TIP3P OH reoptimized
1	0.611	0.717	0.724	0.002
2	0.611	0.717	0.724	0.002
3	0.588	0.668	0.667	0.004
4	0.582	0.710	0.716	0.002
5	0.634	0.730	0.738	0.005
6	0.634	0.730	0.738	0.005

[a] Only the magnitude of the vector contribution is given. Columns are in the order given in the text, with the unmodified TIP3P not shown.

be altered, as will the work done in moving a solute through the water. Further work is required to make it possible to use potentials consonant with these results in a simulation. In particular, methods for tracking the number and type of nearest neighbors are being developed. In addition, a consistent set of potentials for other atom types in molecules (with biomolecules of special interest) must be worked out. We have not mentioned charged clusters at all here. However, work is underway to study these, as well as to keep track of individual atoms, in order to make possible the simulation of proton transfer. When the entire system of potentials and other software is complete, it should be possible to simulate a system containing hydrogen bonds accurately, with the cooperative effects, and include in the simulation the exchange of protons, whether or not the species are charged.

Conclusions

Hydrogen bonds have been found to have significantly different bonding strength from that in bulk water in a type of small cluster that emphasizes the role of cooperative effects involving nearest neighbors. Energy, bond length, electron density at the bond critical point, force constants, dipole moments, and Wiberg bond indices are strongly correlated with the number and type of nearest neighbors. Forces that differ from those in bulk water would have significant effects on the results of MD simulations. Because the clusters were in a vacuum, further work will be required to extend these results to proteins with small clusters of water in clefts or cavities. However, it is now clear that it is dangerous to ignore cooperative effects in simulations using point-charge potentials under nonbulk conditions.

Acknowledgment. This work has been supported in part by grants from NIH (SCORE grant to City College of New York, subgrant to MEG) and PSC−CUNY university grants.

References

(1) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potentials for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926−35.

(2) Jorgensen, W. L.; Madura, J. D. Temperature and Size Dependence for Monte Carlo Simulations of TIP4P Water. *Mol. Phys.* **1985**, *56*, 1381−1392.

(3) Jorgensen, W. L.; Tirado-Rives, J. The OPLS (Optimized Potentials for Liquid Simulations) Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.

(4) Mahoney, M. W.; Jorgensen, W. L. Diffusion Constant of the TIP5P Model of Liquid Water. *J. Chem. Phys.* **2001**, *114*, 363−366.

(5) Mahoney, M. W.; Jorgensen, W. L. A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid Non-Polarizable Potential Functions. *J. Chem. Phys.* **2000**, *112*, 8910−8922.

(6) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces;* Pullman, B., Ed.; Reidel: Dordrecht, Holland, 1981; p 331.

(7) van der Spoel, D.; van Maaren, P. J.; Berendsen, H. J. C. A Systematic Study of Water Models for Molecular Simulation Derivation of Water Models Optimized for Use with a Reaction Field. *J. Chem. Phys.* **1998**, *108*, 10220−10230.

(8) Ahlstrom, P.; Wallqvist, A.; Engstrom, S.; Jonsson, B. A Molecular Dynamics Study of Polarizible Water. *Mol. Phys.* **1989**, *68*, 563−581.

(9) Burnham, C. J.; Xantheas, S. S. Development of Transferable Interaction Models for Water. IV. A Flexible, All-Atom Polarizable Potential (TTM2-F) Based on Geometry Dependent Charges Derived from an ab Initio Monomer Dipole Moment Surface. *J. Chem. Phys.* **2002**, *116*, 5115−5124.

(10) Fanourgakis, G. S.; Xantheas, S. S. The Flexible, Polarizable, Thole-Type Interaction Potential for Water (TTM2-F) Revisited. *J. Phys. Chem. A* **2006**, *110* (4), 100−4106.

(11) Hartke, B. Size-Dependent Transition from All-Surface to Interior-Molecule Structures in Pure Neutral Water Clusters. *Phys. Chem. Chem. Phys.* **2003**, *5*, 275−284.

(12) Donchev, A. G.; Galkin, N. G.; Illarionov, A. A.; Khoruzhii, O. V.; Olevanov, M. A.; Ozrin, V. D.; Subbotin, M. V.; Tarasov, V. I. Water Properties from First Principles: Simulations by a General-Purpose Quantum Mechanical Polarizable Force Field. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8613−8617.

(13) Mo, O.; Yanez, M.; Elguero, J. Cooperative (Nonpairwise) Effects in Water Trimers: An ab Initio Molecular Orbital Study. *J. Chem. Phys.* **1992**, *97*, 6628−663. 8.

(14) Kuo, J.-L.; Ciobanu, C. V.; Ojamae, L.; Shavitt, I.; Singer, S. J. Short H-Bonds and Spontaneous Self-Dissociation in $(H_2O)_{20}$: Effects of H-Bond Topology. *J. Chem. Phys.* **2003**, *118*, 3583−3588.

(15) Tissandier, M. D.; Singer, S. J.; Coe, J. V. Enumeration and Evaluation of the Water Hexamer Cage Structure. *J. Phys. Chem. A* **2000**, *104*, 752−757.

(16) Tsai, C. J.; Jordan, K. D. Theoretical Study of Small Water Clusters: Low-Energy Fused Cubic Structures for $(H_2O)_n$, $n = 8$, 12, 16, 20. *J. Phys. Chem.* **1993**, *97*, 5208−5210.

(17) Anick, D. J. Polyhedral Water Clusters, II: Correlations of Connectivity Parameters with Electronic Energy and Hydrogen Bond Lengths. *THEOCHEM* **2002**, *587*, 97−110.

(18) Anick, D. J. Application of Database Methods to the Prediction of B3LYP-Optimized Polyhedral Water Cluster Geometries and Electronic Energies. *J. Chem. Phys.* **2003**, *119*, 12442−12456.

(19) Hodges, M. P.; Stone, A. J.; Xantheas, S. S. Contribution of Many-Body Terms to the Energy for Small Water Clusters: A Comparison of ab Initio Calculations and Accurate Model Potentials. *J. Phys. Chem. A* **1997**, *101*, 9163−9168.

(20) Xantheas, S. S. Cooperativity and Hydrogen Bonding Network in Water Clusters. *Chem. Phys.* **2000**, *258*, 225−231.

(21) Ojamae, L.; Hermansson, K. Ab Initio Study of Cooperativity in Water Chains: Binding Energies and Anharmonic Frequencies. *J. Phys. Chem.* **1994**, *98*, 427 1−4282.

(22) Chen, W.; Gordon, M. S. Energy Decomposition Analyses for Many-Body Interaction and Applications to Water Complexes. *J. Phys. Chem.* **1996**, *100*, 14316.

(23) Luck, W. A. P.; Klein, D.; Rangsriwatananon, K. Anti-Cooperativity of the Two Water OH Groups. *J. Mol. Struct.* **1997**, *416*, 287−296.

(24) Wu, Y.; Tepper, H.; Voth, G. A. Flexible Simple Point-Charge Model with Improved Liquid-State Properties. *J. Chem. Phys.* **2006**, *124*, 024503.

(25) Piquemal, J.-P.; Cisneros, G. A.; Reinhardt, P.; Gresh, N.; Darden, T. A. Towards a Force Field Based on Density Fitting. *J. Chem. Phys.* **2006**, *124*, 104101/1−12.

(26) Dahlke, E. E.; Truhlar, D. G. Assessment of the Pairwise Additive Approximation and Evaluation of Many-Body Terms for Water Clusters. *J. Phys. Chem. B* **2006**, *110*, 10595−10601.

(27) Todorova, T.; Seitsonen, A. P.; Hutter, J.; Kuo, I.-F. W.; Mundy, C. J. Molecular Dynamics Simulation of Liquid Water: Hybrid Density Functionals. *J. Phys. Chem. B* **2006**, *110*, 3685−3691.

(28) Simon, S.; Duran, M.; Dannenberg, J. J. Effect of Basis Set Superposition Error on the Water Dimer Surface Calculated at Hartree−Fock, Moller−Plesset, and Density Functional Theory Levels. *J. Phys. Chem. A* **1999**, *103*, 1640−1643.

(29) Berweger, C. D.; Van Gunsteren, W. F.; Mueller-Plathe, F. Force Field Parameterization by Weak Coupling. Re-Engineering SPC Water. *Chem. Phys. Lett.* **1995**, *232*, 429−36.

(30) Rablen, P. R.; Lockman, J. W.; Jorgensen, W. L. Ab Initio Study of Hydrogen-Bonded Complexes of Small Organic Molecules, with Water. *J. Phys. Chem. A* **1998**, *102*, 3782−3797.

(31) Anishkin, A.; Sukharev, S. Water Dynamics and Dewetting Transitions in the Small Mechanosensitive Channel MscS. *Biophys. J.* **2004**, *86*, 2883−2895.

(32) Beckstein, O.; Sansom, M. S. P. Liquid-Vapor Oscilations of Water in Hydrophobic Nanopores. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7063−7068.

(33) Tuckerman, M. E.; Marx, D.; Klein, M. L.; Parrinello, M. On the Quantum Nature of the Shared Proton in Hydrogen Bonds. *Science* **1997**, *275* (8), 17−820.

(34) Suresh, S. J.; Satish, A. V.; Choudary, A. Influence of Electric Field on the Hydrogen Bond Network of Water. *J. Chem. Phys.* **2006**, *124*, 074506−1/9.

(35) Arsic, J.; Kaminski, D.; Poodt, P.; Vlieg, E. Liquid Ordering at the Brushite-(0 10)-Water Interface. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2004**, *69*, 245406/1−4.

(36) Uchida, H.; Takaiyama, H.; Matsuoka, M. Molecular Dynamics Simulation of the Solution Structure near the Solid−Liquid Interface between the NaCl(100) and NaCl−KCl−H_2O Solutions. *Cryst. Growth Des.* **2003**, *3*, 209−213.

(37) Green, M. E.; Lu, J. Simulation of Water in a Small Pore: Effect of Electric Field and Density. *J. Phys. Chem. B* **1997**, *101*, 65 12−6524.

(38) Lu, J.; Green, M. E. Simulation of Water in a Pore with Charges: Application to a Gating Mechanism for Ion Channels. *Prog. Colloid Polym. Sci.* **1997**, *103*, 121−129.

(39) Lu, J.; Green, M. E. Simulation of Water in a Small Pore: Effect of Electric Field and Density II: Immobilized Molecules. *J. Phys. Chem. B* **1999**, *103*, 2776−2780.

(40) Guidoni, L.; Torre, V.; Carloni, P. Water and Potassium Dynamics inside the KcsA K^+ Channel. *FEBS Lett.* **2000**, *477*, 37−42.

(41) Bezrukov, S. M.; Kasianowicz, J. J. The Charge State of an Ion Channel Controls Neutral Polymer Entry into Its Pore. *Eur. J. Biophys.* **1997**, *26* (47), 1−475.

(42) Johnston, J. M.; Cook, G. A.; Tomich, J. M.; Sansom, M. S. P. Conformation and Environment of Channel-Forming Peptides: A Simulation Study. *Biophys. J.* **2006**, *90*, 1855−1864.

(43) Scheiner, S. Calculation of Isotope Effects from First Principles. *Biochem. Biophys. Acta* **2000**, *1458*, 28−42.

(44) Scheiner, S. Proton Transfers in Hydrogen-Bonded Systems: Cationic Oligomers of Water. *J. Am. Chem. Soc.* **1981**, *103*, 3 15−320.

(45) Hillebrand, E. A.; Scheiner, S. Effects of Molecular Charge and Methyl Substitution on Proton Transfer between Oxygen Atoms. *J. Am. Chem. Soc.* **1984**, *106*, 6266−6273.

(46) Kirchner, B. Cooperative vs. Dispersion Effects: What Is More Important in an Associated Liquid Such as Water. *J. Chem. Phys.* **2005**, *123*, 204116/1−13.

(47) Ohno, K.; Okimura, M.; Akai, N.; Katsumoto, Y. The Effect of Cooperative Hydrogen Bonding on the OH Stretching-Band Shift for Water Clusters Studied by Matrix-Isolation Infrared Spectroscopy and Density Functional Theory. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3005−3014.

(48) Znamenskiy, V. S.; Green, M. E. Topological Changes of Hydrogen Bonding of Water with Acetic Acid: AIM and NBO Studies. *J. Phys. Chem. A* **2004**, *108*, 6543−6553.

(49) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Oxford U. P.: Oxford, U. K., 1990.

(50) Lin, J.; Frey, P. A. Strong Hydrogen Bonds in Aqueous and Aqueous-Acetone Solutions of Dicarboxylic Acids: Activation Energies for Exchange and Deuterium Fractionation Factors. *J. Am. Chem. Soc.* **2000**, *122*, 11258−11259.

(51) Cleland, W. W.; Frey, P. A.; Gerlt, J. A. The Low Barrier Hydrogen Bond in Enzymatic Catalysis. *J. Biol. Chem.* **1998**, *273*, 25529−25532.

(52) Frey, P. A.; Cleland, W. W. Are There Strong Hydrogen Bonds in Aqueous Solution. *Bioorg. Chem.* **1998**, *26*, 175−192.

(53) Weinhold, F. *NBO*; Theoretical Chemistry Institute, University of Wisconsin, Madison, WI, 2001.

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

(55) Popelier, P. *Atoms in Molecules, An Introduction;* Prentice Hall/Pearson Education: Harlow, England, 2000.

(56) Biegler-Konig, F.; Schoenbohm, J. *Title of Software*, v. 2.0; Buro fur Innovative Software: Bielefeld, Germany, 2002.

CT600139D

JCTC Journal of Chemical Theory and Computation

Molecular Structures and Energetics Associated with Hydrogen Atom Addition to the Guanine−Cytosine Base Pair

Jun D. Zhang* and Henry F. Schaefer III

Center for Computational Chemistry, University of Georgia, Athens, Georgia 30602-2525

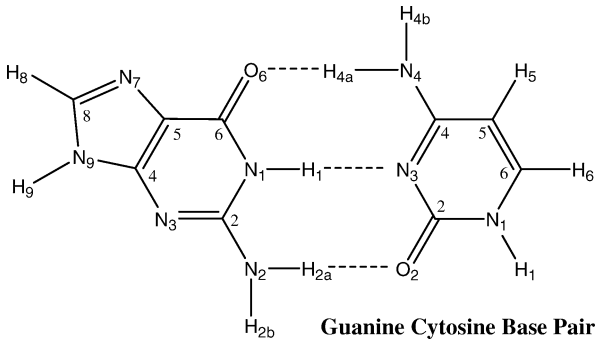
Received August 10, 2006

Abstract: The radicals generated by hydrogen-atom addition to the Watson−Crick guanine−cytosine (G−C) DNA base pair were studied theoretically using an approach that has proved effective in predicting molecular structures and energetics. All optimized structures were confirmed to be minima via vibrational frequency analysis. The dissociation energies of the base-pair radicals are predicted and compared with that of the neutral G−C base pair. The lowest-energy base-pair radical is that with the hydrogen atom attached to the C8 position of guanine, resulting in the nitrogen radical designated G(C8)−C. In this, the most favorable radical, the G−C pair C8═N7 distance of 1.310 Å increases to 1.453 Å when the π bond is broken upon hydrogen-atom addition. This radical has a dissociation energy of 28 kcal/mol, which may be compared with 27 kcal/mol for neutral G−C. The other (GC + H)• radical dissociation energies range downward to 8 kcal/mol. Significant structural changes were observed when the hydrogen was added to the sites where the interstrand hydrogen bonds are formed. For example, "butterfly"-shape structures were found when the hydrogen atom was added to the C4 or C5 sites of guanine. The formation of radical G(C2)−C may cause a single-strand break because of significant strain in the closely stacked base pairs. Radical G(C8)−C is of biological importance because it may be an intermediate in the formation of 8-oxo guanine.

Introduction

With the obvious exception of normal cellular DNA metabolism, lesions resulting from external agents such as ionizing radiation are believed to be the most important source of DNA damage with respect to the preservation of genetic integrity.[1] To gain insight concerning the processes and mechanisms of radiation-induced damage to DNA, two underlying chemical pathways, the direct type and indirect type, have been introduced.[2] The direct-type damage involves energy which mainly comes from monoenergetic photons[3−6] or electrons,[4,7−10] being deposited directly on the nucleotide of the DNA strand.[11] For example, as a primary product of direct damage, the base radical cation $B^{\bullet+}$ often deprotonates to yield the different radicals $B^{\bullet}(-H^{+})$.[12] A distinctly different effect involves indirect damage pathways related to the radiosensitive environment. As a tightly surrounding medium, water can be irradiated to form free radicals such as atomic hydrogen and reactive oxygen species (ROS) like $O_2^{\bullet-}$, $OH^{\bullet}$, and H_2O_2.[13] These radicals may then react with nucleotides

Chart 1. Numbering Scheme for the Guanine−Cytosine Base Pair



Guanine Cytosine Base Pair

* Corresponding author e-mail: jzhang@chem.uga.edu.

10.1021/ct600262p CCC: $37.00
Published on Web 12/13/2006

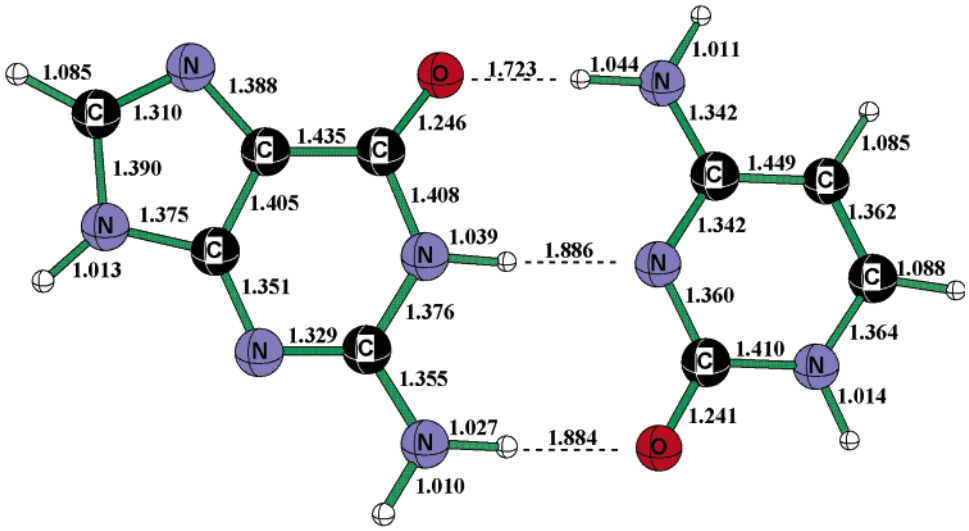


Figure 1. Optimized geometry of the neutral guanine−cytosine (G−C) base pair at the B3LYP/DZP++ level of theory. The unit of bond length in this figure and the following figures is the angstrom (Å).

Table 1. Total and Relative Energies of (G + H)•−C and G−(C + H)• Base-Pair Radicals (See Chart 1 for Atom Numbering)

	non-ZPVE		ZPVE-corrected	
radicals	total energy (Eh)	relative energy (kcal/mol)	total energy (Eh)	relative energy (kcal/mol)
G(C8)−C	−938.28079	0.0	−938.05436	0.0
G−C(N3)	−938.27585	3.1	−938.04959	3.0
G−C(C5)	−938.27503	3.6	−938.04835	3.8
G−C(C6)	−938.27152	5.8	−938.04528	5.7
G(N7)−C	−938.26148	12.1	−938.03528	12.0
G(C4)−C	−938.25296	17.5	−938.02651	17.5
G(C5)−C	−938.24890	20.0	−938.02309	19.6
G(C2)−C	−938.24606	21.8	−938.01945	21.9
G(O6)−C	−938.24108	24.9	−938.01452	25.0
G−C(O2)	−938.23572	28.3	−938.01054	27.5
G(N3)−C	−938.23540	28.5	−938.00920	28.3
G−C(C4)	−938.22578	34.5	−937.99867	34.9

to induce structural damage such as single-strand breaks (SSBs)[14] and double-strand breaks,[15] as well as genetic information changes like pairing mismatches and mutations.[16] There are extensive studies both theoretical[17−19] and experimental[20−23] of OH• radicals reacting with purines and pyrimidines by removing one hydrogen atom[12a] or directly attacking double bonds.[24,25] However, the role played by non-ROS such as hydrogen atoms in DNA radiation damage is less clear.

Some earlier studies of isolated hydrogenated nucleic acid base (NAB) radicals must be noted. In 1998, Wetmore et al.[26] showed H-atom addition to guanine to be preferable at position C8 (See Chart 1 for the standard atom numbering in guanine) from electron paramagnetic resonance (EPR) and electron nuclear double resonance (ENDOR) experiments, as well as density functional theory (DFT). Hydrogenated cytosine has also been explored at the DFT level theoretically[27,28] and by EPR/ENDOR[29,30] experiments. As transient intermediates, the resultant NAB radicals either (a) undergo

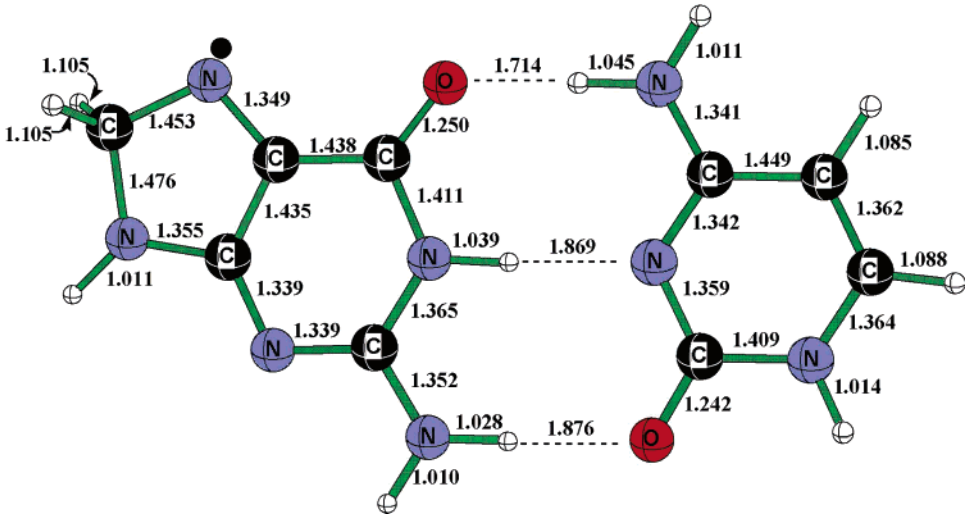


Figure 2. Optimized geometry of the G(C8)−C base-pair radical at the B3LYP/DZP++ level of theory. Relative to neutral G−C, there is a hydrogen atom attached to guanine C8. The dot (•) shows the formal radical center position in the qualitative valence structure.

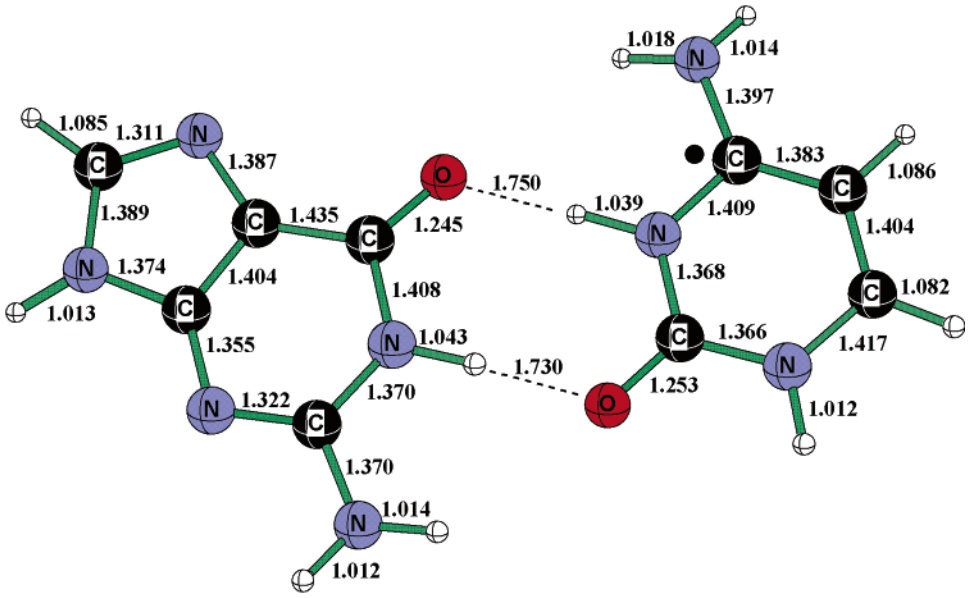


Figure 3. Optimized geometry of the G−C(N3) base-pair radical at the B3LYP/DZP++ level of theory. A hydrogen atom is attached to cytosine N3 when compared to the neutral G−C pair. The dot (•) shows the formal radical center position in the qualitative valence structure.

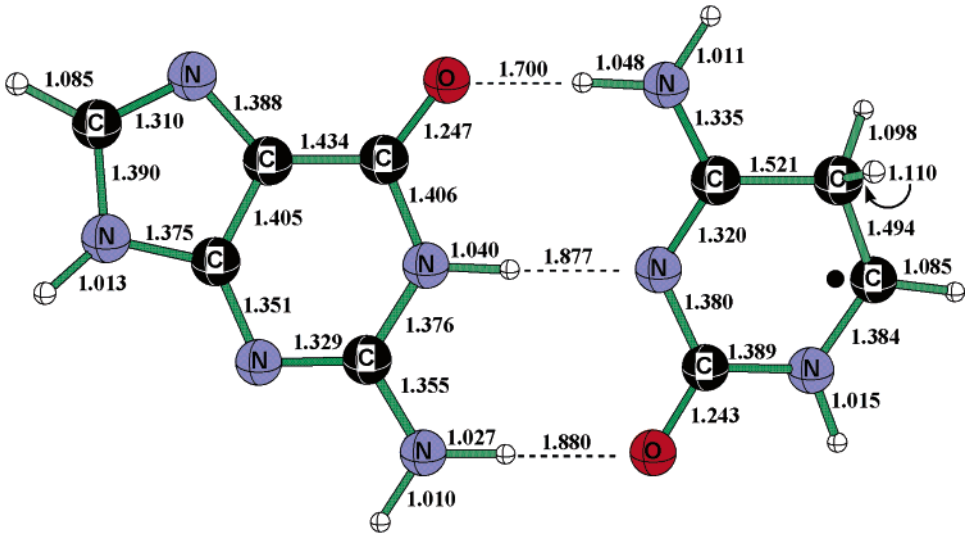


Figure 4. Optimized geometry of the G−C(C5) base-pair radical at the B3LYP/DZP++ level of theory. There is a hydrogen atom attached to cytosine C5 when compared to neutral G−C pair.

fast proton rearrangement reactions to form more thermodynamically stable species that will induce DNA tandem lesions[31a] and interstrand cross-links[31b] after trapping electrons or (b) are repaired[24,32] by thiols in biological systems.

Further investigations of damage in DNA have been extended to the Watson−Crick base pairs and the stacked double helix. Mechanisms of radical formation and further reaction channels have been discussed.[33−41] The formation of radicals sometimes involves excited triplet states of deoxyribose, which are expected to dissociate efficiently into radicals, leading to subsequent strand breaks.[35] Electron transfer through a π-stacked helix increases the chance of exciting closed-shell neutrals,[36,37] and captured electrons may cause strand breaks and mutations.[38−40] Recent research has also shown that the electron capturing probability scales with the number of guanines in the single-strand DNA oligomer.[41] Proton transfer also plays a crucial role in DNA biological processes.[42−46] However, studies of hydrogen addition to base *pairs*, the structures and energies of the different hydrogenated base-pair radicals, and how the deformed single NABs affect the pairing pattern as well as the stacked helix have not been reported.

Hydrogen bonds in the neutral guanine cytosine (G−C) base pair have been discussed extensively. Various theoretical methods have been used to determine DNA base-pair structural features.[47−51] Complex environments, including the interactions between the NAB pair and metal cations[47,48] and water molecules,[48] are required to explain the discrepancies between X-ray crystal structures and the structure of the gas-phase base pairs.[51] Combining Monte Carlo and DFT [B3LYP/6-31+G(d)] methods, Coutinho et al.[49] concluded in 2004 that the G−C interaction in an aqueous environment is weakened to about 70% of the value obtained for the isolated G−C complex. Recently, Sponer et al.[48] examined

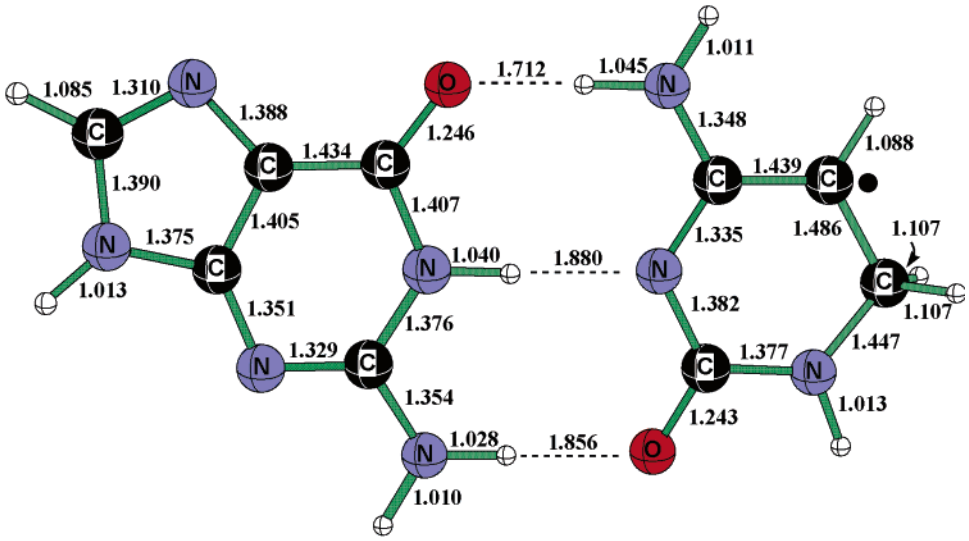


Figure 5. Optimized geometry of the G−C(C6) base-pair radical at the B3LYP/DZP++ level of theory. Relative to neutral G−C, there is a hydrogen atom attached to cytosine C6.

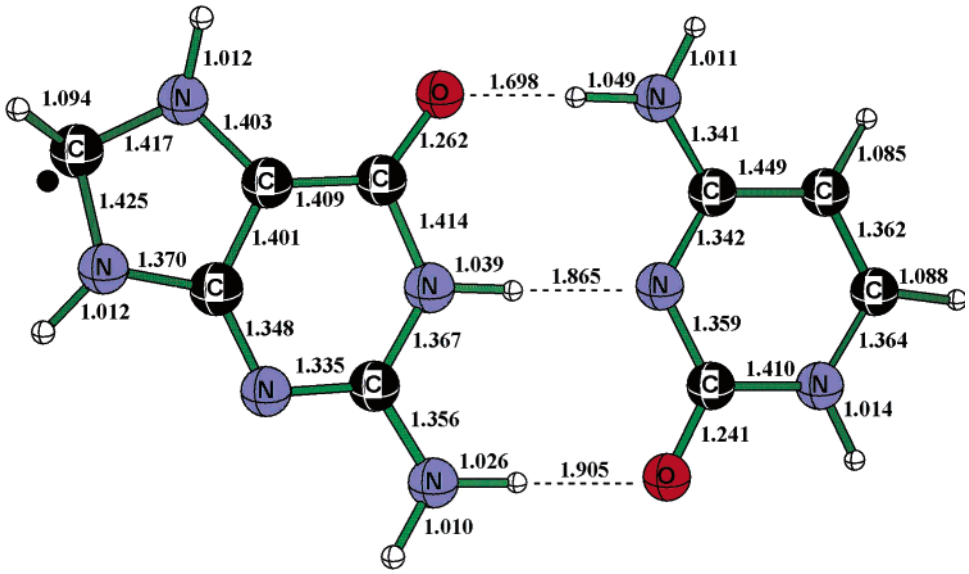


Figure 6. Optimized geometry of the G(N7)−C base-pair radical at the B3LYP/DZP++ level of theory. There is a hydrogen atom attached to guanine N7 position.

the different nucleic acid base pairs using resolution of identity (RI) MP2 methods and coupled-cluster corrections with the inclusion of noniterative triple contributions [CCSD-(T)], combined with complete basis set extrapolations.

In this research, we focus on the neutral radical isomers generated by adding one hydrogen atom to the G−C base pair, the latter depicted in Figure 1. The hydrogen atom may break C═C, C═N, or C═O double bonds on the G−C pair to yield radicals. By using the B3LYP functional combined with DZP++ basis sets, we predict the geometrical structures of 12 possible (G + H)•−C and G−(C + H)• radical isomers, as well as the dissociation energies of each pair. Vibrational frequencies were employed to characterize all optimized geometrical structures as stationary points on the potential energy surfaces. Upon attachment of one hydrogen atom to the G−C pair, the original planarity, demonstrated by gas-phase theoretical methods,[50] is destroyed. For example, a "butterfly" conformation[52] is adopted for the guanine base rings when a hydrogen atom is added at the C4 or C5 positions.

Theoretical Methods

A carefully calibrated DFT approach[28,33,50,53] has been used in this research to optimize geometries and predict vibrational frequencies. The B3LYP method is a combination of the Becke's three-parameter exchange functional (B3),[54] and the dynamic correlation functional of Lee, Yang, and Parr (LYP).[55] The Gaussian 94 system of DFT programs[56] was used for all computations.

This research was carried out using double-ζ-quality basis sets with polarization and diffuse functions, denoted as DZP++. The DZP++ basis sets were constructed by augmenting the Huzinaga−Dunning[57,58] set of constructed double-ζ Gaussian functions with one set of p-type polarization functions for each H atom and one set of five d-type

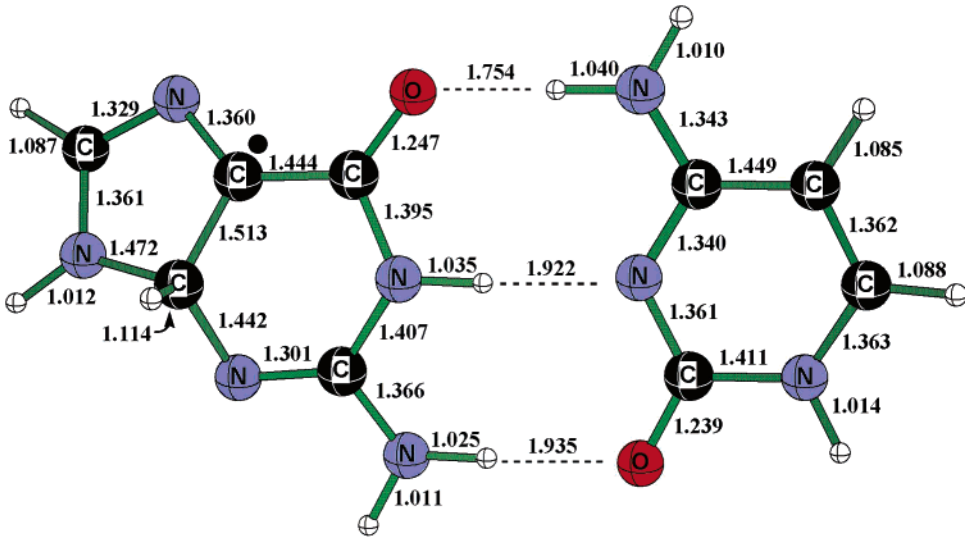


Figure 7. Optimized geometry of the G(C4)−C base-pair radical at the B3LYP/DZP++ level of theory. Relative to neutral G−C, there is a hydrogen atom attached to guanine C4.

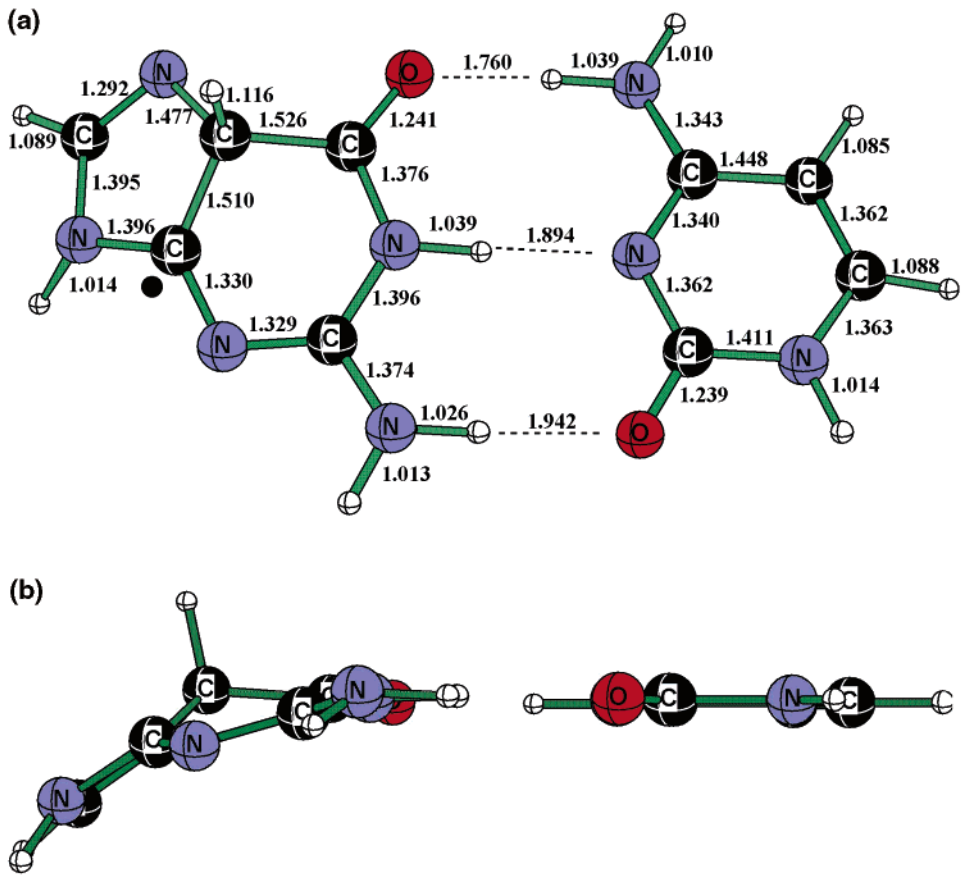


Figure 8. (a) Optimized geometry of the G(C5)−C base-pair radical at the B3LYP/DZP++ level of theory. Relative to the neutral G−C, there is a hydrogen atom attached to guanine C5. (b) Side view of the G(C5)−C radical, showing that the planarity of isolated guanine−cytosine is lost.

polarization functions for each C, N, and O atom [$\alpha_p(H) = 0.75$, $\alpha_d(C) = 0.75$, $\alpha_d(N) = 0.80$, and $\alpha_d(O) = 0.85$]. To complete the DZP++ basis, one even-tempered diffuse s function was added to each H atom while sets of even-tempered diffuse s and p functions were centered on each heavy atom. The even-tempered orbital exponents were determined by the following convention:[59]

$$\alpha_{\text{diffuse}} = \frac{1}{2}\left(\frac{\alpha_1}{\alpha_2} + \frac{\alpha_2}{\alpha_3}\right)\alpha_1$$

in which α_1, α_2 and α_3 are the three smallest primitive Gaussian orbital exponents for a given atom ($\alpha_1 < \alpha_2 < \alpha_3$). The final DZP++ set contains six functions per H atom (5s1p/3s1p) and 19 functions per C, N, or O atom (10s6p1d/5s3p1d), yielding a total of 427 contracted Gaussian functions

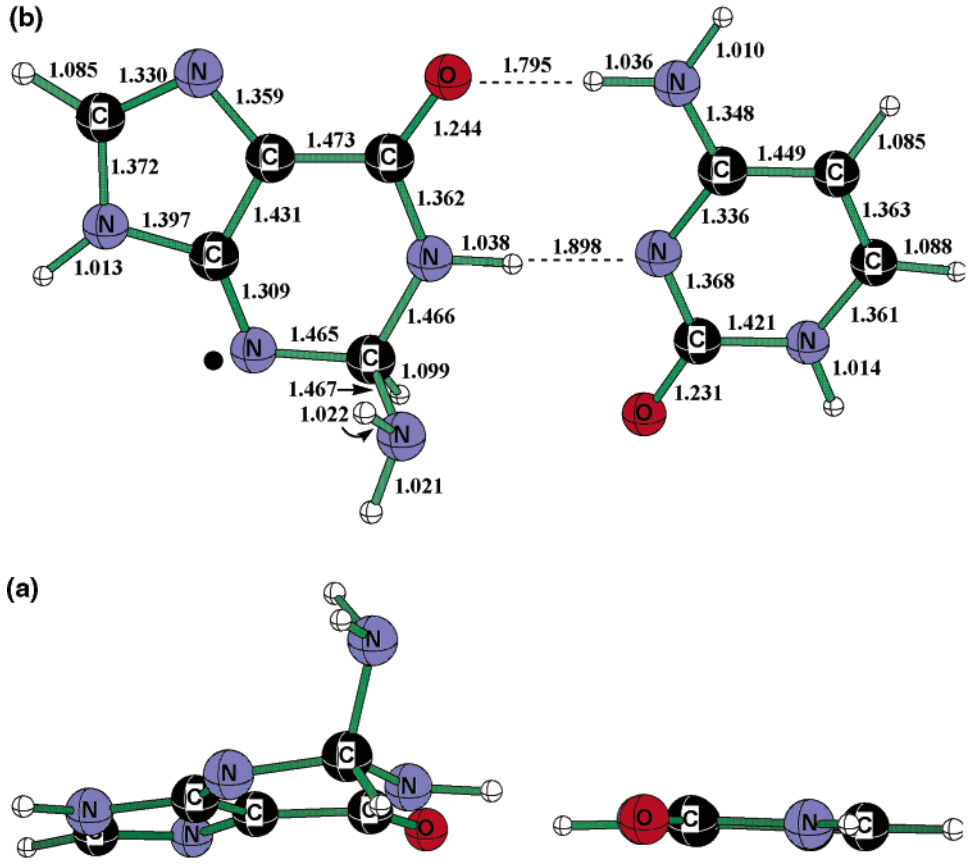


Figure 9. (a) Optimized geometry of the G(C2)−C base-pair radical at the B3LYP/DZP++ level of theory. There is a hydrogen atom attached to the guanine C2 position compared to G−C. (b) Side view of the G(C2)−C radical, showing that the amino group of guanine is significantly out of plane.

for each $(G + H)^{\bullet}-C$ and $G-(C + H)^{\bullet}$ hydrogenated base-pair radical.

The dissociation energies of the hydrogenated G-C radical base pairs were defined as

$$D_e = E[(G + H)^{\bullet}-C] - E[(G + H)^{\bullet}] - E(C) \qquad \text{(i)}$$

To analyze the unpaired electron distributions, Kohn−

$$D_e = E[G - (C + H)^{\bullet}] - E(G) - E[(C + H)^{\bullet}] \qquad \text{(ii)}$$

Sham molecular orbitals and spin-density plots were constructed from the appropriate B3LYP/DZP++ density. Natural population atomic charges were determined using the natural bond order analysis of Reed and co-workers.[60−63]

Results and Discussion

In this work, we use generic labels[33] to represent the 12 possible radical isomers resulting from hydrogen addition to the G−C base pair (Chart 1). The label G(N3)−C is meant to imply that a hydrogen atom has been added to the N3 site of the guanine moiety. The total and relative energies of the 12 radicals are reported in Table 1. The optimized geometries of the closed-shell G−C pair and the $(G + H)^{\bullet}-C$ and $G-(C + H)^{\bullet}$ radicals are schematically represented from Figures 1−13. Generally, hydrogen addition causes substantial distortions to the original C_s-symmetry G−C pair. Most H-atom addition products are of C_1 symmetry, except for G(C8)−C and G−C(C6), which retain planarity. For hydrogen-atom addition to the carbonyl group of each base, we discuss only the results associated with hydrogen-atom appendage to the oxygen atom. When a hydrogen atom adds to the carbon side of the carbonyl bond, the resulting oxygen radical will extract a hydrogen atom from its H-bond pair. The total energy of this radical is higher than that of most isomers we considered. Furthermore, as a complicated biochemical processes, the barrier-free hydrogen and proton transfers[42−46] are beyond the scope of this research.

1. Closed-Shell Guanine−Cytosine (G−C) Base Pair. The Watson−Crick model of the G−C base pair has been explored in this research using the B3LYP/DZP++ method. The optimized structure with C_s symmetry is shown in Figure 1. The three heavy-atom−heavy-atom distances associated with H bonds are predicted to be 2.767, 2.924, and 2.911 Å. It is well-known that these gas-phase H-bond distances do not match quantitatively with the crystallographic findings 2.91, 2.95, and 2.86 Å. The discrepancies between theoretical H-bond lengths and the crystal structure are due to the environmental effects, described in the important paper by Guerra et al.[47] The dissociation energies of G−C, $(G + H)^{\bullet}-C$, and $G-(C + H)^{\bullet}$ are reported in Table 2. The closed-shell G−C dissociation energy in this research is in good agreement with the definitive theoretical result of Sponer et al.[48]

2. Radicals. a. G(C8)−C. The lowest-energy radical in this series is the radical generated by the addition of hydrogen to atom C8 of guanine (Chart 1 and Figure 2). Qualitatively,

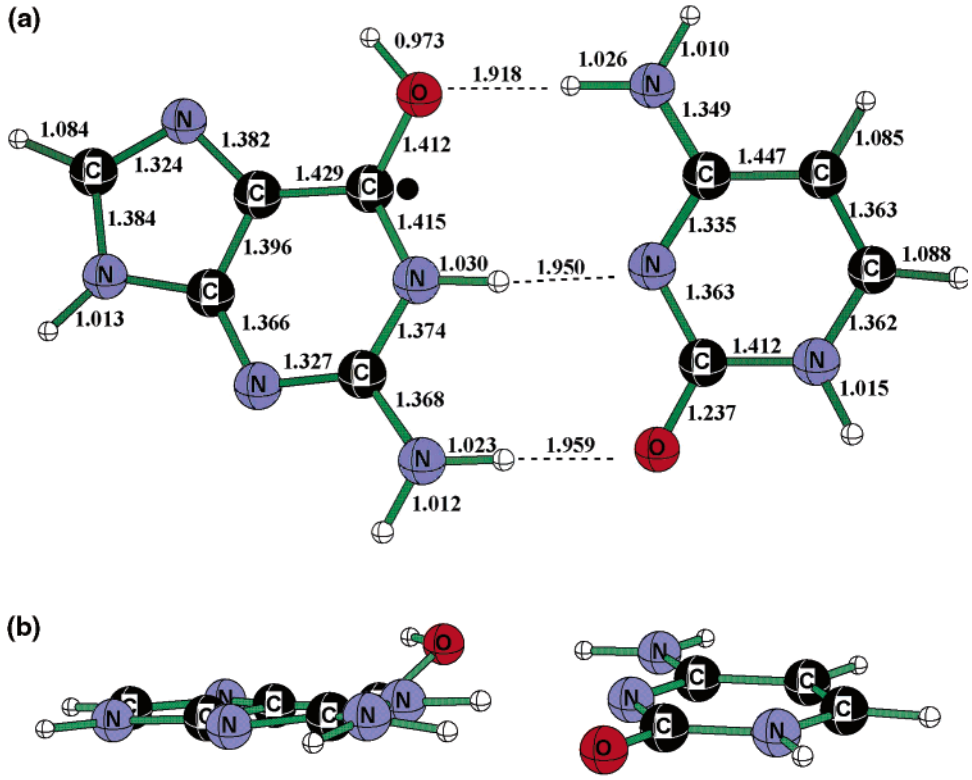


Figure 10. (a) Optimized geometry of the G(O6)−C radical at the B3LYP/DZP++ level of theory. A hydrogen atom is attached to guanine O6 position compared to G−C. (b) Side view of the G(O6)−C radical, showing the twisting of the guanine and cytosine rings.

atom C8 is changed to sp^3 hybridization from sp^2 in G−C after the C8═N7 double bond is broken. The resulting radical pair has relatively minor structural perturbations, keeping the C_s symmetry of closed-shell G−C. The (G + H)• radical is favorable to the formation of H bonds with cytosine, and the three H bonds are shortened by 0.009, 0.017, and 0.008 Å. Although the hydrogen-bonded distances are little changed from G−C to G(C8)−C, some heavy-atom−heavy-atom distances are substantially changed. Most notably, the C8═N7 distance (1.310 Å in G−C) is increased to 1.453 Å for G(C8)−C. This result is of course consistent with breaking the C8═N7 π bond of G−C. However, other structural changes are not so trivially explained. For example, the distance N9−C8 increases from 1.390 to 1.476 Å in going from G−C to the hydrogen-addition radical.

The G(C8)−C dissociation energy is 28.1 kcal/mol, slightly greater than the 27.2 kcal/mol predicted for closed-shell G−C.

b. G−C(N3). When a hydrogen atom adds to the N3 position of cytosine to break the N3═C4 double bond, the radical designated G−C(N3) is formed. Adding the hydrogen to this nitrogen atom necessarily results in the loss of the hydrogen bond N1H1···N3. The entire backbone shifts to form two new hydrogen bonds O6···H3N3 and N1H1···O2 (Figure 3). The new structure favors the formation of H bonds since the new N1H1···O2 bond length is significantly shorter (0.154 Å) than the old N2H2···O2 bond length of the original closed shell G−C pair. The pairing energy is 7.3 kcal/mol greater than that of radical G(C2)−C (Figure 9a), the other radical with two hydrogen bonds. As the second-lowest-energy radical, the total energy of the G−C(N3) radical is only 3.1 kcal/mol above that of the global minimum radical G(C8)−C (Figure 2), despite the reduction from three to two hydrogen bonds. The π bond N3═C4 increases in distance from 1.342 Å for G−C to 1.409 Å for the radical with the formal π bond broken.

c. G−C(C5). Structure G−C(C5) (Figure 4) is obtained when a hydrogen atom attaches to the cytosine C5 site. G−C(C5) is predicted to lie only 3.6 kcal/mol above the global minimum G(C8)−C (Figure 2). The G−C planarity disappears when the C5═C6 π bond is disrupted, increasing this distance from 1.362 to 1.494 Å. The adjacent C6−N1 bond length increases by 0.020 Å relative to G−C, possibly because the N1 lone-pair electrons interact with the unpaired electron formally located on cytosine C6 (with a spin density of 0.84). The overall geometrical structure of G−C(C5) preserves the original G−C pairing pattern. The two H-bond lengths O6···H4aN4 and N1H1···N3 are decreased by 0.023 and 0.009 Å, respectively, and the N2H2a···O2 H-bond distance is increased by 0.004 Å. The pairing energy is 27.9 kcal/mol, only 0.7 kcal/mol larger than that for closed-shell G−C at the same level of theory.

d. G−C(C6). This radical, shown in Figure 5, is achieved by adding a hydrogen atom to site C6 of cytosine. The total energy of G−C(C6) lies 5.8 kcal/mol above the lowest-energy radical isomer G(C8)−C (Figure 2). G−C(C6) maintains the original C_s symmetry of the closed-shell G−C pair. The spin density at atom C5 is predicted to be 0.87, indicating that most of the unpaired electron density resides in this region, consistent with our simple Lewis dot structure. As was the case for the previous radical G−C(C5) (Figure 4), the hydrogen-bond pattern is not greatly perturbed. The

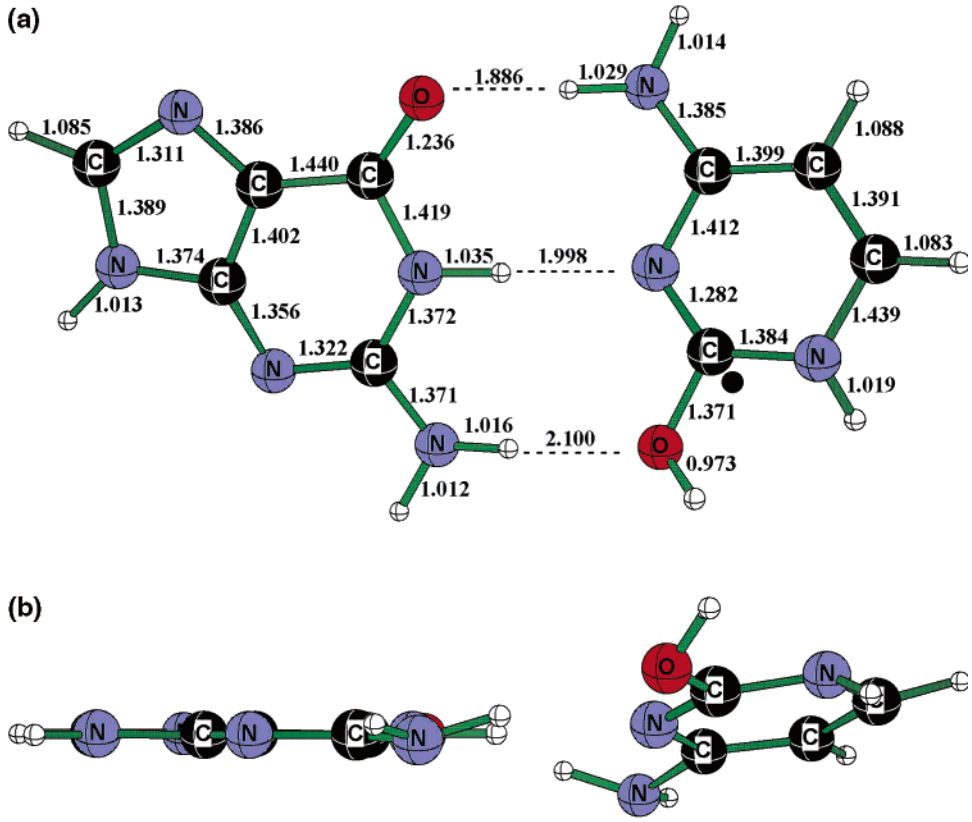


Figure 11. (a) Optimized geometry of the G−C(O2) base-pair radical at the B3LYP/DZP++ level of theory. There is a hydrogen atom attached to the cytosine O2 position compared to G−C neutral pair. (b) Side view of the G−C(O2) radical, showing the twisting of the guanine and cytosine rings.

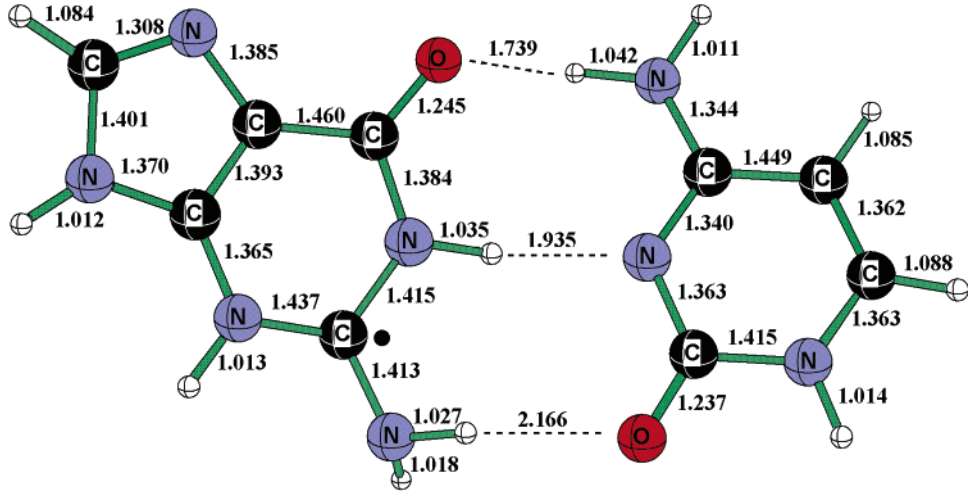


Figure 12. Optimized geometry of the G(N3)−C base-pair radical at the B3LYP/DZP++ level of theory. Relative to the neutral G−C, there is a hydrogen atom attached to guanine N3.

H-bond lengths decrease by only 0.011, 0.006, and 0.028 Å compared to G−C, while the dissociation energy is 27.7 kcal/mol, very close to that for the G−C pair (27.2 kcal/mol). The G−C double-bond distance C5═C6 (1.362 Å) increases to 1.486 Å upon qualitative breakage of the π bond.

e. G(N7)−C. This radical is formed by hydrogen-atom addition to the atom N7 of guanine (Figure 6). The π bond (C8═C7) that is broken lengthens from 1.310 Å (G−C) to 1.417 Å for the radical. Structure G(N7)−C is predicted to lie 12.1 kcal/mol higher than the global minimum G(C8)−C. The out-of-plane angle of the hydrogen attached to C8 is 43.5° (dihedral angle H8−C8−N7−C5 is 136.5°), because of the slight pyramidization of C8. The two hydrogen atoms H7 and H9 are both slightly out of plane. This radical is of biological relevance because it may be an intermediate in the formation of 8-oxo guanine, which has been found to resemble adenine and has been identified as a lesion generator.[63] As seen earlier in this research, the H-bond lengths shorten when they are away from the radical formation center. In this case, two of three H bonds are shorter by 0.025 and 0.021 Å, respectively, compared to the closed-shell G−C pair, and the third is longer by 0.021 Å.

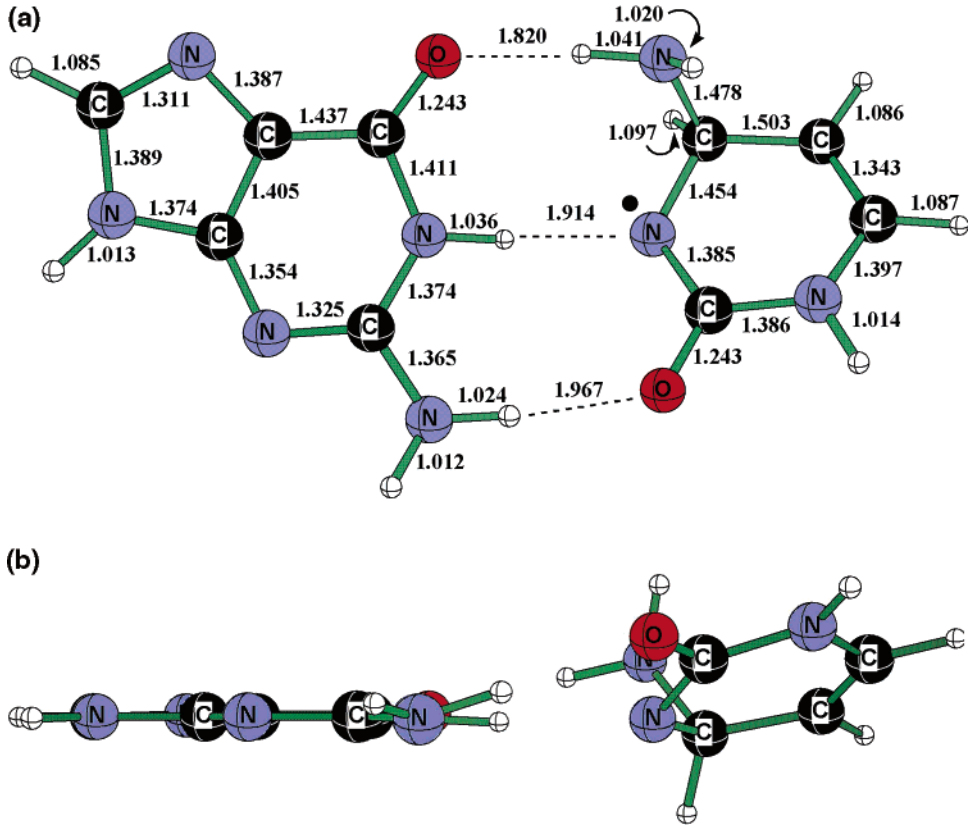


Figure 13. (a) Optimized geometry of the G−C(C4) base-pair radical at the B3LYP/DZP++ level of theory. Relative to neutral G−C, there is a hydrogen atom attached to cytosine C4. (b) Side view of the G−C(C4) radical, showing the twisting of the guanine and cytosine rings.

Table 2. Dissociation Energies of the GC Pair and the (G + H)•−C and G−(C + H)• Base-Pair Radicals (See Chart 1 for Atom Numbering)

		isolated species and energies (Eh)				
radicals	energies (Eh)	guanine/(+H) moiety		cytosine/(+H) moiety		dissociation energies (D_e) (kcal/mol)
G−C	−937.71869	G	−542.66217	C	−395.01323	27.2 27.5[a]
G(C8)−C	−938.28079	G(C8)	−543.22269	C	−395.01323	28.1
G−C(N3)	−938.27585	G	−542.66217	C(N3)	−395.57852	22.1
G−C(C5)	−938.27503	G	−542.66217	C(C5)	−395.56846	27.9
G−C(C6)	−938.27152	G	−542.66217	C(C6)	−395.56521	27.7
G(N7)−C	−938.26148	G(N7)	−543.20692	C	−395.01323	25.9
G(C4)−C	−938.25296	G(C4)	−543.20106	C	−395.01323	24.3
G(C5)−C	−938.24890	G(C5)	−543.19774	C	−395.01323	23.8
G(C2)−C	−938.24606	G(*C2*)	−543.20931	C	−395.01323	14.8
G(O6)−C	−938.24108	G(O6)	−543.19648	C	−395.01323	19.7
G−C(O2)	−938.23572	G	−542.66217	C(O2)	−395.56029	8.3
G(N3)−C	−938.23540	G(N3)	−543.19012	C	−395.01323	20.1
G−C(C4)	−938.22578	G	−542.66217	C(C4)	−395.53060	20.7

[a] This is the RI-MP2/(aDZ → aTZ) result selected by Sponer et al. in their latest ab initio research to compare with other theoretical results (see ref 48).

The pairing energy of 25.9 kcal/mol corresponds to the dissociation of the radical into the hydrogenated guanine radical plus neutral cytosine.

f. G(C4)−C. As a derivative of purine, guanine has both a six-membered ring and a five-membered ring. The two rings share the C4═C5 double bond. The radical isomer G(C4)−C (Figure 7) is produced when a hydrogen atom is added to position C4. The total energy of this isomer is 17.5 kcal/mol higher than the lowest-energy radical G(C8)−C (Figure 2). The geometrical perturbations with respect to closed-shell G−C are large because the planarity of both rings is destroyed when the shared double bond is diminished to a C−C single bond. In this process, the C4═C5 distance (1.405 Å) is increased to 1.513 Å for the radical. However, unlike the Lewis dot picture, the unpaired electron is partially distributed in the vicinity of C4 (spin density 0.42) and C2

Table 3. Relative Energies of Radicals Derived from Cytosine (C + H)• at the B3LYP/DZP++ Level of Theory

	non-ZPVE		ZPVE-corrected
radical	total energy (Eh)	relative energy (kcal/mol)	relative energy (kcal/mol)
C(N3)	−395.57852	0.0	0.0
C(C5)	−395.56846	6.3	6.2
C(C6)	−395.56521	8.4	8.0
C(O2)	−395.56029	11.4	11.0
C(C4)	−395.53060	30.1	29.8

Table 4. Relative Energies of Radicals Derived from Guanine (G + H)• at the B3LYP/DZP++ Level of Theory

	non-ZPVE		ZPVE-corrected
radical	total energy (Eh)	relative energy (kcal/mol)	relative energy (kcal/mol)
G(C8)	−543.22269	0.0	0.0
G(C2)	−543.20931	8.4	9.2
G(N7)	−543.20692	9.9	10.4
G(C4)	−543.20106	13.6	13.9
G(C5)	−543.19774	15.7	15.4
G(O6)	−543.19648	16.4	16.9
G(N3)	−543.19012	20.4	20.9

(spin density 0.26). All three hydrogen-bond distances increase, by 0.031, 0.036, and 0.051 Å compared to the closed-shell G−C, and the dissociation energy (Table 2) of G(C4)−C is 24.3 kcal/mol.

g. G(C5)−C. Structure G(C5)−C (Figure 8a) is the other radical isomer resulting when the shared double bond of the purine is broken. The resulting C4−C5 single bond distance (1.510 Å) is 0.003 Å less than that just discussed. The total energy of this isomer is 20.0 kcal/mol high than the lowest-energy radical G(C8)−C (Figure 2). As found for the previous radical isomer G(C4)−C, the inclined geometry of the two rings in guanine is not a simple open-book geometry but a twisted open-book structure (Figure 8b illustrates the nonplanarity). The three hydrogen-bond distances lengthen by 0.037, 0.008,and 0.058 Å compared to G−C. The predicted dissociation energy of 23.8 kcal/mol is also very comparable to that of G(C4)−C, namely, only 0.5 kcal/mol higher.

h. G(C2)−C. In the formation of this radical, the hydrogen atom breaks the guanine C2═N3 double bond by adding to the C2 site. The total energy of G(C2)−C is predicted to be 21.8 kcal/mol higher than the lowest-energy radical G(C8)−C. The resulting minimum undergoes a qualitative structural change (Figure 9a) compared to closed-shell G−C pair. In the new radical, the two hydrogen atoms of the guanine amino group are so far from the cytosine moiety that no N2H2a···O2 hydrogen bond remains. If this radical can be formed in the DNA damage process, the resulting adduct will cause significant strain (see Figure 9b) in the closely stacked base pairs (∼3.4 Å stacking distance[51]), which may result in an SSB of the DNA backbone. The remaining two H bonds (O6···H4aN4 and N1H1···N3) are somewhat elongated (by 0.072 and 0.012 Å) compared to those for the closed-shell G−C. The radical dissociation energy for G(C2)−C is low, 14.8 kcal/mol, since only the two hydrogen bonds remain (Table 2).

i. G(O6)−C. The radical resulting from hydrogen-atom addition to the carbonyl oxygen of guanine, structure G(O6)−C (Figure 10a), is predicted to lie 24.9 kcal/mol higher than the energy sink G(C8)−C on the potential surface for hydrogen addition to G−C. In structure G(O6)−C, atom O6 is transformed from the oxo to hydroxyl form. As a result, the C6−O6 bond distance is lengthened by 0.166 Å compared to closed-shell G−C. The atom C6 carries a spin density of 0.90, still lying in the purine plane, consistent with the simple Lewis dot structure. The three H-bond distances are 1.918, 1.950 and 1.959 Å. Nevertheless, the twisted pairing geometry (Figure 10b) is not favorable for hydrogen bonding. The pairing (dissociation) energy is only 19.7 kcal/mol, which is the second lowest among the radical isomers retaining three hydrogen bonds.

j. G−C(O2). Structure G−C(O2) (Figure 11a) is formed by hydrogen-atom addition to the oxygen atom of the cytosine carbonyl group. The G−C distance O2═C2 increases from 1.241 to 1.371 Å for the newly formed radical. The attached H2 atom on the O2 site is out of plane by 30.5° (dihedral angle H2−O2−C2−N1). Because of the steric interaction with atom H2, the cytosine H1 atom also shows an obvious distortion from the ring plane by 37.1° (dihedral angle H1−N1−C2−O2). The total energy of G−C(O2) is 28.3 kcal/mol above that of the global minimum G(C8)−C (Figure 2). The NAB pairing interaction also shows interesting features. The theoretical dissociation energy is only 8.3 kcal/mol, much lower than other radical isomers with three H bonds. Compared with the isolated hydrogen-added cytosine isomer C(O2 + H), the geometry of the cytosine moiety in G−C(O2) is different because of the steric interaction between the additional hydrogen atom H2 and the cytosine atom H1. The guanine and cytosine rings twist (Figure 11b), because of the pyramidization of the amino group in cytosine. As for G(O6)−C, the twisted pairing geometry is not favored for H-bond formation, with 0.163, 0.112, and 0.216 Å elongations in bond distances compared to the neutral G−C pair.

k. G(N3)−C. When the hydrogen atom adds to the guanine N3 atom, a radical formally centered at C2 is produced (Figure 12). Breaking the N3═C2 π bond increases this distance by 0.108 Å. This radical is predicted to lie 28.5 kcal/mol above the lowest-energy radical G(C8)−C. The guanine amino group changes from planar to nonplanar partly because of the new sp^3 hybridization at C2. The guanine amino group adjacent to the radical center moves slightly out of plane to diminish the nitrogen lone-pair electron interaction with the localized single electron of C2 (with a spin density of 0.88, following the simple valence picture). As a result, the C2−N2 bond rotation causes the N2H2a••·O2 distance to be lengthened by 0.282 Å. The other two H bonds are increased by 0.016 and 0.049 Å when compared with the closed-shell G−C pair (Figure 1). An energy of 20.1 kcal/mol (Table 2) is required to dissociate this radical base pair, compared to 27.2 kcal/mol for the neutral G−C base pair.

l. G-C(C4). Structure G−C(C4) (Figure 13a) is the isomer with the highest energy of our family of 12 radicals, lying 34.5 kcal/mol higher than the global minimum radical G(C8)−C structure. G−C(C4) is generated by attaching a hydrogen atom to the C4 position of cytosine. Like the previously discussed radical G(C2)−C (Figure 9a), the additional hydrogen causes the carbon atom which connects to the amino group to be transformed qualitatively from sp^2 to sp^3 hybridization. This lengthens the G−C bond N3═C4 from 1.342 to 1.454 Å for the new radical. The amino group thus moves out of the cytosine plane by 78.1° (dihedral angle N4−C4−C5−C6 is 101.9°). The H bond O6···H4aN4 still exists, unlike the vanishing N2H2a···O2 bond in G(C2)−C (Figure 9a). The dissociation energy is 20.7 kcal/mol, about 6.0 kcal/mol higher than that of the radical G(C2)−C with two H bonds. The planarity of the G−C pair has been lost in order to form three H bonds, since the cytosine amino group is significantly out of plane. The guanine and cytosine are twisted as seen in Figure 13b.

Conclusions

In general, the geometrical perturbations are significant in going from the neutral G−C pair to the (G + H)•−C and G−(C + H)• radicals. A total of 10 of the 12 isomers deviate from the original C_s symmetry of the G−C pair in the gas phase. The other two radical isomers, G(C8)−C and G−C(C6), maintain C_s symmetry. Classified by the hydrogen-atom addition site, the 12 radical isomers may be divided into three groups: (1) The hydrogen atom adds to an atom that is close to or directly related to the three interstrand H bonds of the base pair. (2) The hydrogen atom adds to the C4═C5 bridgehead double bond of guanine. (3) The hydrogen atom adds to one of the five atoms away from the G−C hydrogen bonds.

Among the three groups, radicals from the first group undergo the largest structural changes and are the least favored energetically [relative energies ranged 21.8∼34.5 kcal/mol, except for G-C(N3)]. The formation of radicals G(C2)−C (Figure 9a) and G−C(C4) (Figure 13a) may cause significant strain in $\pi-\pi$ stacked DNA bases. Another radical, G−C(N3) (Figure 3), which changes the base pairing sequence, may cause a major lesion of DNA, so as to shift the entire G−C backbone. Compared to the H bonds of neutral G−C (dissociation energy ∼27.0 kcal/mol), all [except G−C(N3)] H bonds in this group show longer bond distances and lower dissociation energies, which can be as small as 8.3 kcal/mol [G−C(O2)].

The two radicals, G(C4)−C (Figure 7) and G(C5)−C (Figure 8), classified in our second group show intermediate total energies (relative energies 17.5 and 20.0 kcal/mol) and pairing energies. The significant geometry disturbances are attributed to the "butterfly" structure of the two-ring guanine system. Since the planarity of guanine is gone, $\pi-\pi$ stacking conjugation may serve to create vertical base−base interactions.

The third group, composed of five radicals, has the lowest relative energies (below 12.1 kcal/mol) and highest dissociation energies (above 25.9 kcal/mol). The planar structure is maintained for radicals G(C8)−C and G−C(C6), since both are formed without qualitative perturbation of the three H bonds. Most H-bond lengths for these radical isomers are even shorter than those of G−C. As the most probable products of NAB lesion following hydrogen-atom attachment, these structures suggest that the formation of hydrogenated G−C pairs may not cause immediate G−C dissociation.

Acknowledgment. This research was supported by the National Science Foundation, Grant CHE-0451445. H.F.S. thanks Professor Martin Quack for hosting him at the ETH Zürich while this manuscript was being prepared.

Appendix

Tables 3 and 4 show the relative energies of radicals derived from cytosine (C + H)• and guanine (G + H)• at the B3LYP/DZP++ level of theory.

References

(1) (a) O'Neill, P.; Fielden, M. *Adv. Radiat. Biol.* **1993**, *17*, 53. (b) Becker, D.; Sevilla, M. D. *Adv. Radiat. Biol.* **1993**, *17*, 121. (c) Colson, A. O.; Sevilla, M. D. *Int. J. Radiat. Biol.* **1995**, *67*, 627. (d) Sanche, L. *Mass Spectrom. Rev.* **2002**, *21*, 349.

(2) von Sonntag, C. *The Chemical Basis of Radiation Biology*; New York, 1987.

(3) Hieda, K. *Int. J. Radiat. Biol.* **1994**, *66*, 561.

(4) Cai, Z.; Cloutier, P.; Hunting, D.; Sanche, L. *J. Phys. Chem. B* **2005**, *109*, 4796.

(5) Steenken, S.; Goldbergerova, L. *J. Am. Chem. Soc.* **1998**, *120*, 3928.

(6) Douki, T.; Angelov, D.; Cadet, J. *J. Am. Chem. Soc.* **2001**, *123*, 11360.

(7) Zheng, Y.; Cloutier, P.; Hunting, D. J.; Sanche, L.; Wagner, J. R. *J. Am. Chem. Soc.* **2005**, *127*, 16592.

(8) Abdoul-Carime, H.; Cloutier, P.; Sanche, L. *Radiat. Res.* **2001**, *155*, 625.

(9) Abdoul-Carime, H.; Sanche, L. *J. Radiat. Res.* **2002**, *78*, 89.

(10) Huels, M. A.; Boudaiffa, B.; Cloutier, P.; Hunting, D.; Sanche, L. *J. Am. Chem. Soc.* **2003**, *125*, 467.

(11) Gohlke, S.; Illenberger, E. *Europhys. News* **2002**, *33*, 207.

(12) (a) Evangelista, F. A.; Paul, A.; Schaefer, H. F. *J. Phys. Chem. A* **2004**, *108*, 3565. (b) Close, D.; Forde, G.; Gorb, L.; Leszczynski, J. *Struct. Chem.* **2003**, *14*, 451.

(13) (a) Slupphaug, G.; Kavil, B.; Krokan, H. E. *Mutat. Res.* **2003**, *531*, 231. (b) Garrett, B. C.; Dixon, D. A. *Chem. Rev.* **2005**, *105*, 335.

(14) Purkayastha, S.; Milligan, J. R.; Bernhard, W. A. *J. Phys. Chem. B* **2005**, *109*, 16967.

(15) Karagiannis, T. C.; El-Osta, A. *Cell. Mol. Life Sci.* **2004**, *61*, 2137.

(16) Llano, J.; Eriksson, L. A. *Phys. Chem. Chem. Phys.* **2004**, *6*, 4707.

(17) Aydogan, B.; Marshall, D. T.; Swarts, S. G.; Turner, J. E.; Boone, A. J.; Richards, N. G.; Bolch, W. E. *Radiat. Res.* **2002**, *157*, 38.

(18) Nikjoo, H.; O'Neill, P.; Goodhead, D. T.; Terrissol, M. *Int. J. Radiat. Biol.* **1997**, *71*, 467.

(19) Fulford, J.; Nikjoo, H.; Goodhead, D. T.; O'Neill, P. *Int. J. Radiat. Biol.* **2001**, *77*, 1053.

(20) Malins, D. C.; Hellstrom, K. E.; Anderson, K. M.; Johnson, P. M.; Vinson, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 5937.

(21) Milligan, J.; Ng, J.; Wu, C.; Augilera, J.; Fahey, R.; Ward, J. *Radiat. Res.* **1995**, *143*, 273.

(22) Jean, C.; Thierry, D.; Didier, G.; Jean-Luc, R. *Mutat. Res.* **2003**, *531*, 5.

(23) Adam, W.; Arnold, M. A.; Nau, W. M.; Pischel, U.; Saha-Moller, C. R. *J. Am. Chem. Soc.* **2002**, *124*, 3893.

(24) Steenken, S. *Chem. Rev.* **1989**, *89*, 503.

(25) Colson, A. O.; Sevilla, M. D. *J. Phys. Chem.* **1996**, *100*, 4420.

(26) Wetmore, S. D.; Boyd, R.; Eriksson, L. A. *J. Phys. Chem. B* **1998**, *102*, 9332.

(27) (a) Turecek, F.; Yao, C. *J. Phys. Chem. A* **2003**, *107*, 9221. (b) Chen, X.; Syrstad, E. A.; Nguyen, M. T.; Gerbaux, P. G.; Turecek, F. *J. Phys. Chem. A* **2005**, *109*, 8121.

(28) Rienstra-Kiracofe, J. C.; Tschumper, G. S.; Schaefer, H. F.; Nandi, S.; Ellison, G. B. *Chem. Rev.* **2002**, *102*, 231.

(29) Debije, M. D.; Bernhard, W. A. *J. Phys. Chem. A* **2002**, *106*, 4608.

(30) Hole, E. O.; Nelson, W. H.; Close, D. M. *J. Phys. Chem.* **1992**, *96*, 8269.

(31) (a) Cater, K. N.; Greenberg, M. M. *J. Am. Chem. Soc.* **2003**, *125*, 13376. (b) Hong, I. S.; Greenberg, M. M. *J. Am. Chem. Soc.* **2005**, *127*, 3692.

(32) Becker, D.; Summerfield, S.; Gillich, S.; Sevilla, M. D. *Int. J. Radiat. Biol.* **1994**, *65*, 537.

(33) Bera, P. P.; Schaefer, H. F. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6698.

(34) Brocklehurst, B. *Radiat. Res.* **2001**, *155*, 637.

(35) Tuppurainen, K.; Lotjonen, S.; Laatikainen, R.; Vartiainen, T.; Maran, U.; Strandberg, M.; Tamm, T. *Mutat. Res.* **1991**, *247*, 97.

(36) (a) Kelley, S. O.; Barton, J. K. *Science* **1999**, *283*, 375. (b) Shao, F.; O'Neill, M. A.; Barton, J. K. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 17914.

(37) Ratner, M. *Nature* **1999**, *397*, 480.

(38) Anusiewicz, I.; Berdys, J.; Sobczyk, M.; Skurski, P.; Simons, J. *J. Phys. Chem. A* **2004**, *108*, 11381.

(39) Gu, J.; Xie, Y.; Schaefer, H. F. *J. Am. Chem. Soc.* **2005**, *127*, 1053.

(40) Ray, S. G.; Daube, S. S.; Naaman, R. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15.

(41) Zoete, V.; Meuwly, M. *J. Chem. Phys.* **2004**, *121*, 4377.

(42) Gorb, L.; Podolyan, Y.; Dziekonski, P.; Sokalski, W. A.; Leszczynski, J. *J. Am. Chem. Soc.* **2004**, *126*, 10119.

(43) Florian, J.; Leszczynski, J. *J. Am. Chem. Soc.* **1996**, *118*, 3010.

(44) Taylor, J.; Eliezer, I.; Sevilla, M. *J. Phys. Chem. B* **2001**, *105*, 1614.

(45) Hutter, M.; Clark, T. *J. Am. Chem. Soc.* **1996**, *118*, 7574.

(46) Gu, J.; Wang, J.; Leszczynski, J. *J. Am. Chem. Soc.* **2004**, *126*, 12651.

(47) Guerra, C. F.; Bicklhaupt, F. M.; Snijders, J. G.; Baerends, E. J. *J. Am. Chem. Soc.* **2000**, *122*, 4117.

(48) Sponer, J.; Jurecka, P.; Hobza, P. *J. Am. Chem. Soc.* **2004**, *126*, 10142.

(49) Coutinho, K.; Ludwig, V.; Canuto, S. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2004**, *69*, 61902.

(50) Richardson, N. A.; Wesolowski, S. S.; Schaefer, H. F. *J. Am. Chem. Soc.* **2002**, *124*, 10163.

(51) Saenger, W. *Principles of Nucleic Acid Structures*; Springer-Verlag: New York, 1984.

(52) Colson, A. O.; Sevilla, M. D. *J. Phys. Chem.* **1996**, *100*, 4420.

(53) Gu, J.; Xie, Y.; Schaefer, H. F. *J. Phys. Chem. B* **2005**, *109*, 13067.

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

(55) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.

(56) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T. A.; Petersson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Cioslowski, J.; Stefanov, B. B.; Nanayakkara, A.; Challacombe, M.; Peng, C. Y.; Ayala, P. A.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. P.; Head-Gordon, M.; Gonzalez, C.; Pople, J. A. *Gaussian 94*, revision c.3; Gaussian, Inc.: Pittsburgh, PA, 1995.

(57) Huzinaga, S. *J. Chem. Phys.* **1965**, *42*, 1293.

(58) Dunning, T. H. *J. Chem. Phys.* **1970**, *53*, 2823.

(59) Lee, T. J.; Schaefer, H. F. *J. Chem. Phys.* **1985**, *83*, 1784.

(60) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735.

(61) Reed, A. E.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 1736.

(62) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899.

(63) Reed, A. E.; Schleyer, P. R. *J. Am. Chem. Soc.* **1990**, *112*, 1434.

(64) Malins, D. C.; Polissar, N. L.; Ostrander, G. K.; Vinson, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 12442.

CT600262P

JCTC Journal of Chemical Theory and Computation

Evaluation of Two Computational Models Based on Different Effective Core Potentials for Use in Organocesium Chemistry

Andrew Streitwieser,* Joe Chih-Yao Liang, E. G. Jayasree, and Faraj Hasanayn

Department of Chemistry, University of California, Berkeley, California 94720-1460

Received July 20, 2006

Abstract: The performance of two computational models was evaluated in describing some aggregated structures, the bond lengths and dimerization energies of cesium halides, aquation energies of the cesium cation, and protonation energies of a range of organocesium compounds. One model used the Hay−Wadt (HW) effective core potential (ECP) and a double-ζ valence basis set on Cs; the other used the Ross ECP with two polarization functions on Cs. In both models, the standard 6-31+G** basis was used for the other atoms. At the Hartree−Fock (HF) level, the Ross ECP was found to give geometries and energies in good agreement with experimental results. Second-order Møller−Plesset calculations with this model gave only modestly improved results compared to HF; the B3LYP level gave variable results with unsatisfactory energies. Although the HW model is generally less satisfactory, it often shows comparable trends to those of the Ross model.

Introduction

Over the course of a number of years, we have contributed a number of quantitative data on organocesium chemistry, particularly in cesium ion-pair carbon acidity,[1] the aggregation of cesium enolates,[2−8] and kinetic acidities with cesium cyclohexylamide.[9] If these results could be modeled effectively with ab initio computations at some reasonable theory level, then they could presumably be extrapolated to other compounds and further extend their usefulness. We are currently engaged in such modeling. To this end, it is important to have an appropriate effective core potential (ECP)[10] for cesium. The use of ECPs reduces the number of electrons to be treated explicitly in electronic structure calculations and, perhaps more importantly, can account for relativistic effects essential in treating the cores of heavy elements. The traditional and most widely used set of ECPs is that developed by Hay and Wadt (HW)[11] for most of the elements in the periodic table. Several other ECPs have been developed in recent years, of which that developed by Ross et al. (Ross)[12] has also been widely used. Both ECPs are calibrated for relativistic effects, but the Ross potential includes spin−orbit effects. The HW ECP has been used, for example, in studies of Cs^+ complexed to ethers,[13,14] liquid cesium,[15] cesium boranes,[16] ^{19}F NMR spectra of Cs salts,[17] and the effect of pressure on cesium sulfide.[18] Similarly, the Ross ECP has been used in studies of diatomic cesium,[19] crown ethers complexed to Cs^+,[20] and the excitation of CsI.[21]

Several years ago, we began calculations to model our ion-pair acidity studies of organocesium compounds. For this purpose, we used Hartree−Fock (HF) calculations using the HW ECP for Cs plus a double-ζ basis set for the Cs valence shell and the standard 6-31+G** basis for the other atoms. Results from this model were not satisfactory, and we repeated the calculations using the larger Ross ECP. This ECP includes a polarization d function with four primitives. We introduced greater flexibility by spitting this d function in two by liberating the penultimate primitive as a separate function. The result was a significant improvement over the HW model. In the present paper, we compare results from both of these computational models symbolized as HW and Ross.

Except for a limited study of arylcesiums,[22] we are not aware of any comparisons of these basis sets for organocesium chemistry.

* Corresponding author fax: (510) 841-6072; e-mail: astreit@berkeley.edu.

10.1021/ct600238v CCC: $37.00
Published on Web 12/16/2006

Table 1. Comparison of ECPs with the Experimental Bond Lengths (Å) for Cesium Diatomics (CsX)

	HW				Ross			
X	HF	MP2	B3LYP	exptl	HF	MP2	B3LYP	Ross HF − d[a]
H	2.762	2.770	2.710	2.494	2.597	2.553	2.527	2.716
F	2.657	2.703	2.664	2.345	2.445	2.434	2.414	2.575
Cl	3.241	3.252	3.214	2.906	3.053	3.007	2.981	3.182
Br	3.377	3.381	3.340	3.072	3.210	3.156	3.136	3.315

[a] Ross ECP with d functions deleted from the ECP basis.

Computational Methods

Calculations were done using various editions of the Gaussian series of programs up to Gaussian 03.[23] Geometries at all theory levels used were optimized using the 6-31+G** basis set on all atoms except cesium, for which the modified HW and Ross ECPs were used. The HW ECP included the double-ζ valence basis set as supplied in Gaussian with the LANL2DZ keyword. The basis set provided by the Ross ECP was split into s 4/4/1, p 3/1/1, and d 3/1 and provides greater flexibility than the HW set. We are not concerned here with spin−orbit effects, and this part of the Ross ECP was not used. The complete sets used are tabulated in Table S1 (Supporting Information).

The study involved HF, second-order Møller−Plesset (MP2), and B3LYP computations. In each case, all structures were fully optimized and shown to be local minima on the potential energy surface by frequency calculations (zero imaginary frequencies). The structure coordinates and energies for all computations are summarized in Table S2 (Supporting Information).

Results and Discussion

A simple study of some cesium diatomics shows immediately that the two computational models give quite different results (Table 1). At the HF level, as well as with MP2 and the popular hybrid density functional theory method, B3LYP, the HW model gives computed bond distances much larger than experimental results. The Ross model does much better and with the incorporation of correlation at the MP2 level gives bond distances that average only 0.08 Å greater than experimental results. The ability of MP2 to give good bond lengths was shown previously for CsH using a different ECP.[24] Computations of the cesium halides with the Ross ECP and a somewhat smaller basis set for the halides at MP2 have been published previously.[25] B3LYP computations with the Ross ECP give good bond distances for the cesium halides, but we'll see below that this promise is not sustained in computations of reaction energies.

Much of the improvement with the Ross ECP appears to be related to the inclusion of d functions. When these are deleted from the ECP, the CsX bond distances are significantly longer and start to resemble the HW results (Table 1). A much larger ECP has recently been proposed that adds both f and g functions to the outermost nine electrons of Cs.[26] We have not compared this new ECP.

Some energy quantities are available for a further comparison. Gas-phase aquation energies, eq 1, are known for cesium cations with up to four waters.[27] Feller et al.,[28] and an earlier paper of Glendening and Feller,[29] have provided an extensive study of water clusters with alkali cations up to eight waters. They compared several theory levels and basis sets but for Cs used the Hay−Wadt ECP augmented with polarization functions. The resulting ECP is thus actually closer to our Ross model. They also showed that the 6-31+G* basis set with the other elements can give satisfactory results. These studies described a number of types of structures including those with hydrogen bonds among waters, particularly for higher numbers of coordinated waters. In our present study, to compare the two Cs ECPs, we have focused on Cs^+ with up to four waters of coordination sufficiently separated so that hydrogen bonding is not involved.

The comparisons in Table 2 show that the Ross ECP gives better quantitative results. The Ross aquation energies are lower than experimental values by only 1 kcal mol^{-1} per water. The MP2 results are only slightly better. With increasing waters, the Cs−O bond distance increases and the aquation energy per water decreases as expected. The HW Cs−O bond distances are systematically larger, and the computed aquation energies correspondingly are lower but only by about another 1−2 kcal mol^{-1}. The MP2 results differ only slightly from those of Ross HF. In all cases, the bonding increases with a shorter Cs−O distance; these are shortest with Ross MP2 and longest with HW. The Ross B3LYP results show aquation energies that are significantly lower.

$$Cs^+ + nH_2O = Cs^+ \cdot nH_2O \qquad (1)$$

Of particular interest are some of the structures. $Cs^+ \cdot 2H_2O$ with the Ross ECP is bent with an O−Cs−O angle of 113.5°. The same result was found by Kaupp and Schleyer[30] using larger basis sets. Similar bent structures were found at several theory levels by Feller et al.[28,29] Such bent structures are normally attributed to polarization of the central cation. In this connection, it is interesting that, with the HW ECP, without additional polarization functions, the minimum structure of $Cs^+ \cdot 2H_2O$ is D_{2d} with O−Cs−O = 180°. Similarly, $Cs^+ \cdot 3H_2O$ with the Ross ECP is pyramidal with O−Cs−O = 103.7°, as found by others,[28−30] whereas the minimum structure with HW is D_3 with oxygens and cesium in the same plane. These results would seem to confirm the role of polarization in cesium structures and the importance of including polarization functions.

Dimerization energies of the cesium halides provide a further opportunity for ECP comparison and are a more complex function of bond distances since the Cs−X distances are different in monomers and dimers. These monomers and dimers have been treated frequently in the past, especially with various ionic models.[31] Thermodynamic measurements are almost all several decades old and were tabulated in 1974 by Kondrat'ev.[32] We know of only one measurement (for CsBr[33]) since then. The available results are compared to computed values at several levels of theory in Table 3. It is difficult to compare the experimental results to theory since results from different laboratories vary by several kilocalories per mole. Moreover, the experiments were all done at high temperatures, although some were extrapolated to lower

Table 2. Enthalpies of Reaction of Cesium Cation with 1−4 Molecules of Water; Energies in kcal mol^{-1}

compound	$-\Delta E_{aq}$ exptl[a]	Ross $-\Delta E$[b]	r(CsO), Å	HW $-\Delta E$[b]	r(CsO), Å	Ross, MP2 $-\Delta E$[c]	r(CsO), Å	Ross, B3LYP $-\Delta E$[d]	r(CsO), Å
$Cs^+ \cdot H_2O$	13.7	12.91	3.058	11.96	3.154	13.22	2.985	12.67	2.996
$Cs^+ \cdot 2H_2O$	26.2	24.29	3.096	22.86	3.190	24.87	3.014	23.85	3.032
$Cs^+ \cdot 3H_2O$	37.4	34.35	3.132	32.59	3.211	35.09	3.040	33.62	3.072
$Cs^+ \cdot 4H_2O$	48.0	43.20	3.166	41.19	3.238	44.35	3.074	41.98	3.103

[a] ref 27. [b] $\Delta E = E(Cs^+ \cdot nH_2O) - E(Cs^+) - nE(H_2O)$ including unscaled ZPE; HF 6-31+g(d,p) + Cs ECP. [c] $\Delta E = E(Cs^+ \cdot nH_2O) - E(Cs^+) - nE(H_2O)$ including unscaled ZPE; MP2 6-31+g(d,p) + Cs ECP. [d] $\Delta E = E(Cs^+ \cdot nH_2O) - E(Cs^+) - nE(H_2O)$ including unscaled ZPE; B3LYP 6-31+g(d,p) + Cs ECP.

Table 3. Dissociation Energies (kcal mol^{-1}) of Cesium Halide Dimers to Monomers, Experimental and Theoretical

compound	$\Delta H_d°$ exptl	ΔE_d calcd[a]	ΔE_d HW[b]	ΔE_d Ross[b]	ΔE_d Ross-MP2[b]	ΔE_d Ross-B3LYP[b]
Cs_2F_2	37.8 ± 1.3 (863 K)[c] 42.3 (298 K)[d] 38.8 ± 2 (298 K) 38.9 (0 K)[e]	43.4	46.35	40.04	39.92	35.56
Cs_2Cl_2	38.9 (815 K)[f] 34.7 ± 1.1 (1300 K)[g] 40.1 (1300 K)[h] 36.2 ± 1.0 (298 K) 36.7 (0 K)[c]	42.2	40.60	36.76	36.58	30.55
Cs_2Br_2	40.4 (1300 K)[h] 39.4 (1600 K)[h] 38.8 ± 2.4 (298 K)[i] 34.5 ± 1.0 (298 K) 34.7 (0 K)[e]	39.3	39.84	38.53	38.23	34.67

[a] 6-31G* + ECP(Cs, Cl, Br) + MP2; ref 25. [b] This work. HF, MP2, or B3LYP, 6-31+G(d) + unscaled ZPE. [c] Ref 39. [d] Ref 40. [e] Ref 35. [f] Ref 41. [g] Ref 42. [h] Ref 34. [i] Ref 33.

temperatures using thermodynamic functions. The theoretical results all pertain to 0 K. We could try to discern trends from the results of two compounds in the same laboratory. The Romanian group[34] reports almost equal values for CsCl and CsBr. Kondrat'ev[32] reports values for all of the cesium halides given in a dissertation[35] but without details. These results show a monotonic decrease in dissociation energies from Cs_2F_2 to Cs_2I_2. The theoretical results of Hilpert et al.[33] also show such a monotonic decrease, but these results are known to contain a basis set superposition error (BSSE) of several kilocalories per mole. Our use of diffuse functions should minimize the importance of BSSE.[28,36] The Ross MP2 results summarized in Table 3 are probably the most reliable values currently available for these energies; they show a nonmonotonic change along the halide series with a dip in ΔE_{diss} at Cs_2Cl_2. The Ross HF values show the same trend and indeed vary from the MP2 results by only 0.1−0.3 kcal mol^{-1}. The HW ECP, however, gives results substantially different and also gives a monotonic series. The Ross B3LYP energies show a nonmonotonic trend but are several kilocalories per mole lower than the other results.

Structure determinations could also provide suitable tests for the two ECPs. Most crystal structures of cesium compounds show extended networks not suitable for modeling with computations of isolated single molecules. A few structures are available, however, that have well-isolated units. An example is the dimer, **1**, of cesium hexamethyldisilazide.[37] This compound crystallized with a molecule of toluene. The cesium salt units show an inversion center with two Cs−N bond lengths of 3.074 and 3.149 Å. The HF/Ross results are reasonably consistent with the Cs−N bond lengths being slightly longer, 3.145 and 3.254 Å, as expected from the cesium halide results above. The calculated N−Cs−N angle of 94.7° is also a little larger than the crystal structure's 90.7°. The bonds to cesium in the HW structure are substantially longer: Cs−N = 3.308 and 3.390 Å and N−Cs−N = 95.8°. The bonds not involving cesium are closer with both ECPs to those measured. The N−Si bond lengths are 1.673 and 1.679 Å (experimental), 1.683 and 1.684 Å (Ross), and 1.676 and 1.677 Å (HW). Similarly, the methyl groups form sets of six with Si−C = 1.872−1.890 Å (experimental), 1.904−1.917 Å (Ross), and 1.905−1.919 Å (HW). The agreement between the crystal structure and the Ross ECP results are clearly satisfactory and only a little less so for the HW ECP.

Unfortunately, this agreement does not extend to the structure of the tetramer, **2**, of cesium (trimethylsilyl)amide, $(CsNHSiMe_3)_4$.[38] The reported crystal structure is cubanelike but with a "remarkably short" Si−N bond length of 1.59 Å, as well as a rather short Cs−N bond length of 2.915 Å. The computational model was that of the tetramer of cesium silylamide, $(CsNHSiH_3)_4$, because of the size of the trimethylsilyl moiety. The computed structure with both ECPs is that of a distorted cube with two Cs−N bond lengths (Ross, 3.231 Å; HW, 3.345 Å) differing from the third (Ross, 3.167 Å; HW, 3.320 Å). Both sets of distances are comparable to those found with **1**. The N−Si bond length is also normal (Ross, 1.691 Å; HW, 1.686 Å) and also comparable to Si−N computed for **1**. The reported unusual structure for **2** finds no support in the computations with either ECP, and we suspect that there is some problem with the crystal structure.

Finally, we compare the HW and Ross protonation energies for a test group of cesium compounds including the Cs salts of both delocalized and localized carbanions as well as compounds with Cs−O, Cs−N, and Cs−S bonds; Cs−H; and some cesium halides. For the present, we compare the two sets of ECP results. Comparison with experimental relative acidities is beyond the scope of the present study and will be discussed in a separate paper. For convenience, the results are considered relative to fluorene in terms of the proton and cesium exchange in eq 2.

$$\text{FluorenylCs} + \text{RH} \xrightarrow{\Delta E} \text{Fluorene} + \text{RCs} \qquad (2)$$

The data are summarized in Table 4 and plotted in Figure 1. Since we are comparing electronic energy calculations with two ECPs, these data do not include the zero-point energy (ZPE). These energies are, in any event, quite similar to the

Table 4. Ross and HW ΔE Values (eq 2, kcal mol^{-1}) for Various RCs Compounds

RH	RCs	ΔE(Ross)	ΔE(HW)
	Delocalized C−Cs		
cyclopentadiene	CpCs	−15.073	−11.759
indene	IndCs	−7.473	−6.314
fluorene	FlCs	0.000	0.000
benzo[b]fluorene	Benzo[b]FlCs	3.362	2.856
benzo[def]fluorene	Benzo[def]FlCs	1.716	1.096
toluene	BnCs	24.338	25.101
p-methylbiphenyl	*p*-PhBnCs	23.421	23.312
diphenylmethane	Ph_2CHCs	16.224	15.850
triphenylmethane	Ph_3CCs	16.033	15.407
propylene	AllylCs	23.872	27.040
1,3-pentadiene	PentadienylCs	14.259	15.707
1,3-diphenylpropene	1,3-DiPhAllylCs	11.367	10.752
	Localized C−Cs		
methane	MeCs	44.351	48.009
ethane	EtCs	49.175	53.198
ethylene	VinylCs	34.187	38.085
benzene	PhCs	35.074	37.568
m-fluorobenzene	*m*-FPhCs	31.050	32.971
p-fluorobenzene	*p*-FPhCs	33.052	35.227
2,6-difluorobenzene	2,6-DiFPhCs	11.684	12.037
acetylene	EthynylCs	2.775	4.444
dimethyl sulfide	CH_3SCH_2Cs	28.388	30.618
dimethyl sulfoxide	CH_3SOCH_2Cs	3.881	5.825
dimethyl sulfone	$CH_3SO_2CH_2Cs$	−4.024	−4.235
$PhSCH_3$	PhSCH2Cs	25.278	26.823
$PhSOCH_3$	PhSOCH2Cs	3.811	5.270
$PhSO_2CH_3$	PhSO2CH2Cs	−2.367	−2.738
dithiane	DithianylCs	14.272	15.778
phenyldithiane	PhDithianylCs	10.351	10.126
methyldithiane	MeDithianylCs	18.693	20.267
	O−Cs		
water	CsOH	−1.845	4.582
methanol	CsOMe	−2.008	3.792
phenol	Cs phenoxide	−17.254	−15.341
ethenol	Cs vinyloxide	−16.700	−14.137
2-propenol	Cs α-MeVinyloxide	−16.135	−15.364
2-methyl-1-propenol	Cs $\beta\beta$-DiMeVinyloxide	−13.010	−10.123
1-phenylethenol	Cs 1-PhVinyloxide	−17.730	−16.092
α-tetralone enol	Cs tetralone-enolate	−16.068	−14.102
isobuyrophenone enol	CsIBP	−14.631	−12.889
	N−Cs		
ammonia	$CsNH_2$	20.666	26.704
CH_3NH_2	Cs MeAmide	23.598	29.707
Me_2NH	Cs DiMeAmide	23.468	29.036
$PhNH_2$	Cs anilide	5.656	7.404
Ph_2NH	Cs diphamide	−1.906	−2.016
carbazole	Cs carbazolide	−10.371	−11.126
	S−Cs		
H_2S	CsSH	−21.653	−20.820
CH_3SH	CsSMe	−15.428	−14.078
PhSH	CsSPh	−25.933	−25.713
	Miscellaneous		
H_2	CsH	20.339	26.560
HF	CsF	−22.623	−16.036
HCl	CsCl	−46.644	−44.898
HBr	CsBr	−55.608	−53.103

two ECPs, and the overall results are not dependent on this treatment. For carbonyl compounds, the comparison was made for the corresponding enols. In this way, all of the Cs−O results reflect the change RO−H → RO−Cs.

Despite the generally longer bond lengths with the HW ECP, the two calculation models give results that correlate quite closely. When all of the data in Figure 1 is used, the plot of ΔE(HW) versus ΔE(Ross) gives an intercept of 1.82 kcal mol^{-1} and a slope close to unity. An alternative approach is to focus on the cesium salts of delocalized carbanions since these are expected to have generally longer bonds to cesium and polarization of the cesium cation, a significant difference between the two ECPs, might be less important. This correlation is quite similar but with a smaller intercept of 0.54 kcal mol^{-1} and a slope just slightly less than unity. Both

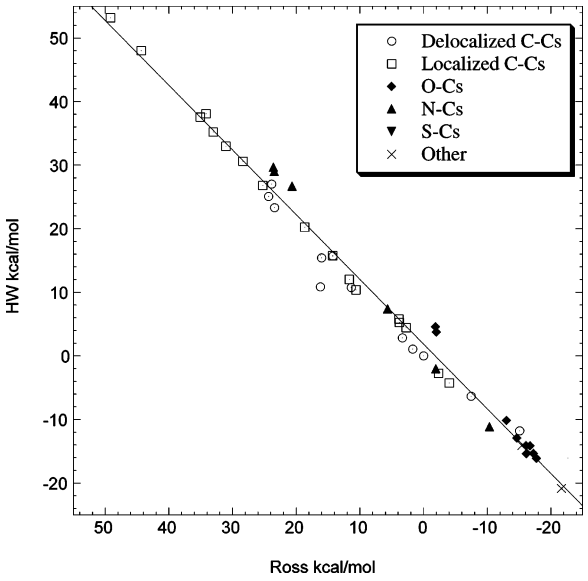


Figure 1. Plot of ΔE(HW) vs ΔE(Ross) for eq 2 (kcal mol^{-1}). The line shown is the regression line through all of the points: $\Delta E(\text{HW}) = 1.82 \pm 0.31 + (1.017 \pm 0.015)\Delta E(\text{Ross})$, $R^2 = 0.990$. A similar regression line for the delocalized C−Cs compounds gives $\Delta E(\text{HW}) = 0.54 \pm 0.82 + (0.960 \pm 0.054)\Delta E(\text{Ross})$, $R^2 = 0.970$.

correlations have high R^2 values. Although the general correlation is good, some compounds deviate substantially but in only a few systematic ways. The exception that the delocalized carbanion salts fall generally below the correlation given by the other more localized cesium salts does suggest that polarization functions on cesium are somewhat less important for those carbanions with effectively longer distances between positive and negative charge.

Conclusions

A computational model using the HW ECP without d functions on Cs gives bond lengths to cesium that are generally too long and corresponding bond strengths that are significantly too small. An alternative model based on the Ross ECP with d functions on Cs gives results that are generally more satisfactory. Nevertheless, for some qualitative work, the smaller HW basis set can give satisfactory results. For more quantitative work, however, the larger Ross ECP is to be preferred. B3LYP with the Ross ECP gives bond energies that are too low and should only be used after further calibration.

Acknowledgment. This work was supported in part by grants from the National Science Foundation and the Petroleum Research Fund of the American Chemical Society.

Supporting Information Available: Tables of basis sets, energies, and structural coordinates. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Streitwieser, A.; Ciula, J. C.; Krom, J. A.; Thiele, G. *J. Org. Chem.* **1991**, *56*, 1074−6.

(2) Facchetti, A.; Streitwieser, A. *J. Org. Chem.* **2004**, *69*, 8345−8355.

(3) Gareyev, R.; Ciula, J. C.; Streitwieser, A. *J. Org. Chem.* **1996**, *61*, 4589−4593.

(4) Krom, J. A.; Streitwieser, A. *J. Org. Chem.* **1996**, *61*, 6354−6359.

(5) Streitwieser, A.; Husemann, M.; Kim, Y. J. *J. Org. Chem.* **2003**, *68*, 7937−7942.

(6) Streitwieser, A.; Wang, D. Z.; Stratakis, M. *J. Org. Chem.* **1999**, *64*, 4860−4864.

(7) Wang, D. Z.; Streitwieser, A. *J. Org. Chem.* **2003**, *68*, 8936−8942.

(8) Wang, D. Z. R.; Kim, Y. J.; Streitwieser, A. *J. Am. Chem. Soc.* **2000**, *122*, 10754−10760.

(9) Streitwieser, A.; Keevil, T. A.; Taylor, D. R.; Dart, E. C. *J. Am. Chem. Soc.* **2005**, *127*, 9290−9297.

(10) For recent histories and reviews, see: Andrae, D. www.uni-bielefeld.de/chemie/tc/Andrae/bonn03.pdf (accessed Nov 1 2006) and Balasubramanian, K. *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Schreiner, P. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P., Schaefer, H. F., III, Eds.; Wiley: Chichester, U. K., 1998; Vol. 4 (Q−S), pp 2471−2480.

(11) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299−310.

(12) Ross, R. B.; Powers, J. M.; Atashroo, T.; Ermler, W. C.; LaJohn, L. A.; Christiansen, P. A. *J. Chem. Phys.* **1990**, *93*, 6654−6670.

(13) Hill, S. E.; Feller, D.; Glendening, E. D. *J. Phys. Chem. A* **1998**, *102*, 3813−3819.

(14) Hill, S. E.; Glendening, E. D.; Feller, D. *J. Phys. Chem. A* **1997**, *101*, 6125−6131.

(15) Ghatee, M. H.; Niroomand-Hosseini, F. *J. Phys. Chem. B* **2004**, *108*, 10034−10040.

(16) Li, Q. S.; Gong, L. F. *J. Phys. Chem. A* **2004**, *108*, 4322−4325.

(17) Cai, S. H.; Chen, Z.; Wan, H. L. *J. Phys. Chem. A* **2002**, *106*, 1060−1066.

(18) Schon, J. C.; Cancarevic, Z.; Jansen, M. *J. Chem. Phys.* **2004**, *121*, 2289−2304.

(19) Jeung, G. H. *J. Mol. Spectrosc.* **1997**, *182*, 113−117.

(20) Golebiowski, J.; Lamare, V.; Ruiz-Lopez, M. F. *J. Comput. Chem.* **2002**, *23*, 724−731.

(21) Stibbe, D. T.; Charron, E.; Brenner, V.; Millie, P.; Suzor-Weiner, A. *J. Chem. Phys.* **2002**, *116*, 10753−10759.

(22) Streitwieser, A.; Abu-Hasanyan, F.; Neuhaus, A.; Brown, F. *J. Org. Chem.* **1996**, *61*, 3151−3154 (Corr. p 5700).

(23) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, J., 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. G.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Pittsburgh, PA, 2003.

(24) Lambert, C.; Kaupp, M.; Schleyer, P. v. R. *Organometallics* **1993**, *12*, 853−859.

(25) Törring, T.; Biermann, S.; Hoeft, J.; Mawhorter, R.; Cave, R. J.; Szemenyei, C. *J. Chem. Phys.* **1996**, *104*, 8032−8042.

(26) Lim, I. S.; Schwerdtfeger, P.; Metz, B.; Stoll, H. *J. Chem. Phys.* **2005**, *122*, 104103 (1−12.).

(27) Dzidit, I.; Kebarle, P. *J. Phys. Chem.* **1970**, *74*, 1466−1474.

(28) Feller, D.; Glendening, E. D.; Woon, D. E.; Feyereisen, M. W. *J. Chem. Phys.* **1995**, *103*, 3526−3542.

(29) Glendening, E. D.; Feller, D. *J. Phys.Chem.* **1995**, *99*, 3060−3067.

(30) Kaupp, M.; Schleyer, P. v. R. *J. Phys. Chem.* **1992**, *96*, 7316−7323.

(31) For some references that also discuss earlier work, see: Törring, T.; Biermann, S.; Hoeft, J.; Mawhorter, R.; Cave, R. J.; Szemenyei, C. *J. Chem. Phys.* **1996**, *104*, 8032−8042 and Chauhan, R. S.; Sharma, S. C.; Sharma, S. B.; Sharma, B. S. *J. Chem. Phys.* **1991**, *95*, 4395−4406.

(32) *Energii Razryva Khimicheskikh Sujazei Potencialy Ionizacii i Srodstvo k Elekttronu*; Kondrat'ev, V. N., Ed.; Nauka: Moscow, 1974.

(33) Hilpert, K.; Miller, M.; Feltrin, A. *Schr. Forschungszent. Juelich, Reihe Energietech./Energy Technol.* **2000**, *15*, 463−466.

(34) Murgulescu, I. G.; Topor, L. *Rev. Roum. Chim.* **1968**, *13*, 1109−18.

(35) Gorokhoy, L. N. Dissertation, High Energy Institute, USSR, 1972; quoted by Kondrat'ev, ref 32.

(36) Galano, A.; Alvarez-Itaboy, J. R. *J. Comput. Chem.* **2006**, *27*, 1203−1210.

(37) Neander, S.; Behrens, U. *Z. Anorg. Allg. Chem.* **1999**, *625*, 1429−1434.

(38) Tesh, K. F.; Jones, B. D.; Hanusa, T. P.; Huffmad, J. C. *J. Am. Chem. Soc.* **1992**, *114*, 6590−6591.

(39) Eisenstadt, M.; Rothberg, G. M.; Kusch, P. *J. Chem. Phys.* **1958**, *29*, 797−804.

(40) Chao, J. *Thermochim. Acta* **1970**, *1*, 71−86.

(41) Milne, T. A.; Klein, H. M. *J. Chem. Phys.* **1960**, *31*, 1628−1637.

(42) Datz, S.; Smith, J. W. T.; Taylor, E. H. *J. Chem. Phys.* **1961**, *34*, 558−564.

CT600238V

JCTC Journal of Chemical Theory and Computation

Elucidation of Rate Variations for a Diels−Alder Reaction in Ionic Liquids from QM/MM Simulations

Orlando Acevedo,[†] William L. Jorgensen,*[,†] and Jeffrey D. Evanseck*[,‡]

Department of Chemistry, Yale University, 225 Prospect Street, New Haven, Connecticut 06520-8107, and Center for Computational Sciences and Department of Chemistry and Biochemistry, Duquesne University, Pittsburgh, Pennsylvania 15282-1530

Received August 25, 2006

Abstract: The impact of acidic and basic ionic liquid 1-ethyl-3-methylimidazolium chloride (EMIC) melts upon cyclopentadiene and methyl acrylate Diels−Alder reaction rates has been investigated using QM/MM calculations. The ability of the ionic liquid to act as a hydrogen bond donor (cation effect), moderated by its hydrogen bond accepting ability (anion effect), has been proposed previously to explain observed *endo/exo* ratios. However, the molecular factors that endow ionic liquids with their rate enhancing potential remain unknown. New OPLS-AA force field parameters in conjunction with potentials of mean force (PMF) derived from free energy perturbation calculations in Monte Carlo simulations (MC/FEP) are used to compute activation energies. QM/MM simulations using a periodic box of ions reproduce relative rate enhancements for the EMIC melts compared to water and 1-chlorobutane that reproduce kinetic experiments. Solute−solvent interactions in acidic and basic ionic liquid melts have been analyzed at key stationary points along the reaction coordinate. The reaction rate was found to be greater in the acidic rather than the basic melt due to less-dominant ion-pairing in the acidic melt, enabling the EMI cation to better coordinate to the dienophile at the transition state. The simulations suggest that the hydrogen on C2 of the EMI cation does not contribute to stabilization of the transition state, as previously believed, and the interactions with the more sterically exposed hydrogens on C4 and C5 play a larger role. In addition, the relative stabilization of the transition state through electrostatic interactions with the EMI cation in the acidic melt is also greater than that afforded by the weaker Lewis-acid effect provided by hydrogen bonding with water molecules in aqueous solution.

Introduction

Though the observed effects of ionic liquids on chemical reactions range from weak to powerful,[1] only a few systematic studies have addressed the microscopic details on how ionic liquids influence chemical reactivity and selectivity.[2−5] Large rate accelerations and stereoselectivity enhancements have been observed for the Diels−Alder reaction between cyclopentadiene and methyl acrylate in the ionic liquid 1-ethyl-3-methylimidazolium chloride (EMIC) (Scheme 1) when compared to common organic solvents or water.[6−8] The mechanism behind the ionic liquid's impact on the reaction is not well understood, although high internal solvent pressure,[9] hydrogen bonding,[3,5] solvent polarity,[10] and Lewis acidity[11] are thought to contribute to the phenomena. Rate enhancements in ionic liquids also differ depending on the counteranion, i.e., BF_4^-, PF_6^-, ClO_4^-, and, of particular interest, the chloroaluminates from $AlCl_3$.[6,7,12]

An attractive property of using $AlCl_3$ is that the Lewis acidity of the melt can be varied with the composition of the liquid.[13] Raman,[14] ^{27}Al NMR,[15] and mass spectra[16] all indicate that when $AlCl_3$ comprises <50% mol of the EMIC

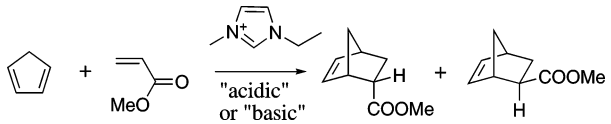
* Corresponding authors e-mail: evanseck@duq.edu (J.D.E.) and william.jorgensen@yale.edu (W.L.J).

[†] Yale University.

[‡] Duquesne University.

10.1021/ct6002753 CCC: $37.00
Published on Web 11/15/2006

Scheme 1. Diels−Alder Reaction between Cyclopentadiene and Methyl Acrylate Giving *Endo* and *Exo* Bicyclic Products in 1-Ethyl-3-methylimidazolium Chloride (EMIC) Solvent



ionic liquid melt, $AlCl_4^-$ is the only chloroaluminate species present. These "basic melts" are composed of EMI^+, $AlCl_4^-$, and with chloride ions that are not bound to aluminum. A ratio >1:1 $AlCl_3$ to EMIC is referred to as an "acidic melt"; ^{27}Al NMR[17] and negative-ion FAB mass spectra[18] have shown that $AlCl_4^-$ and $Al_2Cl_7^-$ are the principal anionic constituents in this case.

When an acidic melt (51% $AlCl_3$) of EMIC was used as the solvent for the Diels−Alder reaction in Scheme 1, the experimental rate of reaction was 10, 175, and 560 times faster than in water, ethyl ammonium nitrate, and 1-chlorobutane, respectively.[6] However, the basic melt (48% $AlCl_3$) of EMIC gave a rate 2.4 times slower than that of water. *Endo* selectivity is also enhanced with good yield when using an acidic EMIC ionic liquid. A 6.7:1 *endo*/*exo* ratio is observed for the solvent ethyl ammonium nitrate (EAN), the basic ionic liquid EMIC melt has a ratio of 5.25:1, while the acidic melt ionic liquid EMIC has a ratio of 19:1; typical nonpolar solvents have *endo/exo* selectivities of 2:1.[6,19]

Welton and co-workers have recently proposed that selectivity enhancements for the reaction between cyclopentadiene and methyl acrylate in a variety of ionic liquids result from a competition between the anion and the dienophile at the transition state for hydrogen bonding [The term hydrogen bonding is used throughout the ionic liquid literature and used in this study as well, despite the fact that the classical definition of a hydrogen bond is not formally followed.] with the cation.[3] Hence, the highest selectivities come from ionic liquids with the strongest hydrogen-bond donor capacity, along with the weakest hydrogen-bond acceptor ability.[3,5] The suggestion has been made that hydrogen bonding may also dictate the rate acceleration, since the increased rate is observed to be concurrent with increased *endo*-selectivity.[3] However, since the correlation is qualitative and does not consider other effects influencing the rate of reaction, a detailed study is necessary to elucidate the intermolecular interactions occurring at the transition state.

To investigate the dramatic kinetic effects of ionic liquids at the atomic level, mixed quantum and molecular mechanical (QM/MM) simulations have been carried out on the Diels−Alder reaction in solution with complete sampling of the geometry of the reacting systems and explicit representation of the solvent components. Reactants, transition structures, and products have been located using force fields in four solvents: acidic and basic chloroaluminate ionic liquids, water, and 1-chlorobutane. The present results indicate that the lowering of the activation barrier does feature enhanced interactions between EMI^+ and the transition state in the acidic melt, but the previously suggested hydrogen bonding with the hydrogen on C2 of EMI^+ is not apparent.[3,4] This investigation provides an understanding of the intermolecular short- and long-range interactions occurring during the reaction, and it sheds light on how to optimize ionic liquids to control other organic reactions.

Computational Methods

Mixed quantum and molecular mechanical (QM/MM) calculations,[20] as implemented in BOSS 4.6[21] were carried out with the reacting system treated with the PDDG/PM3 method.[22] PDDG/PM3 is a semiempirical QM method that has been extensively tested for gas-phase structures and energetics[22] and has given excellent results in solution-phase QM/MM studies of S_N2 reactions,[23] nucleophilic aromatic substitution reactions,[24] decarboxylations,[25,26] and Cope eliminations.[27] The solvent molecules were represented with the TIP4P water model,[28] the all-atom OPLS force field for 1-chlorobutane,[29] and newly developed OPLS parameters for the ionic liquids.[30−32] Binary solvent boxes containing 384 molecules were constructed for model acidic (192 EMI^+ and 192 $Al_2Cl_7^-$ ions) and basic melts (192 EMI^+ and 192 $AlCl_4^-$ ions) and equilibrated for 80M and 65M configurations, respectively. Periodic boundary conditions were used, and the dimensions of the simulation cells were ca. 42 × 42 × 62 Å and 38 × 38 × 56 Å for the acidic and basic melts, respectively. Binary solvent media were also used in the technically similar study for decarboxylation of a biotin model.[26] Partial charges for the ionic liquid cation and anions were obtained from B3LYP/6-31G(d) calculations with the CHELPG method.[33] The EMI cation had full bond and angle flexibility, and torsion angle sampling was included for the ethyl side chain. The anions were kept internally rigid throughout the simulations. Solute−solvent and solvent−solvent intermolecular cutoff distances of 12 Å were employed for the reaction in water based roughly on center-of-mass separations. A feathering of the intermolecular interactions within 0.5 Å of the cutoff is applied. Long-range electrostatics were treated with the Ewald method for the ionic liquids and 1-chlorobutane. In short, Ewald summations calculate the exact electrostatic energy of an infinite lattice of identical copies of the simulation cell. This suppresses artifacts resulting from the simple cutoff of the long-range electrostatic interactions prevalent in the ionic liquid. Total translations and rotations were sampled in ranges that led to overall acceptance rates of about 45% for new configurations.

Free energy perturbation (FEP) calculations were performed in conjunction with Metropolis Monte Carlo (MC) simulations in the NPT ensemble at 25 °C and 1 atm. In this QM/MM implementation, the solutes are treated quantum mechanically using PDDG/PM3; computation of the QM energy and atomic charges was performed for every attempted solute move, occurring every 100 configurations. Electrostatic contributions to the solute−solvent energy were calculated using CM3 charges,[34] with a scale factor of 1.08. The initial geometry of the reacting system was obtained from density functional theory calculations, and the *endo*-cis addition mode was chosen in all cases. FEP windows were run simultaneously using multiple processors on Pentium-based clusters located at Yale University and the Center for Computational Sciences (CCS) at Duquesne University.

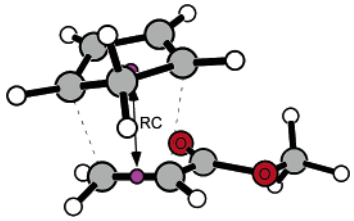


Figure 1. Reaction coordinate, RC, for the Diels−Alder reaction is defined as the distance between the midpoints of the C═C bond of the dienophile and the diene's terminal carbons. The transition structure from gas-phase B3LYP/6-31G(d) calculations is illustrated.

Results and Discussion

Force Field for the Ionic Liquids. To elucidate the effect of the room-temperature ionic liquid EMIC on the Diels−Alder reaction, a custom force field has been parametrized to model EMIC acidic and basic melts. Recent theoretical work has been reported on force the field development for several ionic liquids.[30,35] Parameters specific to the current EMIC-$AlCl_3$ ionic liquid have been created starting from a force field developed and validated by Lopes et al.[30] from existing OPLS-AA parameters designed for heterocyclic aromatic rings.[31] Lennard-Jones parameters for Cl atoms have been fit to the Born-Mayer potential used to describe dispersive interactions in molten salts.[30] Improvements to the Lopes et al. force field specific to EMIC-$AlCl_3$ have been achieved by using new equilibrium bond lengths and angles based on the B3LYP/6-31G(d) geometries, new partial charges computed with the CHELPG method and B3LYP/6-31G(d), and new torsion potentials for the side-chain rotation of the ethyl group in the EMI cation (see the Supporting Information). Computed densities of 1.22 and 1.20 g/cm^3 for the acidic and basic melt, respectively, agree well with the experimental value of 1.266 g/cm^3 at 60 °C for the 1:1 melt.[36] Further validation of the force field including radial pair distributions, g(r), can be found in references by Lopes et al.[30] and de Andrade et al.[35b] which used similar setups and parameters. Complete force field parameters are reported in the Supporting Information.

Table 1. Free Energy Changes (kcal/mol) at 25 °C for the Diels−Alder Reaction between Cyclopentadiene and Methyl Acrylate from QM/MM/MC Simulations

	water	acidic melt	basic melt	BuCl
$\Delta G^{\ddagger}$ (calc)[a]	42.9	40.0	43.5	43.8
$\Delta\Delta G^{\ddagger}$ (calc)	0.0	−2.9	0.6	0.9
$\Delta\Delta G^{\ddagger}$ (exptl)[b]	0.0	−1.36	0.52	2.39

[a] Statistical uncertainties (1σ) are ±0.4 kcal/mol. [b] Reference 6.

Energetics and Structure. Free-energy profiles for the Diels−Alder reaction have been calculated by perturbing a reaction coordinate (RC) defined by the distance between two dummy atoms located at the midpoint of the diene terminal carbon atoms and the midpoint of the dienophile C═C bond, as shown in Figure 1. The range of RC spanned is from 1.42 to 4.20 Å with an increment of 0.02 Å. Each FEP calculation entailed between 5 and 35 million (M) configurations of equilibration followed by 10 M configurations of averaging. The number of single-point QM calculations required for one free energy profile is 20−80 million, clearly showing the need for highly efficient QM methods in such studies. The computed free energy profiles for the reaction in water and in acidic and basic EMIC are shown in Figure 2.

The cycloadduct minimum is found when RC is close to 1.5 Å, while the transition state occurs around 2.12 Å in all solvent systems. The Diels−Alder mechanism was found to be concerted in chloroaluminate ionic liquids, which agrees with previous density functional theory calculations.[4] The free energies are relatively constant for RC beyond 4 Å. The differential solvent effects can be seen primarily near the transition state. The principal point is that the QM/MM/MC approach qualitatively reproduces the observed rate acceleration by acidic ionic liquids and the rate retardation by basic ionic liquids for the Diels−Alder reaction in comparison to aqueous solution. The computed and experimental $\Delta\Delta G^{\ddagger}$ values for the acidic melt are −2.9 and −1.4 kcal/mol, while

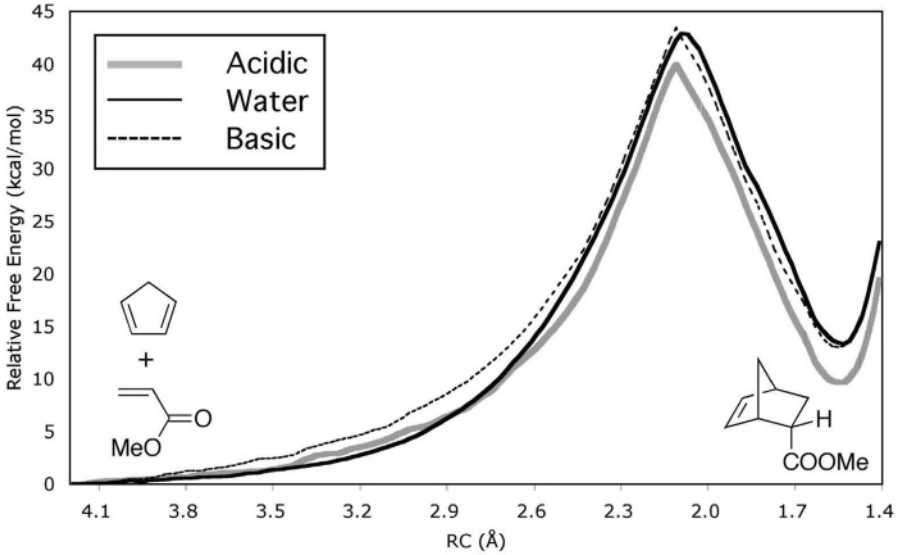


Figure 2. Computed free energy profiles in acidic and basic ionic liquids and water for the cycloaddition between cyclopentadiene and methyl acrylate.

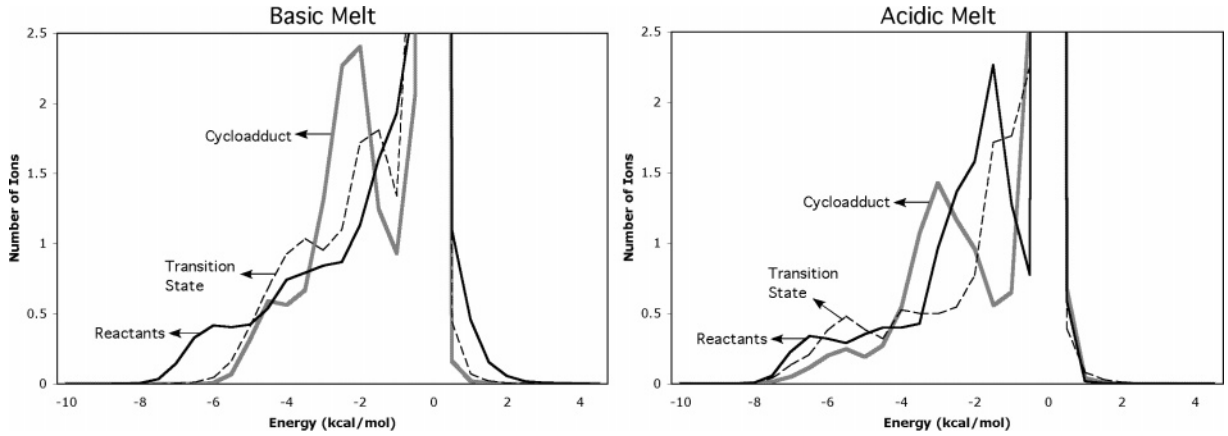


Figure 3. Solute−solvent energy pair distributions for the reaction of cyclopentadiene with methyl acrylate for structures near the reactants, transition state, and product. The ordinate records the number of ions in the melts that interact with the solutes with their interaction energy on the abscissa. Units for the ordinate are number of ions per kcal/mol.

they are 0.6 and 0.52 kcal/mol for the basic melt (Table 1). Since the same computational approach is used for all solvents, most errors are expected to cancel. The computed $\Delta\Delta G^{\ddagger}$ values reproduce well the rate enhancement differences between the basic and acidic melts. Exact agreement could not be expected since the ionic composition of the experimental and modeled systems is not identical.

Uncertainties in the free energy barriers have been calculated by propagating the standard deviation (σ_i) on each individual ΔG_i.[21] Smooth free energy profiles were obtained with statistical uncertainties ($1\sigma_i$) of only 0.005−0.06 kcal/mol in each window; the overall uncertainties have been computed using the equation $\sqrt{\sum_i^N \sigma_i^2}$, where N is the number of ΔG_i values. The computed errors in the free energies of activation, $\Delta G^{\ddagger}$, for water and both ionic liquid melts are computed to be ± 0.4 kcal/mol. For the $\Delta\Delta G^{\ddagger}$ calculations, the propagation of errors through variances is not explicit and cannot be used to compute uncertainties in the relative barriers.

The free energy profile for the Diels−Alder reaction was subsequently also computed in 1-chlorobutane. Before the reaction was modeled, pure liquid properties were also checked through MC simulations. A periodic box of 384 fully flexible 1-chlorobutane molecules was created and equilibrated for 20 M and then averaged for 10 M configurations. The computed heat of vaporization of 7.28 kcal/mol (exptl = 7.26 kcal/mol)[37] and density of 0.83 g/cm^3 (exptl = 0.886 g/cm^3)[38] are both in good accord with experiment. The $\Delta G^{\ddagger}$ of 1-chlorobutane was then computed to be 43.8 kcal/mol (Table 1). The increase relative to the acidic melt is 3.8 kcal/mol, which is in excellent agreement with the experimental result of 3.75 kcal/mol.[6] A plot of the free energy profile in 1-chlorobutane can be found in the Supporting Information.

Although the relative free energies of activation are in good agreement with experiment, the absolute values are overestimated. The experimental $\Delta G^{\ddagger}$ reported for the Diels−Alder reaction in a 50/50 methanol/water mixture was 22.4 kcal/mol.[39] The $\Delta G^{\ddagger}$ values for the ionic liquids should be reasonably close to the water mixture (±2 kcal/mol), so the 40−43 kcal/mol values from PDDG/PM3 are too high. This has been noted for other Diels−Alder reactions,[22] and in previous QM/MM/MC work on aqueous-phase Diels−Alder reactions involving cyclopentadiene and three different dienophiles (acyrlonitrile, methyl vinyl ketone, and 1,4-naphthoquinone), the AM1 semiempirical method also overestimated the absolute activation energies, while leading to good relative values.[40] As noted, computation of each QM/MM/MC free energy profile entailed ca. 20−80 million QM calculations, which make the use of far more computationally taxing ab initio or DFT methods prohibitive from a practical standpoint.

Solvation. The close agreement between the computed and observed changes in free energies of activations (Table 1) indicates that the QM/MM/MC simulations capture the origins of the medium effects at the molecular level. Differential stabilization of the transition structure through hydrogen bonding may be the primary factor affecting the rate, as found in prior simulations of Diels−Alder reactions in water.[40,41] In order to elucidate the differences between the acidic and basic melts, the interaction energies for both were quantified by analyzing the solute−solvent energy pair distributions from QM/MM/MC calculations in three representative FEP windows: near the reactants, transition state, and product. The distributions record the average number of ions in the ionic liquids that interact with the reacting system and their corresponding energies. Very favorable interaction energies between solute and solvent components are reflected in the left-most region with energies more attractive than ca. −5 kcal/mol (Figure 3). The large band near 0 kcal/mol arises from the many ions in outer shells.

In viewing Figure 3, the reactants appear to have similar distributions for both acidic and basic melts leading to comparable ground-state stabilization in both ionic liquids. Integration of the distributions for the reactants from −10.0 to −5.0 kcal/mol confirms the similar number of very favorable interactions (Table 2). However, the distribution for the transition states is clearly different with a larger number of very favorable interactions in the acidic melt (Table 2). Specifically, the number of interactions in the −10 to −5 kcal/mol range decreases by about one in going from the reactants to transition state in the basic melt; however, in the acidic melt, the same number of interactions is retained.

Table 2. Number of Solute−Solvent Interactions for the Reactants, Transition State, and Product Integrating from −10.0 to −5.0 kcal/mol from QM/MM/MC Simulations

	reactants	transition state	cycloadduct
basic	1.8	0.6	0.4
acidic	1.6	1.6	0.8

In addition, there is a shift in the average strength of the most favorable interactions to lower energy by about 2 kcal/mol (Figure 3). The pattern is fully consistent with the faster reaction rate in the acidic melt.

The exact nature of these most favorable solute-ion interactions is of obviously relevant interest. Figure 4 shows snapshots of the transition structures in the acidic and basic melts from the QM/MM/MC simulations. Immediately a major difference in the structure compared to that expected from chemical intuition is observed. For both the acidic and basic melts, hydrogen bonds are not observed between HC2 of the EMI cation and the dienophile's carbonyl group. In the basic melt, the HC2 unit is shielded by the adjacent methyl and ethyl groups and also represents a small volume element. The anions congregate around the diene forming electrostatically favorable interactions with hydrogens, while two EMI cations complex near the carbonyl oxygen of the dienophile. This partitioning of the ions is consistent with the direction of the dipole for the transition state. In the acidic melt, an EMI ring hydrogen makes a short contact with the carbonyl oxygen, while only the methyl and ethyl groups of the two cations are the most proximal in the basic melt. As noted previously,[40−43] hydrogen bonding is sensitive to small charge shifts. In the present case, charges were computed from the final transition state configuration sampled (Figure 4) which is a good representation of the average structure.[40] On the carbonyl oxygen of the methyl acrylate reactant, the charge only declines from −0.288 and −0.291 e in the acidic and basic melt, respectively, to −0.322 and −0.313 e in the transition state (Table 3). However, the charge computed on the carbonyl carbon in the acidic melt *increases* from +0.475 to +0.501 e from reactant to transition structure, while *decreasing* from +0.491 to 0.466 e in the basic melt (Table 3). Thus, the Diels−Alder transition state in the acidic melt has the greatest C^{+}—O^{-} polarization. This is consistent with enhanced hydrogen bonding in the acidic melt and a reduced activation barrier as compared to the basic melt.

Table 3. CM3 Charges for the Carbonyl Oxygen and Carbon on Methyl Acrylate in the Basic and Acidic Melts for the Diels−Alder Reactants, Transition Structure, and Cycloadduct Using PDDG/PM3

	carbonyl O	carbonyl C
	Basic	
reactants	−0.291	0.491
transition state	−0.313	0.466
cycloadduct	−0.339	0.472
	Acidic	
reactants	−0.288	0.475
transition state	−0.322	0.501
cycloadduct	−0.342	0.473

Previous work has shown that the presence of hydrogen-bond donors in the reaction medium leads to enhanced *endo:exo* selectivities Diels−Alder reactions.[3,5,7] This finding has Aggarwal et al. to suggest that the most likely complex determining the selectivity in ionic liquids involves the most acidic proton (HC2) of the EMI cation interacting with the carbonyl oxygen of the dienophile.[3] However, higher selectivities have been observed in 1-alkyl,2,3-dimethylimidazolium ionic liquids, where the 2-position hydrogen has been replaced with a methyl group.[5] The results suggest that the hydrogen-bond donor properties for HC2 of the cation do not exclusively account for the strong *endo* selectivities observed and that other interactions, e.g., with hydrogens on C4 and C5, may play a larger role. The latter hydrogens

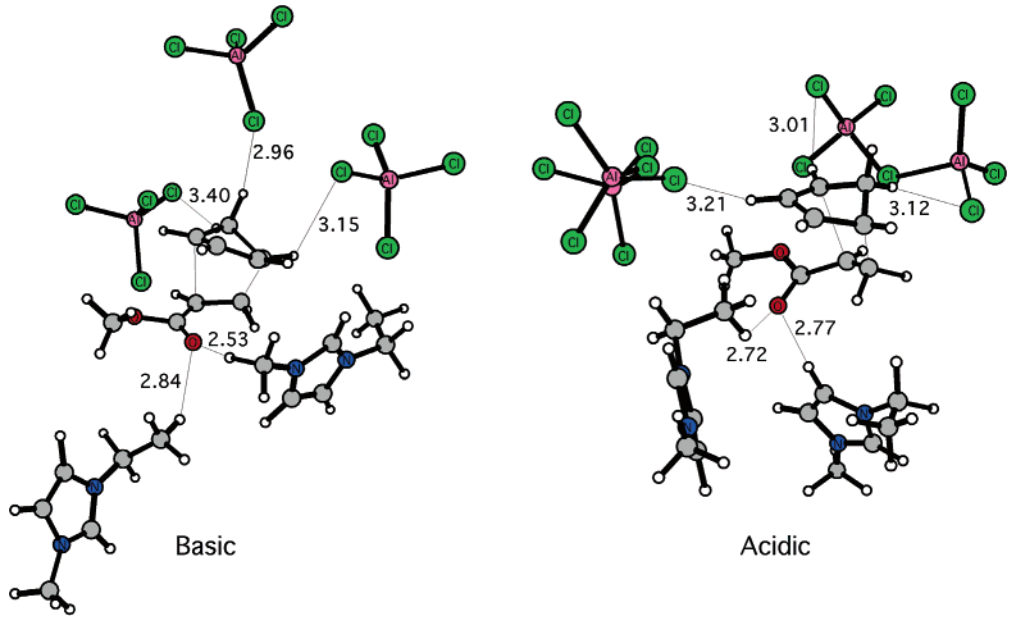


Figure 4. Typical snapshots of transition structures for the Diels−Alder reaction in the basic and acidic melts (only the nearest ions are illustrated). The distances (in Å) are average values over the final 10 million configurations of QM/MM/MC simulations.

are more sterically accessible, and their partial charges are essentially the same as for HC2 (see the Supporting Information). The present computations indicate that the HC2 fragment of the EMI cation is, in fact, not involved in hydrogen bonding with the dienophile at the transition state. The ions in the ionic liquids form ion-pairs with large binding energies, as shown by the DFT calculations,[4] and the interactions weaken gradually as the melt becomes more acidic. Consistently, the acidic melt exhibits enhanced hydrogen bonding with the transition structure, since the energetics of ion-pairing are less dominant; the more acidic hydrogens in the EMI ring are then less engaged in the ion pairing. The basic melt, featuring stronger ion-pairing interactions, prefers to have the ring hydrogens interacting with nearby $AlCl_4^-$ anions allowing only weaker solvation of the transition state through interactions with the methyl and ethyl groups of EMI^+.

In summary, the qualitative picture that emerges is that the reaction rate is greater in the acidic rather than the basic melt because ion-pairing is less dominant in the acidic melt, which enables the EMI cations to better coordinate with the dienophile at the transition state; the coordination does feature hydrogen bonding with the dienophile's carbonyl oxygen, but it is with the more sterically exposed hydrogens on C4 and C5 of the cation rather than the hydrogen on C2. The relative stabilization of the transition state through hydrogen bonding with the EMI cation in the acidic melt is also greater than that afforded by the weaker Lewis-acid effect afforded by hydrogen bonding with water molecules in aqueous solution.[40−42,44]

Conclusions

QM/MM/MC simulations have been carried out for the Diels−Alder reaction in ionic liquids, water, and 1-chlorobutane in good agreement with kinetic experiments. Analysis of the explicit solute−solvent interactions at the transition state indicates that an enhanced coordination of hydrogen bonds at the carbonyl oxygen is largely responsible for the reaction rate being greater in the acidic rather than the basic melt. The rate enhancements in the acidic melt arise from preferential hydrogen bonding with the more sterically exposed C4 and C5 hydrogens on the EMI cation rather than the hydrogen on C2. Strong ion-pairing in the basic melt prevents the ring hydrogens from participating at the transition state. In aqueous solution, a weaker Lewis-acid effect is provided by hydrogen bonding with water molecules compared to the EMI cations in the acidic melt. The findings have a significant impact for predicting the effect of ionic liquids on other organic reactions; improvements in rate are expected to be limited in the absence of differential hydrogen bond stabilization of the transition state.

Acknowledgment. J.D.E. is grateful to the NSF (CHE-0321147, CHE-0354052, AAB/PSC CHE-030008P), Department of Education (P116Z040100 and P116Z050331), and SGI Corporation for the support of this work. Support at Yale was provided by the NSF CHE-0130996.

Supporting Information Available: OPLS-AA force field parameters for EMIC-$AlCl_3$, free energy profile in 1-chlorobutane, DFT energies, zero-point energies, imaginary frequencies, and Cartesian coordinates of all structures. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) (a) Seddon, K. R. *J. Chem. Technol. Biotechnol.* **1997**, *68*, 351−356. (b) Welton, T. *Chem. Rev.* **1999**, *99*, 2071−2083. (c) Wasserscheid, P.; Keim, W. *Angew. Chem. Int. Ed.* **2000**, *39*, 3772−3789. (d) Jain, N.; Kumar, A.; Chauhan, S.; Chauhan, S. M. S. *Tetrahedron* **2005**, *61*, 1015−1060.

(2) (a) Bock, C. W.; Trachtman, M.; Mains, G. J. *J. Phys. Chem.* **1994**, *98*, 478−485. (b) Hébant, P.; Picard, G. *J. Mol. Struct. (Theochem)* **1995**, *358*, 39−50. (c) Picard, G.; Bouyer, F. C.; Leroy, M.; Bertaud, Y.; Bouvet, S. *J. Mol. Struct. (Theochem)* **1996**, *368*, 67−80. (d) Takahashi, S.; Suzuya, K.; Kohara, S.; Koura, N.; Curtiss, L. A.; Saboungi, M. L. *Z. Phys. Chem.* **1999**, 209−222.

(3) Aggarwal, A.; Lancaster, N. L.; Sethi, A. R.; Welton, T. *Green Chem.* **2002**, *4*, 517−520.

(4) Acevedo, O.; Evanseck, J. D. In *Ionic Liquids as Green Solvents: Progress and Prospects*; ACS Symposium Series, 2003; Vol. 856, pp 174−190.

(5) Vidis, A.; Ohlin, C. A.; Laurenczy, G.; Küsters, E.; Sedelmeier, G.; Dyson, P. J. *Adv. Synth. Catal.* **2005**, *347*, 266−274.

(6) Lee, C. W. *Tetrahedron Lett.* **1999**, *40*, 2461−2464.

(7) Fischer, T.; Sethi, A.; Welton, T.; Woolf, J. *Tetrahedron Lett.* **1999**, *40*, 793−796.

(8) (a) Kumar, A.; Pawar, S. S. *J. Org. Chem.* **2004**, *69*, 1419−1420. (b) Kumar, A.; Pawar, S. S. *J. Mol. Catal. A: Chem.* **2004**, *208*, 33−37. (c) Tiwari, S.; Kumar, A. *Angew. Chem. Int. Ed.* **2006**, *45*, 4824−4825.

(9) Eldik, R. V.; Asano, T.; Le, Noble, W. *Chem. Rev.* **1989**, *89*, 549−688.

(10) Berson, J. A.; Hamlet, Z.; Mueller, W. A. *J. Am. Chem. Soc.* **1962**, *84*, 297−304.

(11) Otto, S.; Blokzijl, W.; Engberts, J. B. F. N. *J. Org. Chem.* **1994**, *59*, 5372−5376.

(12) Earle, M. J.; McCormac, P. B.; Seddon, K. R. *Green Chem.* **1999**, *1*, 23−25.

(13) Hussey, C. L. *Pure Appl. Chem.* **1988**, *60*, 1763−1772.

(14) Gale, R. J.; Gilbert, B. P.; Osteryoung, R. A. *Inorg. Chem.* **1978**, *17*, 2728−2729.

(15) Wilkes, J. S.; Frye, J. S.; Reynolds, G. F. *Inorg. Chem.* **1983**, *22*, 3870−3872.

(16) (a) Ackermann, B. L.; Tsarbopoulos, A.; Allison, J. *Anal. Chem.* **1985**, *57*, 1766−1768. (b) Wicelinski, S. P.; Gale, R. J.; Pamidimukkala, K. M.; Laine, R. A. *Anal. Chem.* **1988**, *60*, 2228−2232.

(17) Gray, J. L.; Maciel, G. E. *J. Am. Chem. Soc.* **1981**, *103*, 7147−7151.

(18) Franzen, G.; Gilbert, B. P.; Pelzer, G.; Depauw, E. *Org. Mass Spectrom.* **1986**, *21*, 443−444.

(19) Jaeger, D. A.; Su, D. *Tetrahedron Lett.* **1999**, *40*, 257−260.

(20) (a) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227−249. (b) Kaminski, G. A.; Jorgensen, W. L. *J. Phys. Chem. B* **1998**, *102*, 1787−1796.

(21) Jorgensen, W. L.; Tirado-Rives, J. *J. Comput. Chem.* **2005**, *26*, 1689−1700.

(22) (a) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601−1622. (b) Tubert-Brohman, I.; Guimarães, C. R. W.; Repasky, M. P.; Jorgensen, W. L. *J. Comput. Chem.* **2003**, *25*, 138−150. (c) Tubert-Brohman, I.; Guimarães, C. R. W.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2005**, *1*, 817−823.

(23) Vayner, G.; Houk, K. N.; Jorgensen, W. L.; Brauman, J. I. *J. Am. Chem. Soc.* **2004**, *126*, 9054−9058.

(24) Acevedo, O.; Jorgensen, W. L. *Org. Lett.* **2004**, *6*, 2881−2884.

(25) Acevedo, O.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2005**, *127*, 8829−8834.

(26) Acevedo, O.; Jorgensen, W. L. *J. Org. Chem.* **2006**, *71*, 4896−4902.

(27) Acevedo, O.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2006**, *128*, 6141−6146.

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

(29) Jorgensen, W. L.; Ulmschneider, J. P.; Tirado-Rives, J. *J. Phys. Chem. B* **2004**, *108*, 16264−16270.

(30) Canongia, Lopes, J. N.; Deschamps, J.; Padua, A. H. *J. Phys. Chem. B* **2004**, *108*, 2038−2047.

(31) McDonald, N. A.; Jorgensen, W. L. *J. Phys. Chem. B* **1998**, *102*, 8049−8059.

(32) Acevedo, O. Ph.D. Thesis, Duquesne University, Pittsburgh, PA, 2003.

(33) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361−373.

(34) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **2003**, *24*, 1291−1304.

(35) (a) de Andrade, J.; Böes, E. S.; Stassen, H. *J. Phys. Chem. B* **2002**, *106*, 3546−3548. (b) de Andrade, J.; Böes, E. S.; Stassen, H. *J. Phys. Chem. B* **2002**, *106*, 13344−13351. (c) Margulis, C. J.; Stern, H. A.; Berne, B. J. *J. Phys. Chem. B* **2002**, *106*, 12017−12021. (d) Shah, J. K.; Brennecke, J. F.; Maginn, E. J. *Green Chem.* **2002**, *4*, 112−118. (e) Canongia Lopes, J. N.; Padua, A. H. *J. Phys. Chem. B* **2004**, *108*, 16893−16898. (f) Liu, Z.; Huang, S.; Wang, W. *J. Phys. Chem. B* **2004**, *108*, 12978−12989.

(36) Fannin, A. A., Jr.; Floreani, D. A.; King, L. A.; Landers, J. S.; Piersma, B. J.; Stech, D. J.; Vaughn, R. L.; Wilkes, J. S.; Williams, J. L. *J. Phys. Chem.* **1984**, *88*, 2614−2621.

(37) Shehatta, I. *Thermochim. Acta* **1993**, *213*, 1−10.

(38) *CRC Handbook of Chemistry and Physics*, 52nd ed.

(39) Ruiz-López, M. F.; Assfeld, X.; García, J. I.; Mayoral, J. A.; Salvatella, L. *J. Am. Chem. Soc.* **1993**, *115*, 8780−8787.

(40) Chandrasekhar, J.; Shariffskul, S.; Jorgensen, W. L. *J. Phys. Chem. B* **2002**, *106*, 8078−8085.

(41) Blake, J. F.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1991**, *113*, 7430−7432.

(42) Blake, J. F.; Lim, D.; Jorgensen, W. L. *J. Org. Chem.* **1994**, *59*, 803−805.

(43) Jorgensen, W. L.; Blake, J. F.; Lim, D.; Severance, D. L. *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 1727−1732.

(44) Kong, S.; Evanseck, J. D. *J. Am. Chem. Soc.* **2000**, *122*, 10418−10427.

CT6002753

JCTC Journal of Chemical Theory and Computation

Density Functional Theory Study of the Formation of Naphthalene and Phenanthrene from Reactions of Phenyl with Vinyl- and Phenylacetylene

Jorge Aguilera-Iparraguirre and Wim Klopper*

Lehrstuhl für Theoretische Chemie, Institut für Physikalische Chemie, Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany

Received August 7, 2006

Abstract: Reaction pathways for polycyclic aromatic hydrocarbon growth from reactions of either vinyl- or phenylacetylene with a phenyl radical are proposed and investigated using density functional theory (DFT). B3LYP/TZVP calculations are performed to obtain structures of minima and saddle points as well as kinetic data, supplemented with BMK/TZVP single-point energy calculations. The pathways include a cis−trans isomerization via a radicalic four-membered ring intermediate, which has so far not been considered in the literature. The DFT approach is validated against coupled-cluster calculations of a model system representing this intermediate. The coupled-cluster calculations include single and double excitations as well as perturbative corrections for connected triples and are performed in a correlation-consistent cc-pVTZ basis.

I. Introduction

The chemical vapor deposition (CVD) of carbon from light hydrocarbons involves a large and complex set of homogeneous gas-phase reactions leading to various products including polycyclic aromatic hydrocarbons (PAHs) and heterogeneous surface reactions leading to the deposition of pyrolytic carbon on the substrate.[1−3] Recently, Norinaga et al. studied the product distributions in the CVD of carbon from acetylene at 8 kPa and 900 °C and found that the major products formed under these conditions are benzene, dihydrogen, vinylacetylene, naphthalene, ethylene, and methane.[4] Benzene and vinylacetylene are known to be formed in acetylene pyrolysis, and in ref 4, it was suggested that vinylacetylene is formed by the dimerization of acetylene while benzene is formed by the combination of acetylene and vinylacetylene. Possible reactions for benzene formation from small aliphatics are

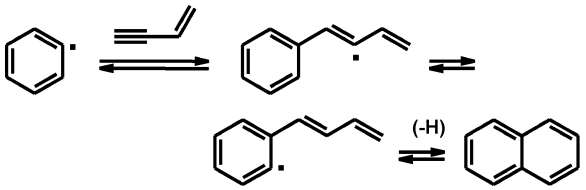
$$n\text{-}C_4H_3 + C_2H_2 \rightarrow \text{phenyl} \quad (1)$$

$$n\text{-}C_4H_5 + C_2H_2 \rightarrow \text{benzene} + H \quad (2)$$

$$C_3H_3 + C_3H_3 \rightarrow \text{benzene} \quad (3)$$

Although still under debate (cf. refs 5−7 for a review of the literature), we shall in this article not be concerned with the formation of benzene but rather with the formation of larger PAHs such as naphthalene and phenanthrene.

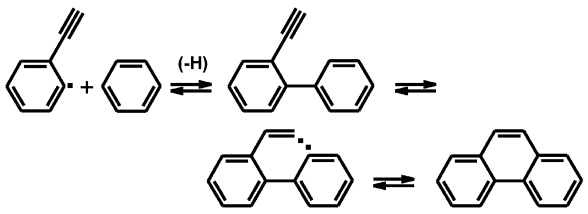
The addition of vinylacetylene to benzene (or more precisely phenyl) is a possible route to naphthalene formation:[8]



(4)

While Appel and co-workers have reported that, in their kinetic modeling, the vinylacetylene addition channel (eq 4) contributed the most to the naphthalene production,[8] Frenklach and others have argued that eq 4 may be ineffective because of the large barrier to rotation about the C=C double bond in the last step.[5,9,10]

Appel and co-workers have furthermore suggested that phenanthrene is formed primarily via ring−ring condensation reactions, such as



(5)

* Corresponding author fax: +49-721-608-3319; e-mail: klopper@chem-bio.uni-karlsruhe.de.

10.1021/ct600255u CCC: $37.00
Published on Web 12/08/2006

Table 1. Parameters Used in the Kinetic Modeling in ref 13

reaction	A (cm^3 mol^{-1} s^{-1})	n	E_a (kJ mol^{-1})	ref
phenyl + vinylacetylene → naphthalene + H	9.900×10^{30}	−5.07	88.29	8
phenyl + phenylacetylene → phenanthrene + H	9.550×10^{11}	0.0	18.03	6

Alternatively, the growth of phenanthrene could be initiated by the formation of biphenyl from benzene, then yielding phenanthrene via the HACA mechanism (hydrogen-abstraction−C_2H_2-addition) of Frenklach and Wang.[5,11,12]

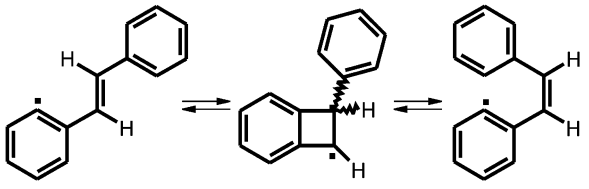
In their kinetic modeling of acetylene pyrolysis (900 °C, 8 kPa, and 0.5 s), Norinaga and Deutschmann[13,14] found that 68% of acetylene is consumed by dimerization to form vinylacetylene ($2C_2H_2 \rightarrow C_4H_4$), 17% by diacetylene formation ($2C_2H_2 \rightarrow C_4H_2 + H_2$), and 7% by benzene formation from vinylacetylene and acetylene ($C_4H_4 + C_2H_2 \rightarrow$ benzene). From the viewpoint of the products, 98% of vinylacetylene is formed by the dimerization of acetylene, and 95% of benzene is formed by the combination of vinylacetylene and acetylene.

In ref 13, modeling was performed with a mechanism involving 227 species and 827 reactions. In the present work, we shall concentrate on two of these reactions that seem to play an important role in the formation of larger PAHs. In ref 4, for example, yields (based on C_1) of ca. 2% of naphthalene and ca. 1% of phenanthrene were observed in acetylene pyrolysis after about 1.0 s under the aforementioned conditions.

The kinetic modeling of ref 13 was based on Arrhenius-like rate expressions:

$$k(T) = AT^n \exp(-E_a/RT) \tag{6}$$

with the kinetic parameters (pre-exponential A, temperature exponent n, activation energy E_a) taken from the literature (R is the gas constant and T is the temperature). Because the modeling of the formation of naphthalene and phenanthrene appeared to be difficult, we decided to reinvestigate—by means of density functional theory (DFT)—the two sets of parameters shown in Table 1. We shall show in this article that very different parameters can be derived from a reaction mechanism that involves a cis−trans isomerization via a radicalic four-membered ring, which has so far not been considered in the literature:



(7)

The present study describes the DFT investigations of the reactions of phenyl with vinylacetylene and phenylacetylene via this four-membered intermediate.

II. Computational Section

Geometry optimizations utilizing analytic nuclear gradients were carried out to locate minima and saddle points using DFT methods as implemented in the Turbomole program package.[15−18] At the initial stage of our work, we used the SV(P) basis[19,20] (split-valence basis with a set of d-type polarization functions on C) and the BP86 functional[21−23] in combination with the RI-J approximation[24,25] to explore possible pathways at a very low computational cost. The RI-J approximation reduces the computing time to about 10% of the time needed for the corresponding calculation without this approximation.

Subsequently, the final results were obtained by locating the minima and saddle points using the TZVP basis[20,26] (triple-ζ valence plus polarization) and the B3LYP functional.[22,27,28] Moreover, single-point calculations were carried out in the same TZVP basis using the BMK functional,[29] which has been designed specifically for accurate calculations of barrier heights. The RI-J approximation was not invoked in the B3LYP and BMK calculations. Redundant internal coordinates were used in the geometry optimizations,[30] and the optimization of saddle points was performed using the TRIM method (trust radius image minimization).[31] Harmonic frequencies were calculated for all species at the B3LYP/TZVP level.[32,33] The frequencies of the minima were all real, and the saddle points exhibited only one imaginary frequency. The harmonic frequencies were scaled by a factor of 0.9 and used to compute the zero-point vibrational energies (ZPVE) and vibrational partition functions.

Simple transition state theory (TST) was used to compute reaction rate constants. This theory assumes that the reaction rate is limited by the formation of the transition state, which is considered to be in pseudo-equilibrium with the reactants. The reaction rate constants can be calculated from the thermochemistry of reactants and transition states, that is, by calculating translational, rotational, and vibrational partition functions. In these calculations, the imaginary frequency of the transition structure is ignored and only $3N - 7$ molecular vibrations remain (for a nonlinear molecule).

The expression for the reaction rate constant of a unimolecular reaction is

$$k(T) = \kappa(T)\frac{k_B T}{h}\frac{Q^{\ddagger}_{\text{int}}(T)}{Q^{\text{R}}_{\text{int}}(T)}\exp(-E_a/RT) \tag{8}$$

where $Q^{\ddagger}_{\text{int}}$ and $Q^{\text{R}}_{\text{int}}$ are the partition functions of the transition state and reactants respectively, involving only internal coordinates, and calculated using the module Freeh, available in Turbomole. k_B is the Boltzmann constant and h the Planck constant. The temperature T is varied from 300 to 1300 K in the present work. E_a is the barrier height including ZPVE. $\kappa(T)$ is the transmission coefficient accounting for tunneling effects, computed in the present work from the formula[34]

$$\kappa(T) = 1 - \frac{1}{24}\left(\frac{h\upsilon^{\ddagger}}{k_B T}\right)^2\left(1 + \frac{RT}{E_a}\right) \tag{9}$$

Only the imaginary frequency associated with the reaction

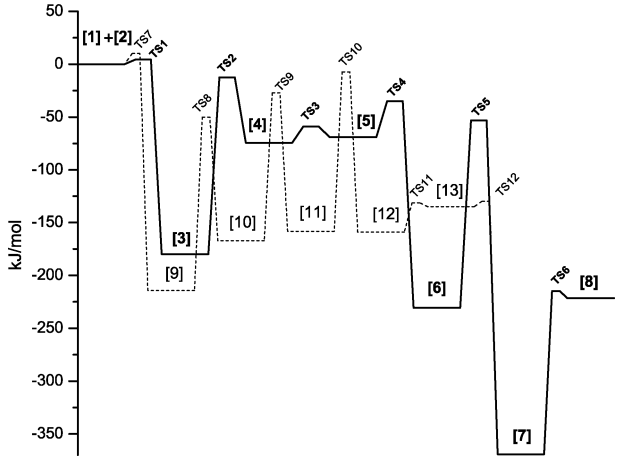


Figure 1. Formation of naphthalene, C_6H_5 [**1**] + C_4H_4 [**2**] → $C_{10}H_{10}$ [**8**], calculated at the B3LYP/TZVP level.

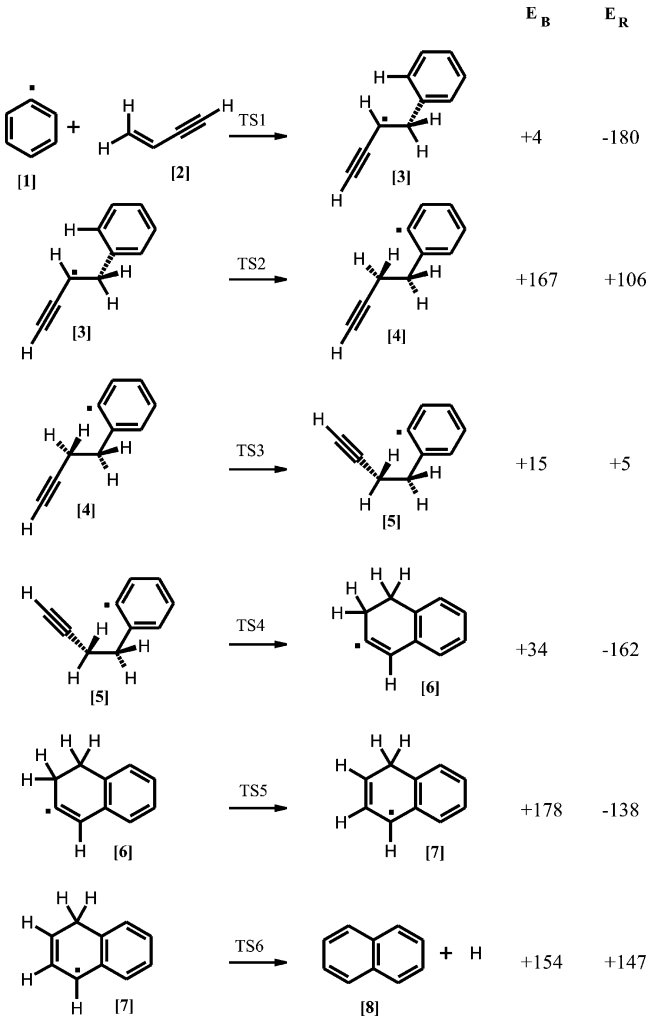


Figure 2. Intermediates and transition states on the first reaction path (solid line in Figure 1) toward the formation of naphthalene, C_6H_5 [**1**] + C_4H_4 [**2**] → $C_{10}H_{10}$ [**8**], calculated at the B3LYP/TZVP level (electronic energies in kJ/mol).

coordinate $\upsilon^{\ddagger}$ is required to calculate $\kappa(T)$. Notice that the minus sign of the second term on the right-hand side of eq 9 is cancelled by the square of the imaginary frequency $\upsilon^{\ddagger}$, such that $\kappa(T) > 1$.

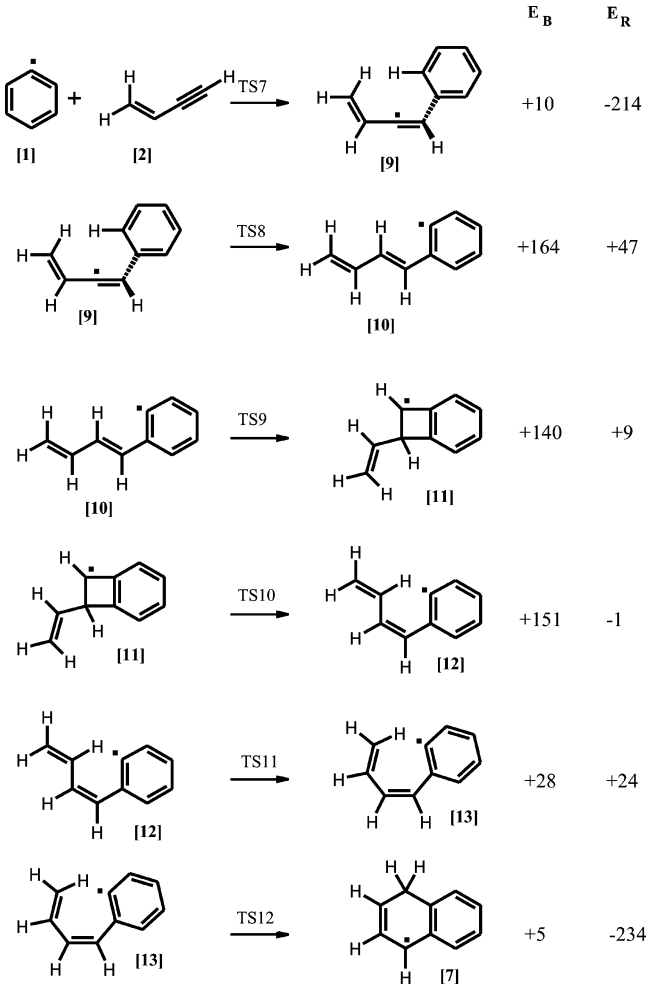


Figure 3. Intermediates and transition states on the second reaction path (dashed line in Figure 1) toward the formation of naphthalene, C_6H_5 [**1**] + C_4H_4 [**2**] → $C_{10}H_{10}$ [**8**], calculated at the B3LYP/TZVP level (electronic energies in kJ/mol).

For a bimolecular reaction, the expression for the reaction rate constant is

$$k(T) = \kappa(T)\,\bar{V}(T)\frac{k_B T}{h}\frac{Q^{\ddagger}(T)}{Q^{R_1}(T)Q^{R_2}(T)}\exp(-E_a/RT) \quad (10)$$

where $Q^{\ddagger}(T)$, $Q^{R_1}(T)$ and $Q^{R_2}(T)$ are the partition functions (including translation) of the activated complex and the two reactants R_1 and R_2, respectively, and $\bar{V}(T)$ is the molar volume of an ideal gas at temperature T.

Arrhenius-like expressions (eq 6) are fitted to the computed rate constants for a series of temperatures in the range from 300 to 1300 K, treating A and n as fitting parameters. E_a is taken as the energy difference (including ZPVE) between the transition state and reactants, and this energy difference is taken directly from the DFT calculation. If necessary, it could be replaced with a more accurate value, for example, from advanced coupled-cluster single-point calculations at the DFT geometries. We have performed single-point calculations with the functional BMK at the B3LYP geometries to obtain good estimates of E_a. It has been shown before that the BMK functional,[29,35] which has been designed and parametrized for kinetics, yields more reliable barrier

Table 2. Calculated TST Rate Constants for the Formation of Naphthalene

	T							
reaction	300	400	500	600	700	800	900	1000
1 + **2** → **3**[a]	5.94×10^{-15}	2.08×10^{-14}	5.23×10^{-14}	1.06×10^{-13}	1.90×10^{-13}	3.12×10^{-13}	4.73×10^{-13}	6.93×10^{-13}
3 → **4**[b]	4.16×10^{-16}	2.34×10^{-9}	2.58×10^{-5}	1.28×10^{-2}	1.07×10^{0}	3.03×10^{1}	4.13×10^{2}	3.32×10^{3}
4 → **5**[b]	1.28×10^{10}	5.95×10^{10}	1.52×10^{11}	2.79×10^{11}	4.33×10^{11}	6.04×10^{11}	7.80×10^{11}	9.57×10^{11}
5 → **6**[b]	8.43×10^{5}	2.11×10^{7}	1.43×10^{8}	5.32×10^{8}	1.32×10^{9}	2.66×10^{9}	4.60×10^{9}	7.13×10^{9}
6 → **7**[b]	8.81×10^{-16}	1.06×10^{-8}	1.91×10^{-4}	1.32×10^{-1}	1.45×10^{1}	4.94×10^{2}	7.86×10^{3}	7.19×10^{4}
7 → **8** + H[b]	8.74×10^{-12}	1.00×10^{-5}	4.67×10^{-2}	1.33×10^{1}	7.84×10^{2}	1.70×10^{4}	1.87×10^{5}	1.31×10^{6}
1 + **2** → **9**[a]	8.02×10^{-15}	4.87×10^{-14}	1.68×10^{-13}	4.25×10^{-13}	8.72×10^{-13}	1.62×10^{-12}	2.68×10^{-12}	4.23×10^{-12}
9 → **10**[b]	1.28×10^{-14}	5.06×10^{-8}	4.38×10^{-4}	1.85×10^{-1}	1.40×10^{1}	3.61×10^{2}	4.51×10^{3}	3.49×10^{4}
10 → **11**[b]	6.55×10^{-12}	5.36×10^{-6}	1.91×10^{-2}	4.47×10^{0}	2.29×10^{2}	4.34×10^{3}	4.26×10^{4}	2.66×10^{5}
11 → **12**[b]	2.18×10^{-13}	6.29×10^{-7}	4.78×10^{-3}	1.84×10^{0}	1.30×10^{2}	3.25×10^{3}	3.97×10^{4}	2.88×10^{5}
12 → **13**[b]	8.54×10^{7}	1.23×10^{9}	6.28×10^{9}	1.82×10^{10}	3.92×10^{10}	7.00×10^{10}	1.09×10^{11}	1.60×10^{11}
13 → **7**[b]	5.23×10^{11}	9.74×10^{11}	1.53×10^{12}	2.13×10^{12}	2.79×10^{12}	3.48×10^{12}	4.19×10^{12}	4.92×10^{12}

	T			fit: $AT^n \exp(-E_a/RT)$			
reaction	1100	1200	1300	*A*	*n*	E_a^{B3LYP}	E_a^{BMK} [c]
1 + **2** → **3**[a]	9.71×10^{-13}	1.28×10^{-12}	1.70×10^{-12}	2.09×10^{-20}	2.61	6.0	6.9
3 → **4**[b]	1.86×10^{4}	7.82×10^{4}	2.62×10^{5}	2.53×10^{13}	−0.55	157.0	157.4
4 → **5**[b]	1.13×10^{12}	1.33×10^{12}	1.49×10^{12}	2.71×10^{12}	0.11	14.8	15.3
5 → **6**[b]	1.00×10^{10}	1.34×10^{10}	1.72×10^{10}	6.85×10^{11}	−0.10	32.5	34.2
6 → **7**[b]	4.43×10^{5}	2.04×10^{6}	7.38×10^{6}	1.05×10^{13}	0.10	162.1	170.9
7 → **8** + H[b]	6.31×10^{6}	2.37×10^{7}	7.40×10^{7}	1.97×10^{10}	0.97	135.8	146.3
1 + **2** → **9**[a]	6.18×10^{-12}	8.62×10^{-12}	1.19×10^{-11}	1.07×10^{-19}	2.72	10.7	13.4
9 → **10**[b]	1.84×10^{5}	7.48×10^{5}	2.46×10^{6}	1.19×10^{14}	−0.51	152.8	157.7
10 → **11**[b]	1.20×10^{6}	4.23×10^{6}	1.21×10^{7}	5.54×10^{11}	0.23	134.4	135.9
11 → **12**[b]	1.48×10^{6}	5.75×10^{6}	1.82×10^{7}	7.00×10^{11}	0.41	145.7	177.3
12 → **13**[b]	2.13×10^{11}	2.72×10^{11}	3.38×10^{11}	8.85×10^{11}	0.20	25.7	22.5
13 → **7**[b]	5.65×10^{12}	6.47×10^{12}	7.23×10^{12}	1.49×10^{10}	0.91	4.2	4.0

[a] In $cm^3\ s^{-1}$ molecule^{-1}, with *T* in K and E_a in kJ/mol. [b] In s^{-1}, with *T* in K and E_a in kJ/mol. [c] This value includes a ZPVE correction calculated at the B3LYP level.

heights than B3LYP (but not necessarily more accurate geometries or vibrational frequencies).

Parts of the pathways were investigated with a variety of different functionals in order to assess the accuracy of the BMK results. Single-point calculations (at the B3LYP/TZVP geometries) were performed with the functionals BLYP,[28] TPSS,[36,37] TPSSh,[36−38] and B97-K.[29] It turned out that the computed energies were quite different for the four-membered ring intermediate, and we therefore studied the isomerization of the penta-1,3-dien-1-yl radical [**21**] as a model for the formation of this intermediate, using the MP2,[39] CC2,[40] and RCCSD(T)[41−43] methods, the latter as implemented in the Molpro[44−46] program package. The MP2 and CC2 calculations were performed with the Turbomole[15] program using the resolution-of-the-identity approximation.[39,40] The innershell electrons (carbon 1s) were not correlated (frozen core, FC, approximation) in any of the correlation treatments.

III. Results and Discussion

A. Formation of Naphthalene. We have computed two pathways for the formation of naphthalene from phenyl and vinylacetylene. On one pathway, the phenyl radical attacks the vinylacetylene at its double bond (solid line in Figure 1, see also Figure 2), and on the other pathway, the phenyl radical attacks the vinylacetylene at its triple bond (dashed line in 1, see also Figure 3). The attack at the double bond yields [**3**] and is followed by a 1,4 hydrogen shift, bringing the radical center back to the ring ([**4**]). Then, further reactions to naphthalene ([**8**]) are straightforward. The other possible pathway occurs after the phenyl's attack of the triple bond of vinylacetylene, yielding [**9**]. Here also, a 1,4 hydrogen shift brings the radical center back to the ring ([**10**]). To proceed further, a rotation about a double bond seems required to prepare for the closure of the second six-ring of naphthalene, and it was this rotation about a double bond that, in previous work, let the second pathway seem inefficient, because barriers to rotation about a double bond are usually very high.[5,9,10]

In our study, however, this rotation takes place in a two-step reaction via the intermediate [**11**]. This four-membered ring helps to decrease the barrier to "rotation" about the double bond significantly. From [**12**] onwards, the ring closure to yield naphthalene is straightforward. We stress that, in contrast to the work by Moriarty and Frenklach,[9] we do not find any structure (minimum or saddle point) with an electronic energy above the transition state of the first, bimolecular reaction step (TS1 or TS7, cf. Figure 1).

In Table 2, we provide the reaction rates obtained from TST for all of the steps of the two pathways in the temperature range 300−1300 K, along with the fitted Arrhenius parameters *A* and *n* and computed values of E_a (including ZPVE). We give the activation energies obtained from calculations with the density functionals B3LYP and

Table 3. Calculated TST Rate Constants for the Formation of Phenanthrene

	T							
reaction	300	400	500	600	700	800	900	1000
1 + **14** → **15**[a]	3.83×10^{-15}	2.65×10^{-14}	9.87×10^{-14}	2.61×10^{-13}	5.57×10^{-13}	1.06×10^{-12}	1.79×10^{-12}	2.80×10^{-12}
15 → **16**[b]	8.02×10^{-13}	1.00×10^{-6}	4.49×10^{-3}	1.21×10^{0}	6.59×10^{1}	1.35×10^{3}	1.43×10^{4}	9.51×10^{4}
16 → **17**[b]	1.11×10^{-10}	3.23×10^{-5}	6.18×10^{-2}	9.57×10^{0}	3.55×10^{2}	5.39×10^{3}	4.45×10^{4}	2.46×10^{5}
17 → **18**[b]	1.80×10^{-15}	1.65×10^{-8}	2.49×10^{-4}	1.54×10^{-1}	1.53×10^{1}	4.87×10^{2}	7.17×10^{3}	6.18×10^{4}
18 → **19**[b]	6.17×10^{8}	2.80×10^{9}	7.07×10^{9}	1.31×10^{10}	2.02×10^{10}	2.87×10^{10}	3.73×10^{10}	4.64×10^{10}
19 → **20** + H[b]	5.70×10^{-2}	2.03×10^{2}	2.91×10^{4}	8.36×10^{5}	9.50×10^{6}	5.85×10^{7}	2.49×10^{8}	7.96×10^{8}

	T			fit: $AT^n \exp(-E_a/RT)$			
reaction	1100	1200	1300	*A*	*n*	E_a^{B3LYP}	E_a^{BMK} [c]
1 + **14** → **15**[a]	4.27×10^{-12}	6.02×10^{-12}	8.30×10^{-12}	6.96×10^{-20}	2.74	11.8	14.7
15 → **16**[b]	4.51×10^{5}	1.63×10^{6}	4.89×10^{6}	8.82×10^{13}	−0.51	141.8	142.8
16 → **17**[b]	9.67×10^{5}	3.12×10^{6}	8.32×10^{6}	3.34×10^{11}	0.13	124.5	126.9
17 → **18**[b]	3.59×10^{5}	1.56×10^{6}	5.50×10^{6}	3.60×10^{11}	0.48	156.9	190.6
18 → **19**[b]	5.46×10^{10}	6.32×10^{10}	7.18×10^{10}	2.71×10^{11}	0.01	15.3	17.5
19 → **20** + H[b]	2.08×10^{9}	4.62×10^{9}	9.05×10^{9}	1.38×10^{11}	0.65	80.1	89.3

[a] In $cm^3\ s^{-1}molecule^{-1}$, with *T* in K and E_a in kJ/mol. [b] In s^{-1}, with *T* in K and E_a in kJ/mol. [c] This value includes a ZPVE correction calculated at the B3LYP level.

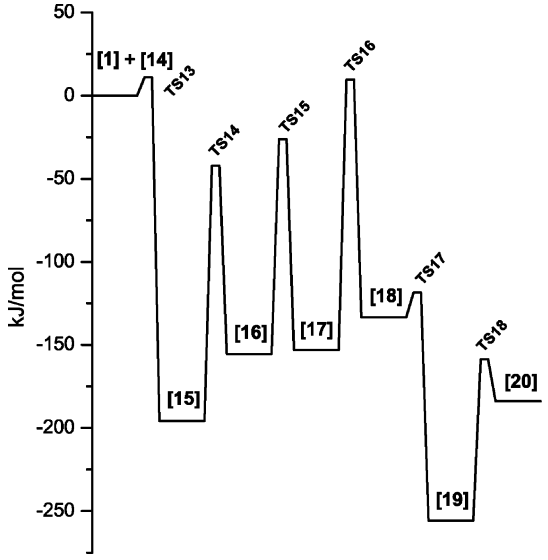


Figure 4. Formation of phenanthrene, C_6H_5 [**1**] + C_8H_6 [**14**] → $C_{14}H_{12}$ [**20**], calculated at the B3LYP/TZVP level.

BMK and recommend the latter for use in the mechanisms on which pyrolysis simulations such as those of Norinaga and Deutschmann are based.[13,14]

B. Formation of Phenanthrene. In Table 3, we provide the reaction rates obtained from TST for all of the steps of the reaction (cf. Figures 4 and 5) in the temperature range 300−1300 K, along with the fitted Arrhenius parameters *A* and *n* and computed values of E_a (including ZPVE). Not only for the formation of naphthalene but also for the formation of phenanthrene, we recommend the BMK activation energies for modeling purposes.

In the case of vinylacetylene, the phenyl radical can be added to both ends of the molecule, but in the case of phenylacetylene, the phenyl can only react with the triple bond. The addition of the phenyl radical is followed by a 1,4 hydrogen shift to yield the trans isomer [**16**], which isomerizes to the cis isomer [**18**] via the four-membered ring intermediate [**17**]. The electronic energy of the transition structure TS16 (10 kJ/mol, cf. Figure 4) lies just below the barrier (TS13, 11 kJ/mol) of the first, bimolecular reaction step. A new six-membered ring is formed in [**18**] to give phenanthrene [**20**].

C. Four-Membered Ring Intermediates. The radicalic intermediates [**11**] and [**17**] with a four-membered ring play a key role in the proposed reaction mechanisms for the formations of naphthalene and phenanthrene from reactions of phenyl with vinylacetylene and phenylacetylene, respectively. Unfortunately, we observe large discrepancies between the B3LYP and BMK results for the barriers between the four-membered ring intermediates and the cis isomers. For example, the BMK barrier (177.3 kJ/mol, including ZPVE) for the step [**11**] → [**12**] is more than 30 kJ/mol higher than the corresponding B3LYP barrier (145.7 kJ/mol, cf. Table 2). Also for the formation of phenanthrene, the BMK barrier (190.6 kJ/mol) for the step [**17**] → [**18**] is more than 30 kJ/mol higher than the corresponding B3LYP barrier (156.9 kJ/mol, cf. Table 3).

Similar data for the functionals BP86, BLYP, TPSS, TPSSh, and B97-K are collected in Table 4, showing that it is mainly the energy of the intermediates [**11**] and [**17**] that causes the large variations in computed activation energies for the [**11**] → [**12**] and [**17**] → [**18**] steps.

In an attempt to understand the different behavior of various exchange-correlation functionals, we have investigated the isomerization of the penta-1,3-dien-1-yl radical [**21**] as a model system (cf. Figure 6). For this model system, we have performed single-point RCCSD(T)(FC) calculations at the B3LYP/TZVP geometries with a restricted open-shell Hartree−Fock reference state (ROHF) using the correlation-consistent cc-pVTZ basis. According to the T1 diagnostic criteria,[47] the RCCSD(T) results do not suffer from a strong multireference character and represent a reliable set of reference data. The equilibrium and transition structures were determined at the B3LYP/TZVP level (Figure 6), and the results from various single-point calculations are shown in

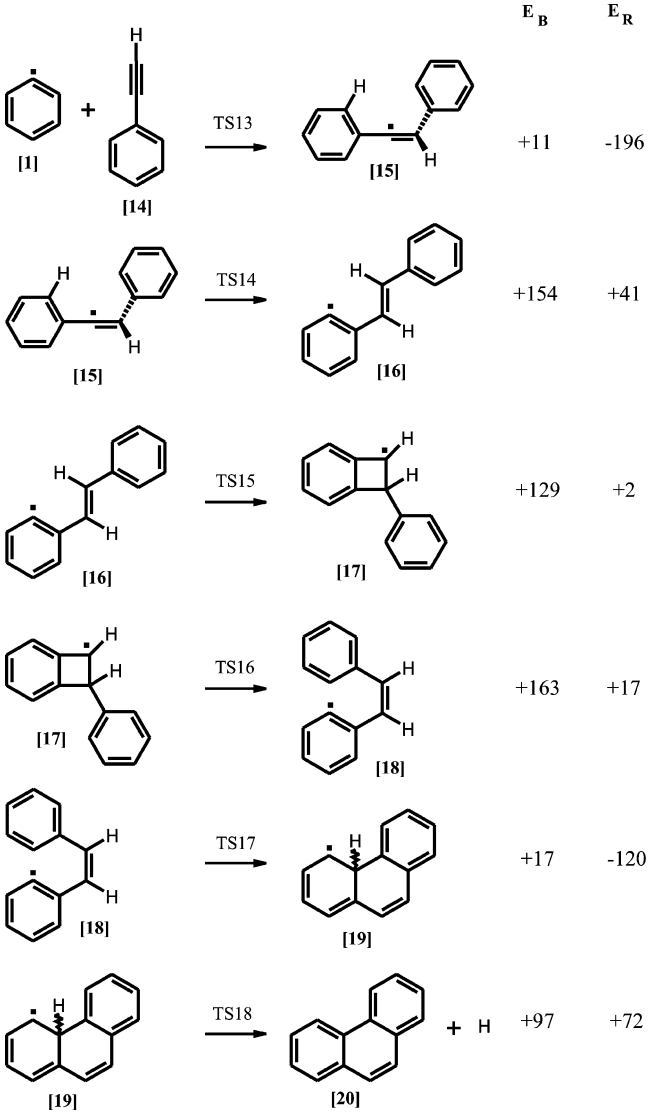


Figure 5. Intermediates and transition states on the reaction path (Figure 4) toward the formation of phenanthrene, C_6H_5 [**1**] + C_8H_6 [**14**] → $C_{14}H_{12}$ [**20**], calculated at the B3LYP/TZVP level (electronic energies in kJ/mol).

Table 4. Calculated (in the TZVP Basis) Relative Energies (in kJ/mol, with Respect To [**10**] and [**16**], Respectively) for the Formation of the Four-Membered Ring Intermediates [**11**] and [**17**] at the B3LYP/TZVP Geometries

method	TS9	[**11**]	TS10	TS15	[**17**]	TS16
BP86	122	−1	141	111	−8	146
BLYP	129	18	149	118	11	154
TPSS	125	0	144	115	−7	149
B3LYP	140	9	160	129	2	165
TPSSh	130	−4	149	119	−11	154
B97-K	142	−10	162	132	−16	169
BMK	141	−21	162	132	−29	168

Table 5. All of the DFT and coupled-cluster calculations indicate that the energy of the four-membered ring seems to be the most difficult to compute accurately. For the four-membered ring [**22**], the deviations of the results from the ROHF-CCSD(T)(FC) results vary between 51 (ROHF) and −40 kJ/mol [UHF-UMP2(FC)]. The corresponding deviations of the B3LYP (14 kJ/mol) and BMK (−16 kJ/mol) results are not so large but have different signs.

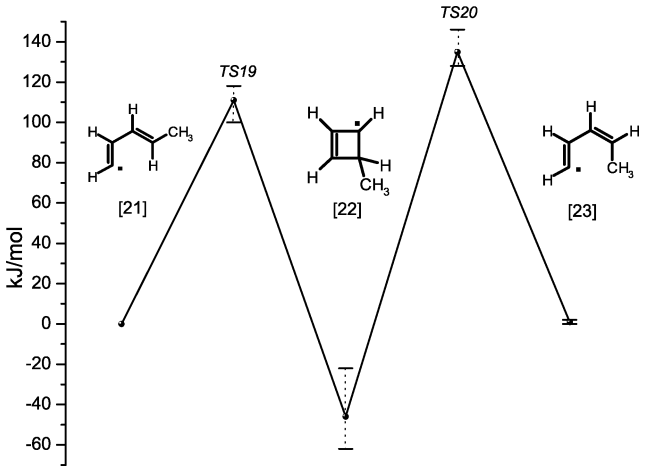


Figure 6. Energy profile of the isomerization of penta-1,3-dien-1-yl [**21**], see Table 5). The range of calculated DFT energies is indicated by error bars.

Table 5. Calculated (in the TZVP Basis) Relative Energies (in kJ/mol, with Respect To [**21**]) of the Model Reaction with Penta-1,3-dien-1-yl (cf. Figure 6) at the B3LYP/TZVP Geometries[a]

method	TS19	δ_{TS19}[b]	[**22**]	$\delta_{[22]}$[b]	TS20	δ_{TS20}[b]	[**23**]	$\delta_{[23]}$[b]
BP86	100	−11	−43	3	128	−7	0	−1
BLYP	110	−1	−22	24	136	1	2	1
TPSS	102	−9	−45	1	128	−7	1	0
B3LYP	117	6	−32	14	145	10	2	1
TPSSh	105	−6	−49	−3	132	−3	1	0
B97-K	118	7	−49	−3	146	11	2	1
BMK	116	5	−62	−16	144	9	0	−1
ROHF	182	71	5	51	211	76	6	5
UHF	137	26	−15	31	155	20	7	6
ROHF-UMP2(FC)	103	−8	−65	−19	129	−6	1	0
UHF-UMP2(FC)	124	13	−86	−40	152	17	0	−1
ROHF-UCC2(FC)	100	−11	−54	−8	124	−11	0	−1
UHF-UCC2(FC)	115	4	−67	−21	143	8	0	−1
ROHF-RCCSD(T)(FC)	111	0	−46	0	135	0	1	0

[a] A cc-pVTZ basis was used for the ROHF-RCCSD(T)(FC) calculations. [b] Deviation from the ROHF−RCCSD(T)(FC) result.

Spin contamination of the doublet states may be a reason for the differences between the DFT results. Indeed, the spin-unrestricted Hartree−Fock (UHF) expectation values of $\hat{S}^2$ amount to 1.30, 0.97, and 1.28 for the minima [**21**], [**22**], and [**23**], respectively, and as much as 1.45 and 1.53 for the transitions structures TS19 and TS20. The UHF and UHF-based values seem unreliable. On the other hand, all of the DFT values are much closer to 0.75 and are in good mutual agreement. For all of the minima, the DFT values range from 0.76 to 0.78 for all of the functionals studied, and the values for TS19 and TS20 vary between 0.77 and 0.81 and between 0.77 and 0.83, respectively. Overall, the DFT values agree to within 25 kJ/mol with the ROHF-CCSD(T)(FC) results, as can be seen from the error bars in Figure 6.

IV. Conclusions

We have proposed three reaction pathways for PAH growth from reactions of either vinyl- or phenylacetylene with a phenyl radical. We have found pathways via four-membered ring intermediates that enable a cis−trans isomerization. These intermediates have so far not been investigated in the literature.

Arrhenius parameters (pre-exponential A, temperature exponent n, and activation energy E_a) have been derived for all of the reaction steps on the three different pathways. We suggest the use of Arrhenius-like rate expressions such as eq 6 with our parameters for the first bimolecular steps in the kinetic modeling of refs 13 and 14.

Acknowledgment. This work was conducted in the Sonderforschungsbereich (SFB) 551 "Carbon from the Gas Phase: Elementary Reactions, Structures, Materials", which is funded by the Deutsche Forschungsgemeinschaft (DFG). We gratefully acknowledge the Fonds der Chemischen Industrie for additional financial support. We thank O. Deutschmann, K. Norinaga, and M. Olzmann for many fruitful discussions.

References

(1) Hüttinger, K. J. *Chem. Vap. Deposition* **1998**, *4*, 151.

(2) Hüttinger, K. J. In *Fibers and Composites*; Delhaès, P., Ed.; Taylor & Francis: Oxford, U. K., 2003; p 75.

(3) Becker, A.; Hüttinger, K. J. *Carbon* **1998**, *36*, 177.

(4) Norinaga, K.; Deutschmann, O.; Hüttinger, K. J. *Carbon* **2006**, *44*, 1790.

(5) Frenklach, M. *Phys. Chem. Chem. Phys.* **2002**, *4*, 2028.

(6) Richter, H.; Howard, J. B. *Phys. Chem. Chem. Phys.* **2002**, *4*, 2038.

(7) Klippenstein, S. J.; Miller, J. A. *J. Phys. Chem. A* **2005**, *109*, 4285.

(8) Appel, J.; Bockhorn, H.; Frenklach, M. *Combust. Flame* **2000**, *121*, 122.

(9) Moriarty, N. W.; Frenklach, M. *Proc. Combust. Inst.* **2000**, *28*, 2563.

(10) Bauschlicher, C. W., Jr.; Ricca, A. *Chem. Phys. Lett.* **2000**, *326*, 283.

(11) Frenklach, M.; Wang, H. *Proc. Combust. Inst.* **1991**, *23*, 1559.

(12) Kislov, V. V.; Mebel, A. M.; Lin, S. H. *J. Phys. Chem. A* **2002**, *106*, 6171.

(13) Norinaga, K.; Deutschmann, O. *Carbon* **2006**, submitted.

(14) Norinaga, K.; Deutschmann, O. In *Proceedings of the 15th European Conference on Chemical Vapor Deposition*; Electrochemical Society Proceedings: Pennington, NJ, 2005; Vol. 2005-09, p 348.

(15) *Turbomole*, v. 5.8; Cosmologic GmbH & Co.: Leverkusen, Germany, 2005. www.turbomole.com (accessed Nov 2006).

(16) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.

(17) Häser, M.; Ahlrichs, R. *J. Comput. Chem.* **1989**, *10*, 104.

(18) Treutler, O.; Ahlrichs, R. *J. Chem. Phys.* **1995**, *102*, 346.

(19) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571.

(20) All basis sets and auxiliary basis sets are available from ftp://ftp.chemie.uni-karlsruhe.de/pub (accessed Nov 2006).

(21) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *59*, 1200.

(22) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.

(23) Perdew, J. P. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *33*, 8822.

(24) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283.

(25) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *242*, 625.

(26) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.

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

(28) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.

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

(30) Von Arnim, M.; Ahlrichs, R. *J. Chem. Phys.* **1999**, *111*, 9183.

(31) Helgaker, T. *Chem. Phys. Lett.* **1991**, *182*, 503.

(32) Deglmann, P.; Furche, F.; Ahlrichs, R. *Chem. Phys. Lett.* **2002**, *362*, 511.

(33) Deglmann, P.; Furche, F. *J. Chem. Phys.* **2002**, *117*, 9535.

(34) Steinfield, J. I.; Francisco, J. S.; Hase, W. L. *Chemical Kinetics and Dynamics*, 2nd ed.; Prentice Hall: New York, 1999.

(35) Carissan, Y.; Klopper, W. *ChemPhysChem* **2006**, *7*, 1770.

(36) Perdew, J. P.; Wang, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *45*, 13244.

(37) Tao, J.; Perdew, J .P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(38) Staroverov, V. N.; Scuseria, G. E.; Perdew, J. P.; Tao, J. *J. Chem. Phys.* **2003**, *119*, 12129.

(39) Weigend, F.; Häser. M. *Theor. Chem. Acc.* **1997**, *97*, 331.

(40) Hättig, C.; Weigend, F. *J. Chem. Phys.* **2000**, *113*, 5154.

(41) Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1993**, *99*, 5219.

(42) Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **2000**, *112*, 3106.

(43) Watts, J. D.; Gauss, J.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 8718.

(44) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R.D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklaβ, A.; Palmierir, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. *Molpro*, version 2002.6; University College Cardiff Consultants Limited: Wales, U. K., 2003. See http://www.molpro.net (accessed Nov 2006).

(45) Lindh, R.; Ryu, U.; Liu, B. *J. Chem. Phys.* **1991**, *95*, 5889.

(46) Hampel, C.; Peterson, K. A.; Werner, H.-J. *Chem. Phys. Lett.* **1992**, *190*, 1.

(47) Lee, T. J.; Taylor, P. R. *Int. J. Quantum Chem.* **1989**, *S23*, 199.

CT600255U

JCTC Journal of Chemical Theory and Computation

Computational Investigation of Mechanisms for Ring-Opening Polymerization of ϵ-Caprolactone: Evidence for Bifunctional Catalysis by Alcohols

Nicholas Buis,[†] Samuel A. French,[‡] Giuseppe D. Ruggiero,[†] Bruno Stengel,[§] Arran A. D. Tulloch,[§] and Ian H. Williams*,[†]

Department of Chemistry, University of Bath, Bath, BA2 7AY, United Kingdom, Johnson Matthey Technology Centre, Blounts Court, Sonning Common, Reading, RG4 9NH, United Kingdom, and Johnson Matthey Catalysts, P.O. Box 1, Belasis Avenue, Billingham, Cleveland, TS23 1LB, United Kingdom

Received August 11, 2006

Abstract: Stepwise addition/elimination and concerted mechanisms for the methanolysis of ϵ-caprolactone, as a model for the initiation and propagation of ring-opening polymerization (ROP), have been investigated computationally using the B3LYP/6-31G* density functional method, with assistance from one or two ancillary methanol molecules. The effects of specific solvation by these extra methanols in cyclic hydrogen-bonded clusters are very significant, with barrier height reductions of about 50 kJ mol^{-1}. However, the effects of bulk solvation as treated by the polarized continuum model are almost negligible. Increasing the ring size lowers the barriers for both the addition and elimination steps of the stepwise mechanism but does not do so for the concerted mechanism; a stepwise mechanism is preferred for methanol-assisted ROP. The essential catalytic role of solvent molecules in this reaction is to avoid the unfavorable accumulation or separation of charges.

Introduction

The ring-opening polymerization (ROP) of a variety of cyclic esters, including ϵ-caprolactone (ϵ-CL) to form biocompatible, biodegradable polymers has generated great interest in recent years. This is due to the potential application of these materials in areas of medicine such as drug delivery, implants, and tissue engineering.[1] New polymers are continually being developed, creating a great demand for new materials and for improved syntheses. ROP is the most commonly used synthetic strategy for preparing these macromolecules and can be carried out with overall high efficiency using transition-metal (TM) initiating compounds.[2]

Despite controversy regarding some initiators, ROP has been shown to proceed by a coordination-insertion mechanism with acyl−oxygen cleavage of the monomer and insertion into the metal−oxygen bond of the initiator. Much work has been focused on the synthesis of new tin, aluminum, and TM initiators,[3] with the goal of increasing reactivity and the production of novel polymer structures.[4] Reaction mechanisms involving TMs have been investigated by kinetic, spectroscopic,[5,6] or chromatographic methods.[7] Commercial catalysts for these important processes are currently based on complexes of tin,[8] but with growing environmental concerns, there is a need to develop more benign catalysts.

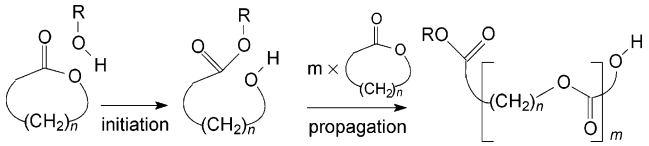
As the first part of an extensive computational modeling study of ROP mechanisms catalyzed by TM complexes,[9] we wish to report now upon alternative mechanisms for the reference reaction of ϵ-CL initiated by alcohol with a view to determining what factors govern the reactivity of these systems.

The first step of lactone ROP requires an initiator ROH, as shown in the initiation step of Scheme 1; subsequent chain-propagation steps involve the ring-opened intermediate

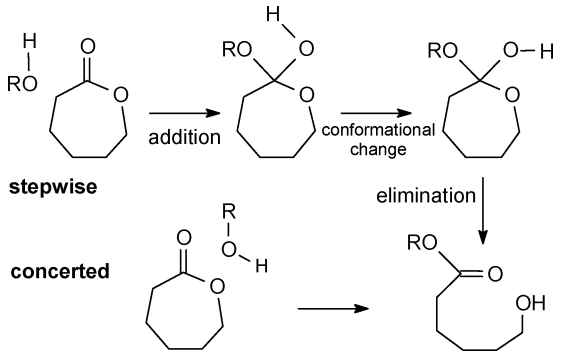
* Corresponding author fax: +44-1225-386231; e-mail: i.h.williams@bath.ac.uk.
[†] University of Bath.
[‡] Johnson Matthey Technology Centre, Sonning Common.
[§] Johnson Matthey Catalysts, Billingham.

10.1021/ct600265c CCC: $37.00
Published on Web 12/09/2006

Scheme 1



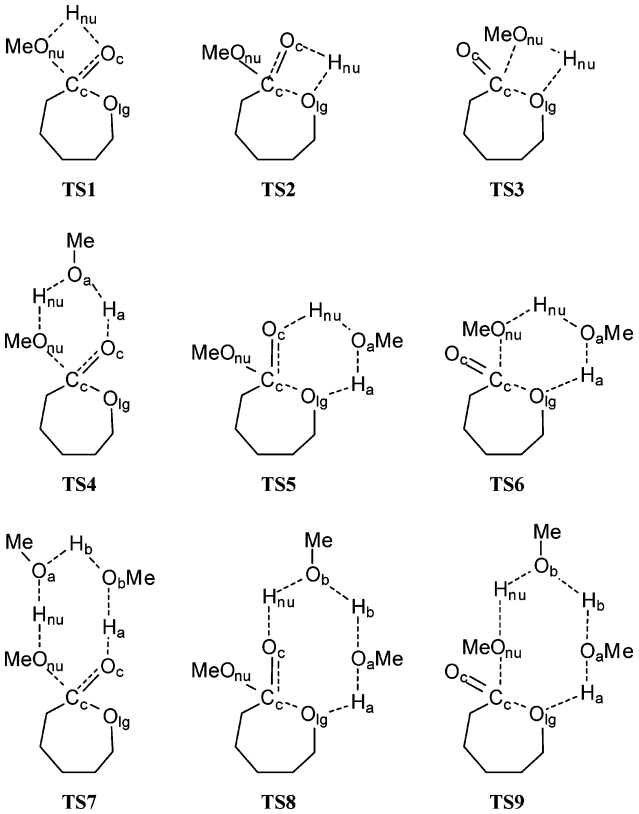
Scheme 2



reacting with further lactone monomers to yield polymers. Both initiation and propagation steps may be subject to catalysis. We have employed density functional theory to perform a computational study of mechanisms (Scheme 2) for the initiation step with ϵ-CL and R═Me (methanol). This serves also as a model for the propagation steps, where ROH is ring-opened ω-hydroxyester. We have considered two overall mechanisms: (a) the stepwise addition−elimination of MeOH across the acyl C═O bond and (b) a concerted addition of MeOH across the acyl C−O bond (metathesis). For each alternative, we have also considered the possibility of either one, two, or three MeOH molecules in cyclic hydrogen-bonded complexes (Chart 1 shows transition structures **TS1**−**TS9** for these complexes), analogous to those found in many other related processes.[10] Analogous mechanisms for the neutral and water-assisted hydrolysis of acyclic esters, methyl formate[11] and ethyl acetate,[12] have been the subjects of previous computational studies, but the present work is the first such study of lactone alcoholysis.

A cyclic transition state involving three molecules of methanol has been proposed by Venkatasubban et al.[13a] for the neutral methanolysis of *p*-nitrophenyl trifluoroacetate in acetonitrile, for which the apparent kinetic order with respect to MeOH is 2.87. Hydrolysis of the same ester in acetonitrile shows a similar apparent kinetic order of 3.22 with respect to both H_2O and D_2O, and proton inventory studies with H_2O/D_2O mixtures suggest a transition state involving the transfer of three protons. These authors recognized that similar cyclic transition states could exist for both the formation and breakdown of a hemiorthoester intermediate. Similar results have been obtained for the hydrolysis of *p*-methylphenyl trichloroacetate in acetonitrile.[13b] The cyclic transition states for reactions in acetonitrile/water mixtures are in contrast to the acyclic three-proton model proposed for ester hydrolysis in water. It is thought that this difference is due to the stabilization of the developing charges on the oxygens by hydrogen bonds to bulk water in aqueous solution; this kind of stabilization is difficult to achieve in a predominantly nonaqueous medium, which forces a cyclic transition state with minimal development of the charges.[13a]

Chart 1. Cyclic Transition Structures for Stepwise and Concerted Mechanisms of ϵ-CL ROP Initiation Involving One, Two, or Three CH_3OH Molecules[a]



[a] Subscripts denote the following: nu, nucleophile; lg, leaving group; c, carbonyl; a, first ancillary molecule; b, second ancillary molecule.

Methods

The calculations of density functional theory using the B3LYP exchange-correlation functional[14,15] were carried out with the GAMESS-UK[16] and Gaussian 03 packages of programs,[17] using the 6-31G* basis set.[18] The geometries and energies of each structure were fully optimized and its nature confirmed by frequency analysis. Transition structures (TSs) were located by means of the EF algorithm[19] and characterized by frequency calculations, both in vacuo and in solution. Intrinsic reaction co-ordinate (IRC) calculations were used to verify the adjacent local minima. Charge distributions were obtained by natural population analysis.[20] Solvent effects were examined by means of the polarized continuum method (PCM),[21] whereby the solvent was characterized by its dielectric constant (ϵ = 32.6 for methanol and ϵ = 78.4 water) around a cavity shaped by a superposition of spherical solute atoms; frequencies and zero-point energies were computed for each stationary structure in PCM methanol and water. Computed bond lengths are given in angströms, bond angles in degrees, total energies in hartrees, and unless otherwise stated zero-point energies and relative energies in kJ mol^{-1}.

Results and Discussion

Table 1 contains relative energies in a vacuum, methanol, and water for each B3LYP/6-31G*-optimized geometry with respect to the reactant complex for the methanolysis of ϵ-CL.

Table 1. B3LYP/6-31G* Relative Energies (kJ mol^{-1}) in a Vacuum (Potential Energy ΔE_{tot}, Zero-Point-Corrected Energy ΔE_0, and Free Energy ΔG_{298} at 298 K and 1 atm) and in PCM Methanol and Water for the Methanolysis of ϵ-Caprolactone

	vacuum			MeOH	H_2O
species	ΔE_{tot}	ΔE_0	ΔG_{298}	ΔE_0	ΔE_0
One Methanol					
ϵ-CL + CH_3OH **(3)**	31	27			
ϵ-CL·CH_3OH **(4)**	0	0	0	0	0
addition TS **(TS1)**	173	163	179	173	176
intermediate **(5a)**	28	33	50	43	47
conformational TS **(5a → 5b)**	42	48	66		
intermediate **(5b)**	18	23	41	32	43
elimination TS **(TS2)**	155	145	162	166	166
product **(6)**	−10	−8	2	8	3
concerted TS **(TS3)**	172	160	176	176	176
Two Methanols					
ϵ-CL + $2CH_3OH$ **(7)**	91	77			
ϵ-CL·$2CH_3OH$ **(8)**	0	0	0		
addition TS **(TS4)**	138	124	138	126	127
intermediate + CH_3OH **(9a)**	42	48	60	59	61
intermediate + CH_3OH **(9b)**	30	35	45	34	39
elimination TS **(TS5)**	119	104	118	114	120
product·CH_3OH **(10)**	−14	−14	−15	−21	−16
concerted TS **(TS6)**	178	160	172	169	167
Three Methanols					
e-CL + $3CH_3OH$ **(11)**	146	126			
e-CL·$3CH_3OH$ **(12)**	0	0	0		
addition TS **(TS7)**	125	101	113	95	100
intermediate + $2CH_3OH$ **(13a)**	33	38	45	61	60
intermediate + $2CH_3OH$ **(13b)**	33	37	44	33	38
elimination TS **(TS8)**	129	110	124	106	109
product·$2CH_3OH$ **(14)**	−9	−10	−14	−15	−15
concerted TS **(TS9)**	174	161	177	165	166

Potential energy differences ΔE_{tot}, zero-point-corrected 0 K energy differences ΔE_0, and free energy differences ΔG_{298} at 298 K and 1 atm are shown for calculation in a vacuum, whereas only the ΔE_0 results are shown for solvated species. The relative free energies for species along the reaction coordinates within the cyclic complexes are independent of the choice of standard state. The discussion below is based on the 0 K energies unless otherwise noted. Figures 1−3 respectively show structures and relative energies for reactions involving one, two, and three methanol molecules. The atom-labeling scheme used throughout is shown in Chart 1.

ϵ-CL ROP with One CH_3OH Molecule. The formation of reactant complex **4** from ϵ-CL **1** and a single methanol **2** (the combined energy of which is denoted by **3**) is energetically favorable by 27 kJ mol^{-1} in a vacuum at 0 K. The addition of $MeO_{nu}H_{nu}$ to the C_c═O_c double bond, the first step of the stepwise mechanism, involves a four-membered-ring **TS1** in which formation of the new O_{nuc}−C_c bond and breaking of the carbonyl π bond are coupled to proton transfer from the nucleophile to the carbonyl oxygen atom O_c, yielding a hemiacetal intermediate **5a** that is 33 kJ mol^{-1} above the reactant complex (Scheme 2). This tetrahedral adduct must undergo a favorable conformational change (−10 kJ mol^{-1}) to **5b** in order to facilitate the elimination step stereoelectronically. The transition structure **5a → 5b** for this conformational change involves an energy barrier of only 15 kJ mol^{-1}. The second step of the stepwise mechanism also involves a four-membered-ring **TS2**, in which cleavage of the C_c−O_{lg} bond and reformation of the carbonyl π bond are coupled to proton transfer from the hemiacetal O_c−H to the ring oxygen atom O_{lg}, giving the ring-opened methyl 6-hydroxyhexanoate product **6**, which lies 8 kJ mol^{-1} in energy below **4**. The energies of **TS1** and **TS2** relative to **4** are, respectively, 163 and 145 kJ mol^{-1}, indicating that the first step (addition) would be rate-determining. The calculated barrier height for the addition of methanol to ϵ-CL is very similar in magnitude to those calculated previously for the addition of one molecule of water or methanol to aldehydes by means of a four-membered-ring TS.[22]

Along the concerted reaction path, the methanol nucleophile attacks the carbonyl C_c but transfers its proton directly to the leaving-group O_{lg} atom of the lactone. This mechanism is comparable to intermolecular concerted S_N2 acyl transfer, but its quasi-intramolecular nature imposes a severe geometrical constraint such that it would be considered as front-side nucleophilic attack. Alternatively, the mechanism may be regarded as a metathesis, with two new σ bonds (O_{nuc}−C_c and H−O_{lg}) being formed simultaneously with breaking of the methanol O−H and lactone C_c−O_{lg} σ bonds, leading directly to the hydroxyester product **6**. The four-membered-ring **TS3** (Figure 1) lies 160 kJ mol^{-1} above the energy of the reactant complex **4**, suggesting that both stepwise and concerted mechanisms might be competitive for ROP involving only one molecule of methanol. However, **TS1**, **TS2**, and **TS3** are each highly strained, and it is unlikely that any significant fraction of the actual gas-phase alcoholysis of ϵ-CL proceeds by means of either of these one-methanol mechanisms since both involve a very high energy barrier. The unfavorable strain energy could be relieved by the incorporation of one or two additional methanol molecules into a cyclic hydrogen-bonded supermolecule, as discussed below.

ϵ-CL ROP with Two CH_3OH Molecules. An additional methanol molecule may be incorporated into a cyclic hydrogen-bonded complex such that it acts both as a proton donor to the ϵ-CL and as a proton acceptor to the original (nucleophilic) methanol. Thus, the ancillary methanol (MeO_a-H_a in Chart 1) may serve as a bifunctional catalyst.[10a,c,23]

The formation of reactant complex **8** from ϵ-CL **1** and 2× MeOH **2** is energetically favorable by 77 kJ mol^{-1} in a vacuum; note that cooperativity among the interactions between the component species in **8** causes the association energy to be greater than twice that in **4**. The methanol-assisted addition of methanol to the C_c═O_c double bond, the first (addition) step of the stepwise mechanism, involves a six-membered-ring **TS4** in which formation of the new O_{nu}−C_c bond and breaking of the carbonyl π bond are coupled to a double proton transfer from the nucleophile via the ancillary MeOH to O_c, yielding a complex **9a** of methanol with a hemiacetal intermediate that is 48 kJ mol^{-1} above **8** (Figure 2). This tetrahedral adduct must undergo a favorable

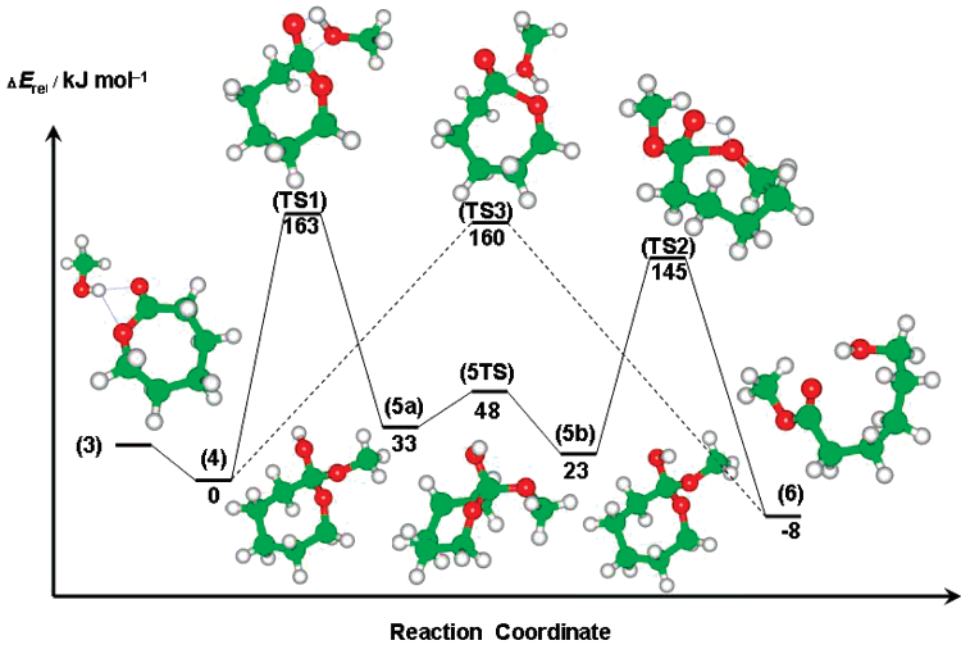


Figure 1. Zero-point-corrected B3LYP/6-31G* energy profile in a vacuum for the stepwise (solid line) and concerted (dashed line) mechanisms of ϵ-CL ROP with one CH_3OH molecule.

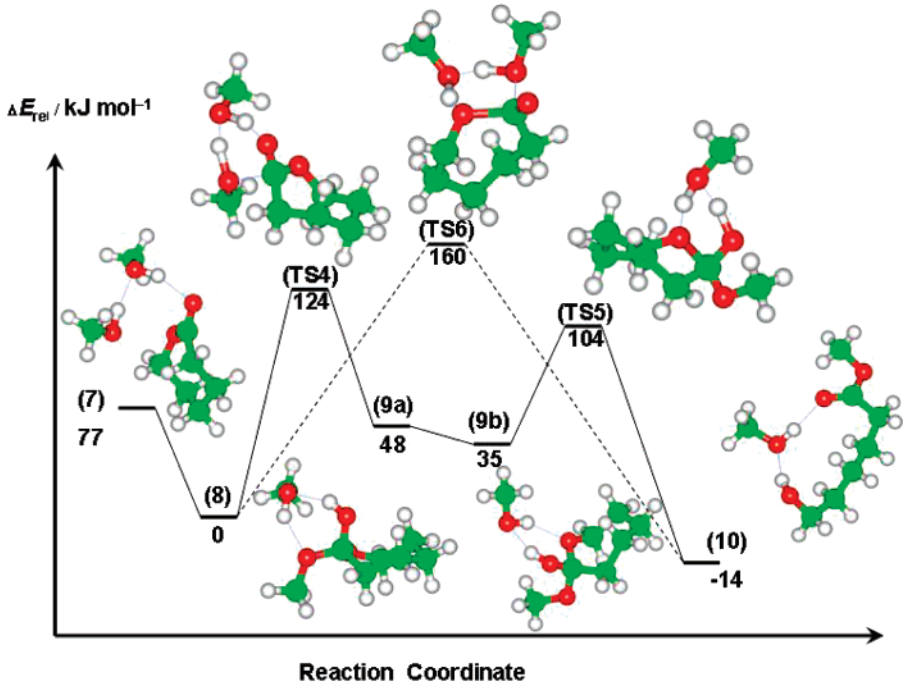


Figure 2. Zero-point-corrected B3LYP/6-31G* energy profile in a vacuum for the stepwise (solid line) and concerted (dashed line) mechanisms of ϵ-CL ROP with two CH_3OH molecules.

conformational change (-13 kJ mol^{-1}) to **9b** in order to facilitate the elimination step stereoelectronically. This change involves not only internal rotation of the hydroxyl group but also a repositioning of the ancillary MeOH; however, as for TS **5a** → **5b**, the energy barrier is not expected to be large, and the transition structure for this conformational rearrangement has not been characterized in this study. The second (elimination) step of this mechanism also involves a six-membered-ring **TS5**, in which cleavage of the C_c-O_{lg} bond and reformation of the carbonyl π bond are coupled to proton transfer from the hemiacetal O_c-H via the ancillary MeOH to O_{lg}, giving the ring-opened product complex **10** of methanol with methyl 6-hydroxyhexanoate, which lies 14 kJ mol^{-1} in energy below **8**. The energies of **TS4** and **TS5** relative to **8** are, respectively, 124 and 104 kJ mol^{-1}, indicating that the addition step would be rate-determining.

Along the reaction path for the concerted mechanism, the nucleophilic methanol attacks C_c with a transfer of its proton via the ancillary MeOH directly to O_{lg}, which is the leaving group. Again, this mechanism resembles concerted S_N2 acyl transfer, but still its quasi-intramolecular nature involves a

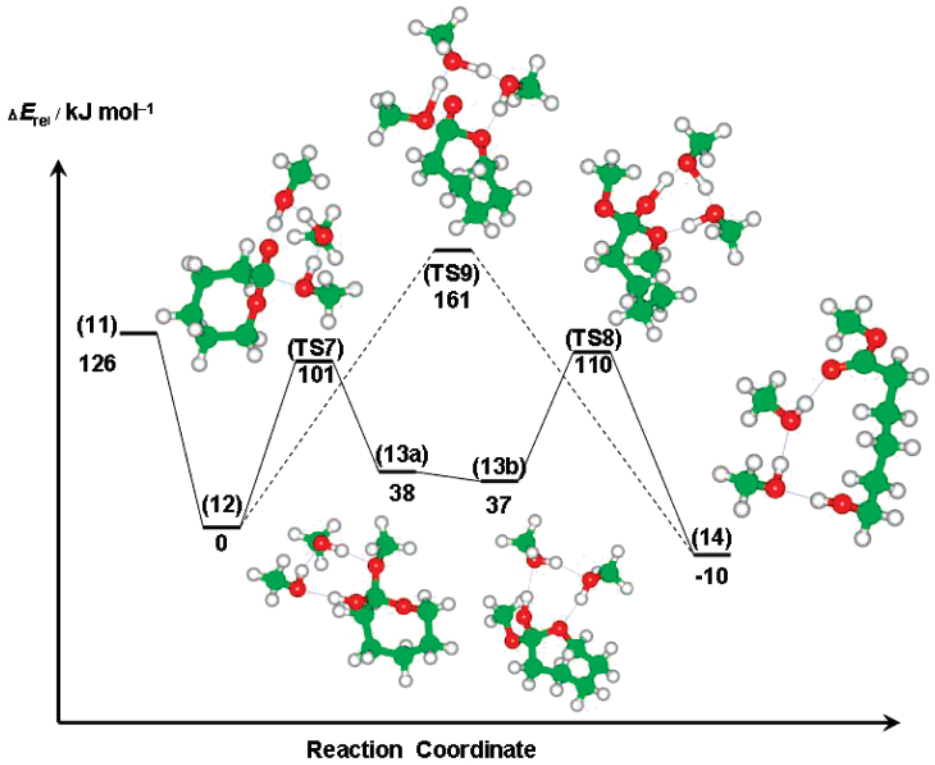


Figure 3. Zero-point-corrected B3LYP/6-31G* energy profile in a vacuum for the stepwise (solid line) and concerted (dashed line) mechanisms of ϵ-CL ROP with three CH_3OH molecules.

very severe geometrical constraint ($O_{nu}C_cO_{lg}$ angle = 93°) despite the expanded size of the cyclic TS. The six-membered-ring **TS6** (Figure 2) once again lies 160 kJ mol^{-1} above the energy of the reactant complex **8**, indicating that expansion from a four-membered ring to a six-membered ring is not accompanied by reduction of the energy barrier for the concerted mechanism. Whereas the reaction involving only one molecule of methanol showed that both the stepwise and concerted mechanisms might be competitive for ROP, it is clear that the methanol-assisted reaction favors the stepwise addition/elimination, for which the rate-determining step is 36 kJ mol^{-1} lower than the barrier for the concerted mechanism.

ϵ-CL ROP with Three CH_3OH Molecules. The formation of reactant complex **12** from ϵ-CL **1** and 3× MeOH **2** is energetically favorable by 126 kJ mol^{-1} in a vacuum; again, cooperativity among the interactions between the component species in **12** causes the association energy to be greater than thrice that in **4**. Incorporation of the third methanol molecule further expands the cyclic hydrogen-bonded complex to an eight-membered ring. Addition to the $C_c{=}O_c$ double bond is coupled to a triple proton transfer: from the nucleophile to the first ancillary methanol, from this to the second ancillary methanol, and from that to O_c. Inspection of the transition vector and the IRC path confirms that **TS7** is the only stationary point between **12** and the tetrahedral intermediate **13a**, which lies 38 kJ mol^{-1} above the reactant complex (Figure 3). Following a conformational change and repositioning of the two ancillary methanols, the elimination phase of the stepwise mechanism proceeds from intermediate **13b** by means of another eight-membered cyclic **TS8**, again involving a triple proton transfer from O_c to O_{lg} via the two ancillary methanols. The energies of **TS7** and **TS8** relative to **12** are, respectively, 101 and 110 kJ mol^{-1}, now indicating that the elimination step would be rate-determining. The barrier for the addition step via **TS7** is 23 kJ mol^{-1} lower than that via **TS4** with one fewer methanol, whereas the barrier for the elimination step via **TS8** is a little higher than that via **TS5** with the smaller ring. Note that **TS8** (as shown in Figure 3) is the lowest of three conformations considered with differing relative orientations of the methyl groups of the ancillary methanols: the other conformations were 6 and 9 kJ mol^{-1} higher in energy.

Along the concerted reaction path, **TS9** again involves a triple proton transfer from O_{nu} to O_{lg} via the two ancillary methanols in an eight-membered ring. At 161 kJ mol^{-1}, the barrier for this mechanism is remarkably similar to those for the concerted mechanisms via **TS3** and **TS6** with the smaller rings. Inclusion of the third methanol favors the stepwise over the concerted mechanism by 51 kJ mol^{-1}. Note that these (relative) energies are zero-Kelvin enthalpy changes; consideration of temperature-dependent entropy contributions within the cyclic complexes does not materially alter any of the conclusions in the above discussion: the relative free energies at 298 K (Table 1) for the various TSs of interest are all raised by a small amount in the range 12−17 kJ mol^{-1} but do not show any changes of mechanistic significance.

Continuum Solvent Effects. The reoptimization of all stationary structures with PCM methanol and PCM water, and the inclusion of zero-point energies obtained from vibrational frequencies computed with these continuum models, leads to very similar energetic trends to those discussed above for the reactions in a vacuum (Table 1).

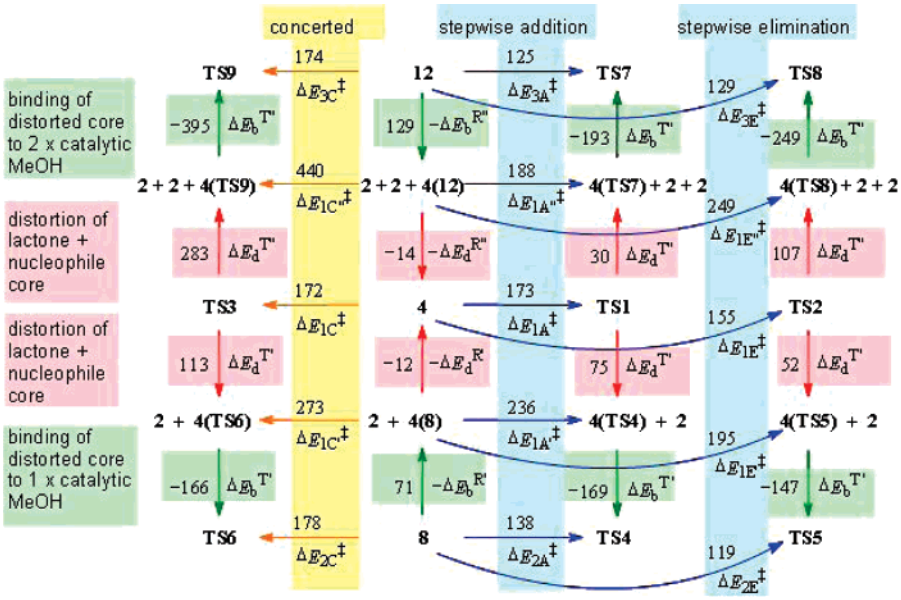


Figure 4. Energetics of ϵ-CL methanolysis analyzed in terms of contributions from the binding (green) of one (lower half) or two (upper half) ancillary methanol molecules and distortion (pink) of the lactone + nucleophilic methanol reacting core for the concerted (yellow, left half) and stepwise addition and elimination (blue, right half) mechanisms.

Although the concerted mechanism is competitive with the stepwise addition/elimination mechanism for the unassisted reaction, the stepwise mechanism is favored for reactions assisted by either one or two ancillary methanols. The addition step is rate-determining for reactions involving four- or six-membered rings, and the barrier height decreases with increasing ring size in both solvents, with the consequence that the elimination step becomes rate-determining for the largest (eight-membered) ring. Most of the barrier heights are slightly larger in solution than in a vacuum. However, the overall effect of PCM solvation, in which methanol or water is treated as a continuum characterized by the dielectric constant of the bulk solvent, upon the reaction energetics is rather small and uninteresting by comparison with the specific effects of the ancillary methanols considered in the next section.

Specific Solvation by Ancillary Methanol Molecules: Analysis of Distortion and Binding Energy Contributions. The purpose of this section is to consider how the barrier-height reductions reported above are achieved. The analysis presented assumes a "fundamentalist" view of catalysis, which focuses upon the differential stabilization of the transition state relative to the reactant state. The source of the catalytic effect of the ancillary methanol molecule(s) may be analyzed most directly in terms of potential energies (i.e., not including zero-point energies). It is convenient to consider the interaction energy of either a reactant complex or a transition structure with a catalyst as the sum of two components. The first is a distortion energy required to alter the geometries of the "core" fragments (lactone + nucleophile) of the uncatalyzed reaction to the structures they possess in the catalyzed reaction. The second is an intrinsic binding energy for bringing the distorted core together with the catalyst, which is either a single ancillary methanol (1× catalytic MeOH) or two ancillary methanols (2× catalytic MeOH). These components are color-coded in Figure 4 (pink for distortion and green for binding) and applied to both concerted (yellow) and stepwise (blue) mechanisms.

The potential energy barrier $\Delta E_{2A}^{\ddagger}$ for the addition step of the two-methanol stepwise mechanism (**8** → **TS4**, lower-right of Figure 4) may be expressed by eq 1.

$$\Delta E_{2A}^{\ddagger} = \Delta E_{1A}^{\ddagger} - \Delta E_b^{R'} - \Delta E_d^{R'} + \Delta E_b^{T'} + \Delta E_d^{T'} \quad (1)$$

$\Delta E_{1A}^{\ddagger}$ is the barrier height (173 kJ mol^{-1}) for the addition step of the one-methanol stepwise mechanism (**4** → **TS1**, middle right of Figure 4). $\Delta E_b^{R'}$ and $\Delta E_b^{T'}$ are the respective binding energies (−71 and −169 kJ mol^{-1}) of one additional methanol **2** with species **4(8)** and **4(TS4)**, which are the distorted "core" (ϵ-CL + nucleophilic methanol) fragments **4** as found in reactant complex **8** and transition structure **TS4**. $\Delta E_d^{R'}$ and $\Delta E_d^{T'}$ are the corresponding respective distortion energies (12 and 75 kJ mol^{-1}) of those core fragments relative to **4** and **TS1**. The ancillary methanol, which acts as both a hydrogen-bond acceptor and a hydrogen-bond donor, has a much stronger interaction with the distorted core in **TS4** than in **8** because of the greater dipolar character that develops as the nucleophilic addition proceeds;[25] the dipole moment of the core fragment increases from 3.44 d in **4(8)** to 4.77 d in **4(TS4).** The distortion energy of the ϵ-CL + nucleophile core from **TS1** to **TS4** is greater than that from **4** to **4(8)**. The difference $\Delta E_b^{T'} - \Delta E_b^{R'} = -98$ kJ mol^{-1} is the contribution to catalysis from the interaction of the reacting core with one ancillary methanol serving as a bifunctional catalyst. This amount is offset by the difference $\Delta E_d^{T'} - \Delta E_d^{R'} = +63$ kJ mol^{-1}, which is the net distortion energy that must be paid in order for the core to attain the necessary geometry to realize its increased interaction with the catalyst. The net catalytic effect $\Delta E_{1A}^{\ddagger} - \Delta E_{2A}^{\ddagger}$ is 35 kJ mol^{-1} (Table 2) for a single ancillary methanol on the addition step of the stepwise mechanism.

A similar analysis follows for catalysis of the stepwise mechanism by two ancillary methanols. The potential energy

Table 2. Analysis of the Effect of the Number *N* of Ancillary Methanol Molecules upon the Relative Energy Barriers (kJ mol^{-1}) for the Stepwise and Concerted Mechanisms

	addition			concerted			elimination		
$N=$	0	1	2	0	1	2	0	1	2
$\Delta E_{core}^{\ddagger}$	173	173	173	172	172	172	155	155	155
$\Delta E_{distortion}^{\ddagger}$	0	63	16	0	101	268	0	40	93
$\Delta E_{binding}^{\ddagger}$	0	−98	−64	0	−95	−266	0	−76	−120
$\Delta E_{total}^{\ddagger}$	173	138	125	172	178	174	155	119	129
$\Delta E_{catalysis}^{\ddagger}$		−35	−48		+6	+2		−36	−26

barrier $\Delta E_{3A}^{\ddagger}$ for the addition step involving a total of three methanols (**12** → **TS7**, upper right of Figure 4) may be expressed by eq 2.

$$\Delta E_{3A}^{\ddagger} = \Delta E_{1A}^{\ddagger} - \Delta E_b^{R''} - \Delta E_d^{R''} + \Delta E_b^{T''} + \Delta E_d^{T''} \quad (2)$$

$\Delta E_b^{R''}$ and $\Delta E_b^{T''}$ are the respective binding energies (−129 and −193 kJ mol^{-1}) of two additional methanols (**2** + **2**) with species **4(12)** and **4(TS7)**, which are the distorted core fragments **4** as found in reactant complex **12** and transition structure **TS7**. $\Delta E_d^{R''}$ and $\Delta E_d^{T''}$ are the corresponding respective distortion energies (14 and 30 kJ mol^{-1}) of those core fragments relative to **4** and **TS1**. The ancillary methanol dimer fragment, which again acts as both a hydrogen-bond acceptor and a hydrogen-bond donor, also has a much stronger interaction with the distorted core in **TS7** than in **12** because of the greater dipolar character that develops as the nucleophilic addition proceeds; the dipole moment of the core fragment increases from 3.33 d in **4(12)** to 4.50 d in **4(TS7)**. The distortion energy of the core from **TS1** to **TS4** is greater than that from **4** to **4(12)**. The difference $\Delta E_b^{T''} - \Delta E_b^{R''} = -64$ kJ mol^{-1} is the contribution to catalysis from the interaction of the reacting core with two ancillary methanols serving bifunctionally as a catalytic dimer. This amount is offset by the difference $\Delta E_d^{T''} - \Delta E_d^{R''} = +16$ kJ mol^{-1}, the net distortion energy for the core to attain the necessary geometry to realize its increased interaction with the catalyst. The net catalytic effect $\Delta E_{1A}^{\ddagger} - \Delta E_{3A}^{\ddagger}$ is 48 kJ mol^{-1} (Table 2) for two ancillary methanols on the addition step of the stepwise mechanism.

In like manner, the effects of either one (lower left) or two (upper left of Figure 4) ancillary methanols on the concerted mechanism may be analyzed in terms of the potential energies of distortion of the core and of binding to it. Now $\Delta E_b^{T'}$ is the binding energy (−166 kJ mol^{-1}) of one additional methanol **2** with species **4(TS6)**, which is the distorted core **4** as found in **TS6**, and $\Delta E_d^{T'}$ is the distortion energy (113 kJ mol^{-1}) of the core relative to **TS3**. The difference $\Delta E_b^{T'} - \Delta E_b^{R'} = -95$ kJ mol^{-1} is the apparent contribution to catalysis from the interaction of the reacting core with one ancillary methanol. However, this amount is more than offset by the difference $\Delta E_d^{T'} - \Delta E_d^{R'} = +101$ kJ mol^{-1}, the net distortion energy. The difference $\Delta E_{1C}^{\ddagger} - \Delta E_{2C}^{\ddagger} = +6$ kJ mol^{-1} shows that a single ancillary methanol actually increases the barrier for the concerted mechanism.

The binding energy $\Delta E_b^{T''}$ (−395 kJ mol^{-1}) of two methanols with **4(TS9)** is clearly very large but so also is the distortion energy $\Delta E_d^{T''}$ (283 kJ mol^{-1}) of the core relative to **TS3**. The apparent contribution to catalysis is the difference $\Delta E_b^{T''} - \Delta E_b^{R''} = -266$ kJ mol^{-1}, which is almost exactly balanced by the net distortion energy $\Delta E_d^{T''} - \Delta E_d^{R''} = +268$ kJ mol^{-1}; the difference $\Delta E_{1C}^{\ddagger} - \Delta E_{3C}^{\ddagger} = +2$ kJ mol^{-1} shows that two ancillary methanols do not serve to catalyze the concerted mechanism. Two extra methanols stabilize the reacting core for the concerted mechanism much more than one extra methanol stabilizes it because the dipole moment in **4(TS9)** is much larger than that in **4(TS6)** (4.70 d vs 3.82 d). In contrast, for the addition step, two extra methanols stabilize the reacting core less well than one extra methanol because the dipole moment in **4(TS7)** is smaller than that in **4(TS4)** (4.50 d vs 4.77 d). However, the distortion energy incurred by the reacting core in order to benefit from the favorable energy of binding to one or two extra methanols is dramatically larger for the concerted mechanism than for the addition step of the stepwise mechanism.

The far right of Figure 4 shows the distortion and binding energies for the elimination step of the stepwise mechanism, which may be instructively compared with those for the concerted mechanism. It is evident from Figure 1 that **TS2** and **TS3** differ only in respect to which oxygen atom (O_c or O_{nu}) is partially bonded to H_{nu}; **TS5** and **TS6** are similarly related, and so also are **TS8** and **TS9**. The binding energy (−147 kJ mol^{-1}) of one additional methanol **2** with species **4(TS5)** is a little less than with **4(TS6)**, but the distortion energy (52 kJ mol^{-1}) to reach **4(TS5)** from **TS2** for the elimination step is markedly lower than that to reach **4(TS6)** from **TS3** for the concerted mechanism (113 kJ mol^{-1}). Similarly, while the binding energy (−249 kJ mol^{-1}) of two additional methanols with **4(TS8)** (dipole moment 1.57 d) is smaller than with **4(TS9)** (dipole moment 4.70 d), the distortion energy (107 kJ mol^{-1}) to obtain **4(TS8)** from **TS2** is much lower than the 283 kJ mol^{-1} required to obtain **4-(TS9)** from **TS3**. Consequently, the net catalytic effects $\Delta E_{1E}^{\ddagger} - \Delta E_{2E}^{\ddagger} = 36$ kJ mol^{-1} for one ancillary methanol and $\Delta E_{1E}^{\ddagger} - \Delta E_{3E}^{\ddagger} = 26$ kJ mol^{-1} for two ancillary methanols cause significant reductions in the barrier heights for the elimination step of the stepwise mechanism.

The bottom line of this analysis is that for the stepwise mechanism the binding energy of the transition state (relative to reactants) is always greater than the distortion energy of the transition state (relative to reactants) for interaction with either one or two ancillary methanols. In contrast, for the concerted mechanism, the binding energy of the transition state (relative to the reactants) is cancelled out by the distortion energy of the transition state (relative to the reactants) for interaction with either one or two ancillary methanols. Thus, the ancillary methanols generate catalysis for the stepwise mechanism but not for the concerted mechanism.

TS Geometries and Charge Changes. Just as in the previous section interactions between the reacting core and the catalyst were analyzed in terms of distortion and binding, so also may the barrier for the uncatalyzed reaction be decomposed. Thus, the 173 kJ mol^{-1} potential energy barrier for the addition step may be broken down into three

Table 3. Selected B3LYP/6-31G*-Optimized TS Bond Lengths (Å) and Bond Angles (Degrees) for Methanolysis of ϵ-Caprolactone in a Vacuum

# MeOH	species	$O_{nu}-C_c$	C_c-O_{lg}	$C_c=O_c$	$O_{nu}-H_{nu}$	O_c-H_{nu}	$O_{lg}-H_{nu}$/ $O_{lg}-H_a$[a]
	ϵ-CL **1**		1.361	1.208			
	CH_3OH **2**				0.969		
1	addition **TS1**	1.883	1.343	1.308	1.291		1.183
	elimination **TS2**	1.336	1.887	1.312		1.173	1.290
	concerted **TS3**	1.802	1.804	1.189	1.201		1.216
2	addition **TS4**	1.819	1.350	1.288	1.266		1.243
	elimination **TS5**	1.355	1.814	1.297		1.208	1.234
	concerted **TS6**	1.733	1.723	1.201	1.230		1.218
3	addition **TS7**	1.853	1.355	1.287	1.241		1.183
	elimination **TS8**	1.385	1.738	1.283		1.325	1.300
	concerted **TS9**	1.598	1.572	1.239	1.576		1.400

[a] $O_{lg}-H_{nu}$ for # MEOH = 1; $O_{lg}-H_a$ for # MeOH = 2 and 3.

components: 32 kJ mol^{-1} to dissociate the ϵ-CL and nucleophilic methanol fragments as found in the reactant complex **4**, 198 kJ mol^{-1} to distort these fragments to the geometry they have in **TS1**, and -58 kJ mol^{-1} to reassociate these fragments to form **TS1**. The distortion energy is the sum of 61 kJ mol^{-1} for ϵ-CL and 137 kJ mol^{-1} for the methanol. The primary cause for the large distortion energy of the nucleophile is the significant stretching of the $O_{nu}-H_{nu}$ bond from 0.969 to 1.291 Å (Table 3). The C_c-O_{lg} bond is shorter, and the $C_c=O_c$ bond longer, in **TS1** than in ϵ-CL, and the sum of the three angles CC_cO_{lg}, $O_{lg}C_cO_c$, and CC_cO_c around the carbonyl C_c atom decreases from 360° in the trigonal planar lactone to 353° in the partly pyramidalized TS for the addition step. In contrast, the ϵ-CL distortion required to reach **TS3** for the concerted mechanism is much larger (178 kJ mol^{-1}): this is due to the very considerable lengthening of the C_c-O_{lg} bond (from 1.361 to 1.804 Å) and a greater degree of pyramidalization (to 349°) in **TS3**. The distortion energy of the methanol in the concerted mechanism is only 86 kJ mol^{-1} since the $O_{nu}-H_{nu}$ bond is stretched much less in **TS3** than in **TS1**. The association energy between the distorted ϵ-CL and methanol fragments to form **TS3** for the concerted mechanism is more favorable (-125 kJ mol^{-1}) than for the addition step of the stepwise mechanism, because the distorted ϵ-CL involves greater charge separation in the highly stretched C_c-O_{lg} bond.

The $O_{nuc}C_cO_c$ angle in the four-membered ring of **TS1** is 86°, and the $O_{nu}H_{nu}O_c$ angle is 128° (Table 4). Correspondingly, the $O_{nuc}C_cO_{lg}$ angle is more acute (78°) in TS3, whereas the $O_{nu}H_{nu}O_{lg}$ angle is more obtuse (137°). These elements of angle strain are subsumed into the association energy of the distorted ϵ-CL and methanol fragments to form the respective TSs: less strain in these interfragment angles is reflected in a more favorable association energy.

Expansion from the four-membered **TS1** to the six-membered **TS4** for the addition step leads to an increase in the $O_{nu}C_cO_c$ angle from 86° to 101°, and the $C_c=O_c$ and $O_{nu}H_{nu}$ bonds are less stretched. These geometrical changes, which one might think would reduce strain in the TS, are nevertheless part of the *unfavorable* distortion energy from **TS1** to **4(TS4)** (Table 2). However this distortion involves not only geometrical factors but also changes in the charge distribution (Table 5). The $O_{nu}H_{nu}O_a$ and $O_aH_aO_c$ hydrogen-bond angles are relaxed from 128° to an average value of 157°, a change that should also reduce strain; this contribution is subsumed into the interaction energy (Table 2) between the distorted core and an (undistorted) ancillary methanol. Further expansion of the hydrogen-bonded cyclic system to an eight-membered ring by inclusion of the second ancillary methanol in **TS7** causes an additional reduction in $O_{nu}H_{nu}$ bond stretching and further relaxation of the $O_{nu}C_cO_c$ angle (to 106°) and the average OHO hydrogen-bond angle (to 171°).

Adding the ancillary methanols to the concerted mechanism expands the four-membered ring in **TS3** first to a six-membered ring in **TS6** and then to an eight-membered ring in **TS9**. The trends in the bond angles are very similar to those described for the addition step and, therefore, do not provide an explanation for the lack of catalysis by the additional methanol molecules. However, although there is a marked decrease in the C_c-O_{lg} bond length, there are also very significant increases in the $C_c=O_c$ and $O_{nu}H_{nu}$ bond lengths as the hydrogen-bonded connection between O_{nu} and O_c expands. The $C_c=O_c$ stretching is accompanied by an increase in electronic charge upon O_c (Table 5), even though

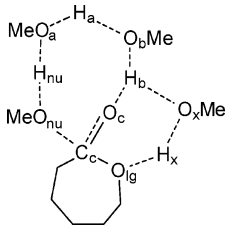
Table 4. Selected B3LYP/6-31G*-Optimized TS Bond Angles (Degrees) for the Methanolysis of ϵ-Caprolactone in a Vacuum

# MeOH	species	$O_{nu}C_cO_c$	$O_{nu}C_cO_{lg}$	$O_cC_cO_{lg}$	<OHO>	ΣC_c
	ϵ-CL **(1)**			118		360
1	addition **TS1**	86	110	114	128	353
	elimination **TS2**	120	106	85	126	
	concerted **TS3**	115	78	111	137	349
2	addition **TS4**	102	102	114	157	348
	elimination **TS5**	117	102	101	158	
	concerted **TS6**	113	93	111	156	344
3	addition **TS7**	106	101	111	171	349
	elimination **TS8**	119	103	104	168	
	concerted **TS9**	112	97	110	170	338

Table 5. Selected B3LYP/6-31G* Natural Population Analysis (NBO) TS Atomic Charges for the Methanolysis of ϵ-Caprolactone in a Vacuum

# MeOH	species	O_{nu}	O_{lg}	O_c	C_c	H_{nu}	H_{lg}	H_c
				reactive complex				
1	**4**	−0.644	−0.431	−0.470	0.611	0.429		
2	**8**	−0.675	−0.457	−0.539	0.641	0.434	0.438	
3	**12**	−0.675	−0.452	−0.507	0.649	0.443	0.446	
				addition				
1	**TS1**	−0.720	−0.542	−0.784	0.886	0.527		
2	**TS4**	−0.706	−0.587	−0.767	0.862	0.521		0.527
3	**TS7**	−−0.722	−0.556	−0.733	0.881	0.527		0.521
				concerted				
1	**TS3**	−0.706	−0.716	−0.563	0.805	0.541		
2	**TS6**	−0.697	−0.702	−0.622	0.829	0.536	0.530	
3	**TS9**	−0.680	−0.687	−0.765	0.845	0.531	0.546	
				elimination				
1	**TS2**	−0.437	−0.648	−0.653	0.712		0.457	
2	**TS5**	−0.457	−0.657	−0.673	0.739		0.475	0.480
3	**TS8**	−0.469	−0.658	−0.686	0.766		0.487	0.480

Chart 2



this degree of freedom does not contribute to the reaction coordinate. The extra methanol molecules are unable to stabilize this negative charge buildup on O_c since they are hydrogen-bonded to the "wrong" atom, O_{nu}.

In contrast, ring expansion in **TS2**, **TS5**, and **TS8** for the elimination step leads to decreases in the C_c-O_{lg} and $C_c=O_c$ bond lengths and increases in the $O_cC_cO_{lg}$ and average OHO angles, with consequent reductions in strain. There is a slight increase in negative charge on O_c, but this is directly involved in the hydrogen-bonding interactions linking this atom with O_{lg}.

One might speculate as to the possible catalytic effect of a third ancillary methanol. In the light of the results discussed above for one and two methanols, to extend the size of a hydrogen-bonded ring system further from 6 to 8 to 10 would probably not be the most effective strategy. Instead, it might be preferable to deploy the extra methanol in a second hydrogen-bonded ring, as shown in Chart 2. This arrangement would ensure that any negative charge increase on O_c would be stabilized by bonding interactions with H_b, either covalently in the intermediate of a stepwise mechanism or noncovalently in a concerted mechanism.

Orbital Interactions. Nucleophilic addition to the carbonyl group of ϵ-CL requires overlap between a lone-pair orbital n_O on O_{nu} and the antibonding $\pi^*(C_c=O_c)$ orbital, with a consequent preference for approach along the Bürgi−Dunitz trajectory, roughly perpendicular to the plane of the lactone, and ideally close to the tetrahedral angle for $O_{nu}C_cO_c$. Expansion of the cyclic TS by the inclusion of two ancillary methanol molecules allows this geometry to be adopted, with $O_{nu}C_cO_c$ increasing from 86° in **TS1** to 106° in **TS7**, while keeping O_{nu} away from O_{lg} at a respectable angle of 101°. On the other hand, nucleophilic substitution at C_c requires overlap between n_O on O_{nu} and the antibonding $\sigma^*(C_c-O_{lg})$ orbital, with a strong preference for a collinear alignment of O_{nu} with the C_c-O_{lg} bond. In practice, steric constraints and the necessity of transferring a proton from O_{nu} to O_{lg} together do not permit the $O_{nu}C_cO_{lg}$ angle to adopt a value anyway close to 180°: the best that can be achieved is 78° in **TS3**, increasing to 97° in **TS9**. In a planar lactone, the $\pi^*(C_c=O_c)$ and $\sigma^*(C_c-O_{lg})$ orbitals are orthogonal, but any distortion away from planarity allows the two orbitals to mix. Thus, any development of overlap between n_O and $\sigma^*(C_c-O_{lg})$ is inevitably accompanied by overlap between n_O and $\pi^*(C_c=O_c)$ with consequent lengthening of the $C_c=O_c$ bond and an accumulation of electron density on O_c. This might possibly be avoided if the nucleophilic attack could be decoupled from the proton transfer, but this would require a different type of catalyst.

Conclusions

Stepwise addition/elimination and concerted mechanisms for methanolysis of ϵ-caprolactone have been investigated computationally as a model for ROP, with assistance from one or two ancillary methanol molecules. The effects of specific solvation by these extra methanols in cyclic hydrogen-bonded clusters are very significant, whereas the effects of bulk solvation as treated by the PCM method are almost negligible. Increasing the ring size lowers the barriers for both the addition and elimination steps of the stepwise mechanism but does not do so for the concerted mechanism; a stepwise mechanism is therefore preferred for methanol-assisted ROP. This is because for the stepwise mechanism the intrinsic binding-energy of the reacting core with the catalyst in the transition state (relative to reactants) is always greater than the distortion energy of those core fragments in the transition state (relative to reactants) for interaction with either one or two ancillary methanols. In contrast, for the concerted mechanism, the intrinsic binding-energy term is cancelled out by the distortion energy term for interaction with either one or two ancillary methanols. The distortion energies are large in the transition states for the concerted mechanism because the geometry required in order for nucleophilic addition to occur simultaneously with proton transfer to the leaving group forces electronic charge onto the carbonyl oxygen, even though this is not part of the reaction coordinate. Thus, the ancillary methanols generate catalysis for the stepwise mechanism but not for the concerted mechanism. The essential role of a catalyst is to avoid unfavorable accumulation or separation of charges. The key requirement for catalyst design is to avoid an unfavorable buildup of charge. It may be noted that the oxyanion hole of serine proteases fulfills this requirement for the enzymic catalysis of biological acyl transfer reactions.

Acknowledgment. We are grateful to the EPSRC and Johnson Matthey for a CASE studentship to N.B. and to the EPSRC National Service for Computational Chemistry Software (http://www.nsccs.ac.uk) for the provision of computer resources.

Supporting Information Available: Coordinates, total energies, zero-point energies for all optimized structures, transition frequencies for transition structures, and full Gaussian citation. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) (a) Langer, R.; Vacanti, J. P. *Science* **1993**, *260*, 920. (b) Albertsson, A.-C.; Varma, I. K. *Biomacromolecules* **2003**, *4*, 1466−1486.

(2) (a) Dubois, P.; Ropsen, N.; Jérôme, R.; Teyssié, P. *Macromolecules* **1996**, *27*, 1965−1975. (b) Ovitt, T. M.; Coates, G. W. *J. Am. Chem. Soc.* **1999**, *121*, 4072−4082. (c) O'Keefe, B. J.; Hillmyer, M. A.; Tolman, W. B. *J. Chem. Soc., Dalton Trans.* **2001**, *15*, 2215−2224.

(3) Penczek, S.; Duda, A.; Libiszowski, J. *Macromol. Symp.* **1998**, *128*, 241−254.

(4) Lofgren, A.; Albertsson, A.-C.; Dubois, P.; Jerome, R.; Teyssie, P. *Macromolecules* **1994**, *27*, 5556−5562.

(5) Ropson, N.; Dubois, P.; Jerome, R.; Teyssie, P. *Macromolecules* **1995**, *28*, 7589−7598.

(6) Eguiburu, J. L.; Fernandez-Berridi, M. J.; Cossio, F. P.; San Roman, J. *Macromolecules* **1999**, *32*, 8252−8258.

(7) Dechy-Cabaret, O.; Martin-Vaca, B.; Bourissou, D. *Chem. Rev.* **2004**, *104* (12), 6147−6176.

(8) Ryner, M.; Stridsberg, K.; Albertsson, A.-C.; von Schenck, H.; Svensson, M. *Macromolecules* **2001**, *34*, 3877−3881.

(9) Musaev, D. G.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 6509−6517.

(10) For example: (a) Williams, I. H.; Spangler, D.; Femec, D. A.; Maggiora, G. M.; Schowen, R. L. *J. Am. Chem. Soc.* **1980**, *102*, 6619−6621. (b) Williams, I. H.; Maggiora, G. M.; Schowen, R. L. *J. Am. Chem. Soc.* **1980**, *102*, 7831−7839. (c) Williams, I. H.; Spangler, D.; Femec, D. A.; Maggiora, G. M.; Schowen, R. L. *J. Am. Chem. Soc.* **1983**, *105*, 31−40. (d) Williams, I. H. *J. Am. Chem. Soc.* **1987**, *109*, 6299−6307.

(11) Kallies, B.; Mitzner, R. *J. Mol. Model.* **1998**, *4*, 183−196.

(12) (a) Schmeer, G.; Sturm, P. *Phys. Chem. Chem. Phys.* **1999**, *1*, 1025−1030. (b) Yamabe, S.; Tsuchida, N.; Hayashida, Y. *J. Phys. Chem. A* **2005**, *109*, 7216−7224.

(13) (a) Venkatasubban, K. S.; Bush, M.; Ross, E.; Schultz, M.; Garza, O. *J. Org. Chem.* **1998**, *63*, 6115−6118. (b) Frasson, C. M. L.; Brandão, T. A. S.; Zucco, C.; Nome, F. *J. Phys. Org. Chem.* **2006**, *19*, 143−147.

(14) (a) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785−789. (b) Becke, A. D. *J. Phys. Chem.* **1993**, *98*, 5648−5652.

(15) Sinclair, P. E.; Catlow, R. A. *J. Phys. Chem. B* **1999**, *103*, 1084−1095.

(16) Guest, M. F.; Van Lenthe, J. H.; Schoffel, K.; Sherwood, P.; Harrison, R. J.; Amos, R. D.; Buenker, R. J.; Dupuis, M.; Handy, N. C.; Hillier, I. H.; Knowles, P. J.; Bonacic-Koutecky, V.; Von Niessen, W.; Saunders, V. R.; Stone, A. J. *GAMESS-UK (1995−1997)*; Daresbury Laboratories: Warrington, U. K., 1997. The package is derived from the original GAMESS code due to M. Dupuis, D. Spangler, and J. J. Wendoloski.

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

(18) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.* **1992**, *70*, 560−571.

(19) (a) Cerjan, C. J.; Miller, W. H. *J. Chem. Phys.* **1981**, *75*, 2800−2806. (b) Simons, J.; Jørgensen, P.; Taylor, H.; Ozment, J. *J. Phys. Chem.* **1983**, *87*, 2745−2753. (c) Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. *J. Phys. Chem.* **1985**, *89*, 52−57.

(20) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735−746.

(21) Miertus, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117−129.

(22) Williams, I. H.; Spangler, D.; Maggiora, G. M.; Schowen, R. L. *J. Am. Chem. Soc.* **1985**, *107*, 7717−7723.

(23) Williams, I. H. *L'actualité chimique* **1991**, 27−31.

(24) Schowen, R. L. In *Transition States of Biochemical Processes*; Gandour, R. D., Schowen, R. L., Eds.; Plenum Press: New York, 1978.

(25) Williams, I. H. *Bull. Soc. Chim. Fr.* **1988**, 192−198.

(26) (a) Henderson, R. *J. Mol. Biol.* **1970**, *54*, 341−354. (b) Hedstrom, L. *Chem. Rev.* **2002**, *102*, 4501−4523.

CT600265C

JCTC Journal of Chemical Theory and Computation

Generalized Born Model with a Simple, Robust Molecular Volume Correction

John Mongan,[†,‡,§] Carlos Simmerling,[#] J. Andrew McCammon,[†,§,||,⊥,▽] David A. Case,[§,★] and Alexey Onufriev*[,◆]

Bioinformatics Program, Medical Scientist Training Program, Center for Theoretical Biological Physics, Department of Chemistry and Biochemistry, and Department of Pharmacology, UC San Diego, La Jolla, California 92093-0365, Department of Chemistry and Center for Structural Biology, Stony Brook University, Stony Brook, New York 11794-3400, Howard Hughes Medical Institute, Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037, and Departments of Computer Science and Physics, Virginia Tech, Blacksburg, Virginia 24060

Received March 3, 2006

Abstract: Generalized Born (GB) models provide a computationally efficient means of representing the electrostatic effects of solvent and are widely used, especially in molecular dynamics (MD). A class of particularly fast GB models is based on integration over an interior volume approximated as a pairwise union of atom spheres—effectively, the interior is defined by a van der Waals rather than Lee-Richards molecular surface. The approximation is computationally efficient but, if uncorrected, allows for high dielectric (water) regions smaller than a water molecule between atoms, leading to decreased accuracy. Here, an earlier pairwise GB model is extended by a simple analytic correction term that largely alleviates the problem by correctly describing the solvent-excluded volume of each pair of atoms. The correction term introduces a free energy barrier to the separation of nonbonded atoms. This free energy barrier is seen in explicit solvent and Lee-Richards molecular surface implicit solvent calculations but has been absent from earlier pairwise GB models. When used in MD, the correction term yields protein hydrogen bond length distributions and polypeptide conformational ensembles that are in better agreement with explicit solvent results than earlier pairwise models. The robustness and simplicity of the correction preserves the efficiency of the pairwise GB models while making them a better approximation to reality.

1. Introduction

The effects of aqueous solvent are critical to the structure and function of biological macromolecules. Commonly, solvent is represented explicitly, by models of multiple water molecules, or implicitly, by a high dielectric region and additional apolar solvation terms. Although explicit solvent is a more physically rigorous representation, implicit solvent models have the advantage of dramatically reducing the degrees of freedom that must be sampled by eliminating those associated with the solvent. Additionally, implicit solvent models are often more computationally efficient than their explicit counterparts.

The solvation effects can be described by ΔG_{solv}: the free energy of transferring a given configuration of a molecule from vacuum to solvent. To facilitate calculation of ΔG_{solv}, it is typically decomposed into polar and nonpolar compo-

* Corresponding author e-mail: alexey@cs.vt.edu.
† Bioinformatics Program, UC San Diego.
‡ Medical Scientist Training Program, UC San Diego.
§ Center for Theoretical Biological Physics, UC San Diego.
|| Department of Chemistry and Biochemistry, UC San Diego.
⊥ Department of Pharmacology, UC San Diego.
Stony Brook University.
▽ Howard Hughes Medical Institute.
★ The Scripps Research Institute.
◆ Virginia Tech.

10.1021/ct600085e CCC: $37.00
Published on Web 12/08/2006

nents: $\Delta G_{solv} = \Delta G_{pol} + \Delta G_{nonpol}$. Here, ΔG_{nonpol} is the free energy of introducing the solute molecule into solvent while electrostatic interactions between the solute and solvent are turned off, and ΔG_{pol} is the free energy change in the system resulting from turning these electrostatic interactions back on. In this work, the focus is on methods for calculating ΔG_{pol}.

Assuming that the solvent can be faithfully represented by a continuum dielectric region, the Poisson−Boltzmann (PB) equation is the most physically correct method of determining ΔG_{pol} and has been widely used over the past decade.[1−7] Application of PB to molecular geometries requires numerical solution of second-order partial differential equations, which is fairly computationally intensive and does not easily yield forces, although recent advances in PB methodology have improved the situation somewhat.[1,8−10] Alternatively, generalized Born (GB) models have become popular as a computationally efficient approximation to numerical solutions of the PB equation,[6,11−23] especially for use in dynamics.[24−34]

GB models evaluate polar solvation free energy as a sum of pairwise interaction terms between atomic charges. When the solute dielectric is 1 and the solvent dielectric is much greater than that of the solute,[35] the interactions can be accurately described by an analytical function first proposed by Still et al.[12] that interpolates between the Coulombic limit at long distances and the Born or Onsager limits at small distances

$$\Delta G_{pol} \approx \Delta G_{GB} = -\frac{1}{2}\sum_{i,j}\frac{q_i q_j}{\sqrt{r_{ij}^2 + R_i R_j \exp\left(\frac{-r_{ij}^2}{4R_i R_j}\right)}}\left(1 - \frac{1}{\epsilon_w}\right) \quad (1)$$

where r_{ij} is the distance between atoms i and j, q_i and q_j are partial charges, and ϵ_w is the dielectric constant of the solvent. The key parameters in this GB function are the effective radii of the interacting atoms, R_i and R_j, which represent each atom's degree of burial within the solute. More specifically, the effective radius of an atom is defined as the radius of a corresponding spherical ion having the same ΔG_{pol} as the self-energy of this atom in the molecule. The self-energy is the polar solvation free energy for the molecule with partial charges set to zero for all atoms except the atom of interest. The effective radius of an atom is larger than the intrinsic radius of its atom sphere because of the descreening effects of surrounding atoms, reducing the extent to which the atom charge is screened by solvent. A computationally inefficient but theoretically interesting method for determining effective radii is to derive them from self-energies calculated using well-converged numerical PB solutions. When these "perfect" effective radii are used, GB results are in close agreement with PB results,[36] which serve as a natural point of reference for assessing the accuracy of GB, since current GB models are an approximation to the more fundamental formalism of the PB equation. Although this form of GB is impractical for application, it suggests that in aqueous solution the GB function introduced by Still et al. has so far been a minor source of error compared with the error introduced by nonperfect methods for estimating effective radii. Consequently, considerable effort has been spent on improving the way effective Born radii are computed.

In practice, effective radii for each atom are generally calculated by integration of an approximate electric field density due to the atom of interest over some definition of the molecule's volume,[5,13,14,21,32,38] although formulations based on surface integrals have also been proposed.[19,23] Here, we focus on volume-based GB models which have traditionally used a Coulomb field integral

$$I_i = \frac{1}{4\pi}\int_{\Omega_i} \mathbf{r}^{-4} d^3\mathbf{r} \quad (2)$$

where the origin is centered on atom i, and Ω_i represents the volume inside the molecule but outside atom i. The effective radius is then calculated according to

$$R_i = (\rho_i^{-1} - I_i)^{-1} \quad (3)$$

where ρ_i is the intrinsic radius of atom i. Within the Coulomb field approximation (CFA) embodied by the integral in eq 2, it is assumed that the electric field generated by an atomic point charge is unaffected by the nonhomogeneous dielectric environment created by the solute, so that the field has the form described by Coulomb's law. The CFA is exact for a point charge at the center of a spherical solute, but it overestimates effective radii for molecular geometries[21] as well as for spherical regions when the charge is off center.[39] Some of the success of early GB models on small molecules may be attributed to fortuitous cancellation of errors in effective radius calculations between the overestimates of a CFA based integrand and the underestimates of a van der Waals (VDW) based region of integration.[32] Improved approximations based on empirical corrections to the CFA[21,38,40] or theoretical derivations originating with the Kirkwood formula[39,41] have significantly better agreement with effective radii calculated from PB self-energies.

The integration in eq 2 can be performed numerically[12,19,21,38,40] or by an analytical pairwise approximation.[13−15,32,33,37] GB methods based on analytically approximated integrals are easily extended to calculate solvation forces and are generally faster than their numerically integrated counterparts,[42] so they have traditionally found greater application in dynamics.

Most pairwise approximations estimate the integral over a region formed by the union of atom spheres, which is equivalent to a VDW surface dielectric boundary. In calculating the effective radius for atom i, the contribution of every other atom $j \neq i$ to the integral is determined as a function of ρ_j and the distance between atoms i and j. Summation of these terms yields an overestimate of the total integral, due to overlap between descreening atoms. To correct for these overlaps, a multiplicative scaling factor, S_x, is introduced to reduce the intrinsic radius of each descreening atom.

In contrast, PB calculations generally use a Lee-Richards molecular surface dielectric boundary, defined by rolling a solvent sphere over the surface of the molecule.[43] Although

there is no uniquely correct definition of the dielectric boundary, a van der Waals surface creates regions of interstitial high dielectrics that are smaller than a water molecule, while the molecular surface has the conceptually attractive advantage of excluding high dielectric from regions into which a water molecule is too large to fit. Differences between the molecular and VDW surface definitions are minimal for small molecules, where all atoms are well solvated, but become more substantial for macromolecules, where inclusion of interstitial high dielectrics in VDW-based models leads to overestimation of the solvation of interior atoms, relative to molecular surface results.[44] This may partially explain why early GB models that had good results for small molecules were less effective when applied to macromolecules.[26,32,37] Additionally, implicit solvent models that allow interstitial high dielectrics produce incorrect potentials of mean force between nonbonded atoms.[44] However, it may not be practical to use the Lee-Richards molecular surface directly in a GB model as it is fairly computationally intensive and can produce unstable or infinite forces for some molecular configurations.[1,9]

Attempts to reduce or eliminate the problems of interstitial high dielectrics in GB models have followed two paths. One approach, embodied by the GBMV2 model developed by Lee et al.,[38] has been to use numerical integration with adaptations for calculating forces in combination with an analytic surface definition that closely approximates the properties of a molecular surface. A CFA correction term is also employed in the integration. This GB model yields stable dynamics while providing excellent agreement with PB Lee-Richards molecular surface results. However, both the analytic surface definition and the numerical integration are relatively slow, such that the fastest PB models approach the performance of GBMV2.[42] Furthermore, the reliance on numerical integration introduces artifacts, such as a lack of rotational invariance.

A different method (*OBC* GB), developed by Onufriev, Bashford, and Case,[32] sought to extend the pairwise integration method (*HCT* GB) of Hawkins, Cramer, and Truhlar[13,14] to reduce the effect of interstitial high dielectrics. Based on the observation that effective radii for buried atoms are larger than for surface atoms, but still much smaller than PB-derived "perfect" effective radii, this method modifies the radius calculation in eq 3 by rescaling the integral from eq 2 according to

$$R_i = (\tilde{\rho}_i^{-1} - \rho_i^{-1}\tanh(\alpha I\tilde{\rho}_i - \beta(I\tilde{\rho}_i)^2 + \gamma(I\tilde{\rho}_i)^3))^{-1} \quad (4)$$

where $\tilde{\rho}_i = \rho_i - 0.09$ Å and α, β, and γ are tunable parameters. When these parameters are set such that most radii are scaled up, the rescaled radii substantially improve agreement with PB solvation free energies, and the computational expense of the rescaling function is minimal so the efficiency of the Hawkins et al. model is retained. In addition, effective radii calculated with eq 4 are smoothly capped at about 30 Å, avoiding problems with numerical instability and negative radii that can be encountered when using eq 3. However, by design, the rescaling function only affects atoms that are sufficiently buried that the interstitial high dielectrics can be accounted for in an averaged, geometry-independent manner. Uncompensated interstitial high dielectrics between more highly solvated surface atoms still affect solvation energies and potentials of mean force.

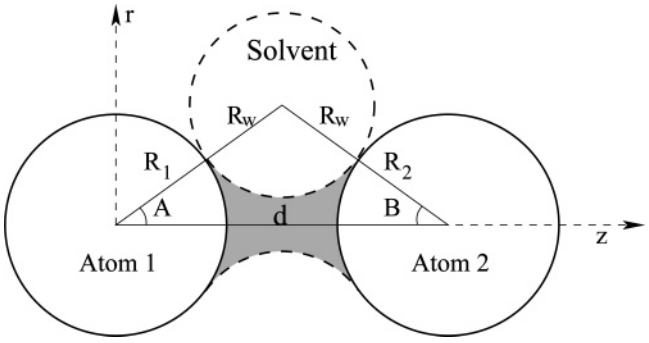


Figure 1. The neck region (shaded) is defined by the radius of atom 1, R_1, the radius of atom 2, R_2, the distance that separates them, d, and the radius of the solvent molecule, R_w. The coordinate system used for performing integration is also illustrated (see the Appendix).

In this paper, we attempt to combine the best aspects of both of these efforts in development of a GB model that adds a geometrically based molecular volume correction term accounting for interstitial high dielectrics to the pairwise approximated integration method. Since the correction term is, itself, a computationally efficient pairwise approximation, the performance and numerical benefits of analytical GB models are retained.

The shortcomings of the CFA are now well-known, but rigorously derived non-CFA pairwise approximated GB models have only recently been described[41] and their stability and performance have not yet been extensively tested on biomolecules, so the model described here extends the Coulomb field-based *HCT* GB model.

2. Theory

An ideal volume correction term for a GB model based on VDW volume and the CFA would yield the integral of $\mathbf{r}^{-4}$ over the region inside the Lee-Richards molecular surface and outside the van der Waals surface. This region is designated the correction region.

$$\int_{LR}\mathbf{r}^{-4}d^3\mathbf{r} = \int_{VDW}\mathbf{r}^{-4}d^3\mathbf{r} + \int_{correction}\mathbf{r}^{-4}d^3\mathbf{r} \quad (5)$$

Since the *HCT* GB integration scheme calculates the value of the integral within the van der Waals surface, adding this correction term would yield an integral over the region within the molecular surface. In the general case, the correction region cannot be analytically defined. However, in the simple case of two closely spaced or overlapping atoms, the correction region forms an analytically definable "neck" region between the two atoms, as seen in Figure 1. The general case of the correction region can be approximated by a union of these neck regions calculated pairwise between atoms. In the simplest form of this approximation, developed here, the integral for each atom includes corrections for only the neck regions in which the atom is directly involved. This simple form is a reasonably good approximation because the value of the integrand ($\mathbf{r}^{-4}$) is much higher in the nearby

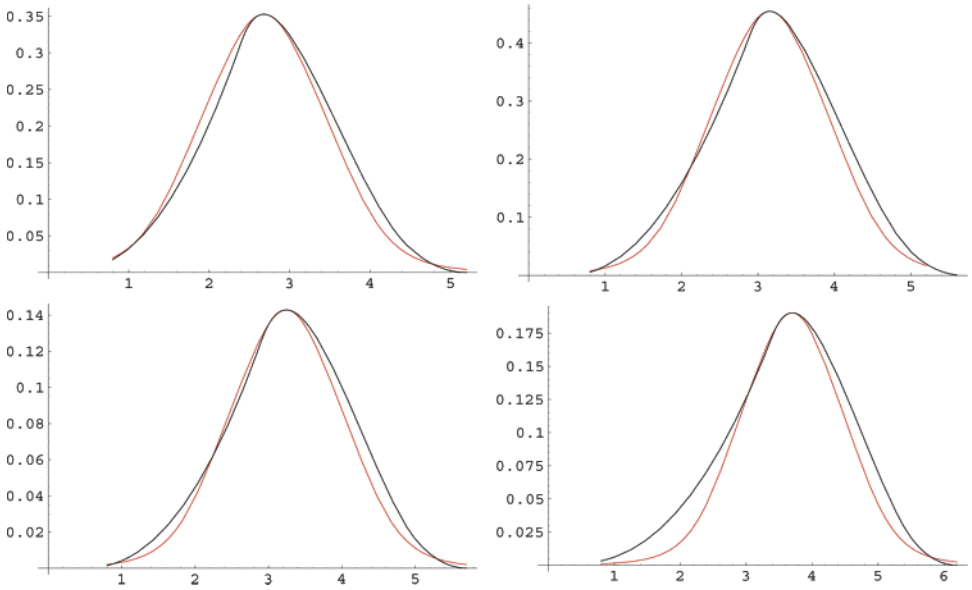


Figure 2. Values of numerical integration over the neck region (black) and analytical approximation (red) as a function of distance between atoms in angstroms. Left to right, top to bottom, radii (in angstroms) for atoms 1 and 2, respectively are 1.2 and 1.2; 1.2 and 1.7; 1.7 and 1.2; and 1.7 and 1.7.

neck regions with which the atom is directly involved than in the distant portions of the correction region formed by interactions between other pairs of atoms.

Figure 1 illustrates how the geometry of the neck region is defined by four parameters: the radii of the two atoms, R_1 and R_2; the radius of the solvent molecule, R_w; and the distance between the two atoms, d. Derivations of the expressions for the CFA integrals over the neck region are given in the Appendix. Although the integrands in these expressions are fairly simple, the limits of integration are sufficiently complex to make analytical solution of the integrals impractical. The problem is simplified by considering that in the GB model, parameters R_1, R_2, and R_w have a relatively small set of discrete values (a single value, in the case of $R_w = 1.4$ Å), and so d is the only parameter with continuous values. With this view in mind, the function in four variables described by these integrals can be evaluated as a family of single variable functions of d, with each function determined by a particular set of values for R_1, R_2, and R_w. These functions of d can be plotted by solving the integrals numerically for a range of values of d, producing curves as shown in Figure 2.

Numerical solution of these integrals is far too computationally costly for application in a GB model. Instead, they are replaced with an empirically determined analytic function shown in eq 6

$$neck\_integral(d) = \frac{m_0}{1 + (d - d_0)^2 + 0.3(d - d_0)^6} \quad (6)$$

This function is parametrized by the position (d_0) and value (m_0) of the maximum, which are determined by numeric optimization (maximization) of the integral of $\mathbf{r}^{-4}$ over the neck region of Figure 1. The values of d_0 and m_0 are dependent on R_1, R_2, and R_w, but since these variables have a small set of discrete values, tabulating all possible values of d_0 and m_0 is quite feasible (see Tables 3 and 4). As illustrated in Figure 2, eq 6 is a very good approximation over the range of atomic radii typically encountered in biomolecules.

Applications of GB solvation models to dynamics require calculation of derivatives with respect to distance. Equation 6 is easily differentiated, yielding eq 7.

$$neck\_integral'(d) = -\frac{\left(2(d - d_0) + \frac{9}{5}(d - d_0)^5\right)m_0}{\left(1 + (d - d_0)^2 + \frac{3}{10}(d - d_0)^6\right)^2} \quad (7)$$

Ideally, neck integrals would be calculated only between atoms that are close enough to define a neck region ($d < R_1 + R_2 + 2R_w$): beyond this distance the neck integral and its first derivative with respect to d should be zero. However, the analytic approximation used here approaches zero asymptotically, and at $d = R_1 + R_2 + 2R_w$ its value is on the order of 10^{-3}. Truncating the function at this point would create a discontinuity which could lead to unstable dynamics. A variety of techniques could be employed to smooth this discontinuity; we have taken the simplest approach of continuing to calculate the neck correction for $d > R_1 + R_2 + 2R_w$ until d is large enough that the value of the function is sufficiently small that the error of truncating it is on the order of rounding error.

The neck correction described by the integrals in the Appendix and approximated by eq 6 is exact for a system of two atoms, but in the usual case of a molecule with more than two atoms, a strict summation of neck integrals calculated pairwise between atoms will tend to overestimate the integral over the correction region. Overestimation of the integral is due to overlap of neck regions with atoms not participating in the neck as well as overlap with other neck regions and must be corrected by scaling the contributions to the total integral.

The *GBn* model ("n" for neck) presented here takes a simple, two-step approach to scaling. First, each neck integral

value calculated in eq 6 is multiplied by a scaling factor S_{neck} ($S_{neck} < 1$). Second, effective radii are calculated using eq 4 which provides descreening dependent scaling as well as numerical stabilization for large effective radii. The descreening dependent scaling of the second step helps to compensate for different molecular geometries, as the effective radii of more deeply buried (more descreened) atoms, which are involved in more necks and thus have more overlaps, can be scaled down to a greater degree than less buried atoms. It appeared likely that a more complex scaling procedure, such as one that employed multiple atom-type dependent values of S_{neck}, would yield a somewhat more accurate estimate of the molecular volume. However, our tests of such scaling procedures yielded insufficient improvements to justify addition of more free parameters to the model (results not shown); we believe this is because the quality of the current model is most severely constrained by the limitations of the CFA, rather than the simple scaling process described above.

The two-step scaling involves four parameters which must be optimized, S_{neck}, α, β, and γ. Since the neck correction alone is expected to bring the integration volume closer to molecular volume, the optimal parameters of eq 4 are different from those used by the *OBC* model. The key difference between *GBn* and *OBC* GB can be best illustrated by a diatomic system such as that in Figure 1: the *OBC* model will produce correct effective radii for only one value of atom−atom separation distance (hence its "geometry-independence"), while the *GBn* model should calculate accurate radii for this simple system across the entire range of interatomic distances.

Additionally, it is necessary to refit the intrinsic radius scaling factors, S_x. Although formally the S_x scaling factors merely correct overlaps, in practice they have been used as free parameters to optimize GB results for agreement with PB and experimental results.[14,26] As a result, the sets of S_x values used in the *HCT* and *OBC* GB models not only correct for atomic overlaps but also correct for some of the effects of the CFA and interstitial high dielectrics (to the extent that this is possible on an averaged, geometry independent basis). Since the *GBn* model already accounts for interstitial high dielectrics with the neck term and has a different degree of CFA error due to the altered region of integration, it would clearly be inappropriate to use S_x sets that were fit for VDW regions of integration with the *GBn* model.

3. Results and Discussion

The ultimate goal of an implicit solvent model is to accurately and efficiently approximate the results of computationally expensive explicit solvent molecular simulations. While agreement with explicit solvent provides a rigorous test of an implicit model, it is often difficult to identify the source of discrepancies due to the dramatic differences between the explicit and implicit solvent formalisms. Therefore the performance of *GBn* is compared to the earlier *OBC* GB[32] model using multiple levels of less approximate solvation models as standards. Comparison to PB is instructive in identifying the source of shortcomings of the current model, while comparison to explicit solvent provides a more useful assessment of the ultimate quality of the model. The *OBC* GB model is selected as a reference for comparison because it is among the most recent and most accurate[42] pairwise GB models that do not have a molecular volume correction beyond the "average" rescaling provided by eq 4.

Table 1. Optimized Scaling Parameters

parameter	value	parameter	value
α	1.095	S_H	1.091
β	1.908	S_C	0.484
γ	2.508	S_N	0.700
S_{neck}	0.362	S_O	1.066

Once the parameters have been optimized, the *GBn* model achieves approximately a 25% improvement over *OBC* GB in accuracy of effective radii relative to PB results. The minimal native-state bias and stable dynamics achieved by *OBC* GB are maintained in this model. A major qualitative improvement of *GBn* is that it reproduces the free energy barrier to separation of hydrogen bonds that is seen in molecular surface PB results but absent in nonmolecular surface implicit solvent models. Although quantitative improvement in agreement of *GBn* solvation energies with PB solvation energies is fairly small, substantial improvements are seen in agreement between *GBn* and TIP3P explicit solvent dynamics. Specifically, hydrogen bond length distributions are significantly more similar, there is improved agreement with the TIP3P ϕ/ψ potential, and a dramatic improvement in the conformational ensemble of deca-alanine. The results that have been summarized here are examined in detail in the following.

3.1. Parametrization. Parameters of the *GBn* model (S_{neck}, α, β, γ, and the S_x parameters for atom types C, H, N, and O) were optimized using the Nelder-Mead simplex algorithm.[45] The objective function that was minimized measured agreement between PB and GB solvation free energies over a training set consisting of structures from denaturation trajectories of apo myoglobin and protein L and structures representing potentials of mean force (PMF) for two hydrogen bonds and a salt bridge (see Methods for details of the objective function). The objective function has multiple local minima, so 100 minimizations were performed starting from random initial points. Optimized parameter values producing the best overall performance are given in Table 1. Treatment of the S_x values as free parameters to optimize GB performance beyond their formal purpose of correcting overlap is made obvious by the values of S_O and S_H, which exceed 1. This represents a continuation of previous practice, although it may at first appear to be a divergence because previous sets of S_x values where all $S_x < 1$ may have been incorrectly interpreted as merely correcting overlaps.

3.2. Comparison with PB Effective Born Radii and Solvation Energies. A common first test of a GB model is comparison of effective radii with "perfect" radii derived from PB calculations. Agreement at this level is generally correlated with the overall quality of a GB model. While use of perfect radii in the GB formalism has been shown to guarantee a very good agreement between the GB and PB solvation free energies,[36] small improvements in agreement

Table 2. RMS Deviation (in Units of Inverse Å) between Inverse Effective Radii, Computed by the *GBn* and *OBC* GB Models Relative to the PB Reference with Significance of Improvement Measured by *F*-Test[a]

	thioredoxin	apomyoglobin-I	apomyoglobin-II	β-hairpin
OBC GB	0.128	0.067	0.046	0.055
GBn	0.092	0.050	0.033	0.045
p-value	10^{-40}	10^{-47}	10^{-60}	10^{-3}

[a] The β-hairpin and thioredoxin structures are in their native states, while apomyoglobin is represented by two partially unfolded states along an acid denaturation trajectory.[46] Both models perform more poorly on thioredoxin than other structures due to a higher number of large effective radius atoms in thioredoxin. The much larger *p*-value for β-hairpin is due to the small number of atoms in the molecule, resulting in fewer degrees of freedom in the *F*-test.

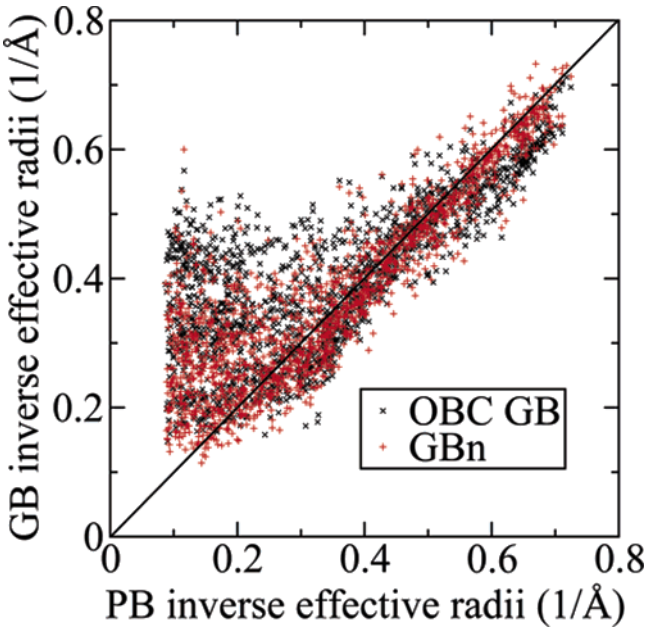


Figure 3. Scatter plot comparison of inverse effective radii calculated by the current GB neck model (red +) and earlier OBC GB model (black X) to inverse "perfect" PB radii for thioredoxin (PDB code 2TRX). Diagonal line indicates perfect agreement.

of the effective radii may or may not translate into significantly improved overall performance of the GB model. Nevertheless, radius comparisons are instructive as rough quality measures and in identifying sources of error that may not be readily apparent when molecular solvation free energies are compared. It is most useful to compare inverse radii, as these more faithfully represent the contribution of the effective radii to the energy in eq 1. Such a comparison for a set of four structures that includes native and partially unfolded proteins and peptides is presented in Table 2. For all structures, the *F*-test shows a highly significant improvement in the accuracy of effective radii calculated by the *GBn* model compared to the *OBC* GB model.

A more detailed analysis of effective radii is illustrated in Figure 3, showing improvement across the whole range of effective radii. This includes an improvement in the accuracy of large effective radii (left portion of the figure), although these radii continue to have the largest errors. Errors seem to be largest for atoms near crevices that are slightly too small for a water molecule; presumably the pairwise approximation is poorest here.

Table 3. Distance between Atoms at Which Integral of $\mathbf{r}^{-4}$ over the Neck Region (Defined in Eqs 13−15) Has the Maximum Value, Tabulated for a Range of Radii for Atoms 1 and 2, Assuming a Solvent Molecule (R_w) Radius of 1.4 Å[a]

	atom 2						
atom 1	1.20	1.25	1.30	1.35	1.40	1.45	1.50
1.20	2.6797	2.7250	2.7719	2.8188	2.8656	2.9125	2.9609
1.25	2.7359	2.7813	2.8281	2.8750	2.9219	2.9688	3.0156
1.30	2.7922	2.8375	2.8844	2.9297	2.9766	3.0234	3.0719
1.35	2.8500	2.8953	2.9406	2.9859	3.0328	3.0797	3.1266
1.40	2.9062	2.9516	2.9969	3.0422	3.0891	3.1359	3.1828
1.45	2.9625	3.0078	3.0531	3.0984	3.1437	3.1906	3.2375
1.50	3.0188	3.0641	3.1078	3.1547	3.2000	3.2469	3.2922
1.55	3.0750	3.1203	3.1641	3.2094	3.2563	3.3016	3.3484
1.60	3.1313	3.1750	3.2203	3.2656	3.3109	3.3563	3.4031
1.65	3.1875	3.2313	3.2766	3.3203	3.3656	3.4125	3.4578
1.70	3.2437	3.2875	3.3313	3.3766	3.4219	3.4672	3.5125
1.75	3.3000	3.3422	3.3875	3.4312	3.4766	3.5219	3.5688
1.80	3.3547	3.3984	3.4422	3.4875	3.5313	3.5766	3.6234

	atom 1					
atom 2	1.55	1.60	1.65	1.70	1.75	1.80
1.20	3.0078	3.0562	3.1047	3.1531	3.2016	3.2500
1.25	3.0641	3.1109	3.1594	3.2078	3.2563	3.3047
1.30	3.1188	3.1672	3.2141	3.2625	3.3109	3.3594
1.35	3.1750	3.2219	3.2703	3.3172	3.3656	3.4141
1.40	3.2297	3.2766	3.3250	3.3719	3.4203	3.4688
1.45	3.2844	3.3313	3.3797	3.4266	3.4750	3.5234
1.50	3.3391	3.3875	3.4344	3.4813	3.5297	3.5781
1.55	3.3953	3.4422	3.4891	3.5359	3.5844	3.6313
1.60	3.4500	3.4969	3.5438	3.5906	3.6391	3.6859
1.65	3.5047	3.5516	3.5984	3.6453	3.6922	3.7406
1.70	3.5594	3.6063	3.6531	3.7000	3.7469	3.7953
1.75	3.6141	3.6609	3.7078	3.7547	3.8016	3.8484
1.80	3.6688	3.7156	3.7625	3.8094	3.8563	3.9031

[a] These are the values used for d_0 in eqs 6 and 7. Distances and atom radii are in angstroms.

A more direct test of GB model performance is comparison of GB solvation free energies with those calculated by PB methods. Minimizing error across multiple conformations of the same system is of particular interest for GB methods that will be used in dynamics, as conformation-dependent errors will bias sampling. Figure 4 plots the difference between GB and PB solvation free energies for a series of conformations obtained from a thermal denaturation molecular dynamics trajectory of protein A. Error is reduced for the *GBn* model (standard deviation 6.4 kcal/mol) relative to the *OBC* GB model (standard deviation 7.2 kcal/mol). The *F*-statistic for this improvement approaches the accepted theshold of significance with $p \approx 0.06$. Solvation free energy errors are plotted as a function of the number of native tertiary contacts for the corresponding conformation to elucidate trends in error with respect to degree of denaturation. The GB model of Hawkins et al.[14] has significantly more negative errors for near-native conformations than for denatured conformations, but this native state bias is almost

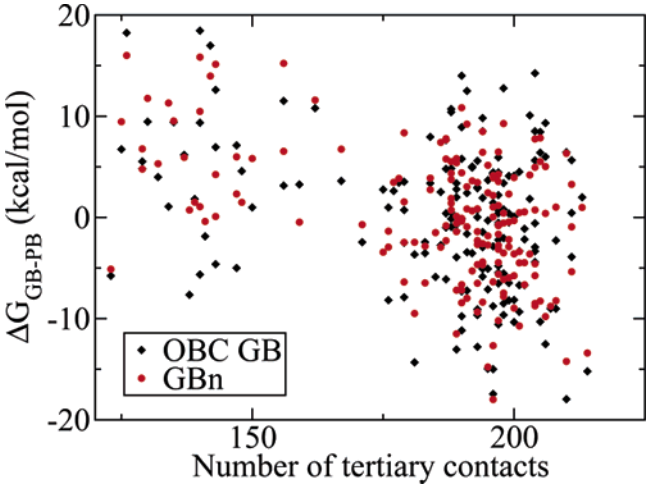


Figure 4. Relative deviation from PB solvation energy for *GBn* and *OBC* GB for a series of snapshots from a denaturation trajectory of protein A. *GBn* has a tighter clustering of points, indicating less random error than *OBC* GB (stdev 6.4 vs 7.2 kcal/mol), while maintaining a similar native state bias (trend of points across the plot). Average errors of −9.2 (*OBC* GB) and 68.9 (*GBn*) kcal/mol removed to facilitate comparison.

entirely corrected by the rescaling function in eq 4 employed by the *OBC* GB model.[32] As seen in Figure 4, the *GBn* model has a very small native state bias, similar to *OBC* GB. Similar, slightly better results are obtained for conformations of protein L and apo-myoglobin along their respective denaturation trajectories; these results are not shown because they were used as part of the objective function in the optimization process and as such are likely to be less indicative of performance on other systems.

3.3. Comparison with PB PMFs. The improvements in effective radius and solvation free energy calculations described above represent useful but still fairly incremental improvement over the existing *OBC* GB model. Indeed, the *OBC* GB model's performance is already quite good on low free-energy conformations, such as those found in crystal structures or sampled from molecular dynamics trajectories, making dramatic improvements on these structures unlikely. However, performance on higher free energy conformations is also important for common applications such as dynamics and docking; here there is ample room for improvement on *OBC* GB. One common high free energy conformation is encountered in the free energy curve for separating a salt bridge or hydrogen bond, referred to here as a PMF to reflect the averaging of solvent degrees of freedom by the implicit solvent model. It has been shown that implicit solvent models that employ a molecular surface dielectric boundary have a free energy barrier to separation of the bond,[44] in qualitative agreement with explicit solvent results,[47] but models based on traditional pairwise integration, even with average molecular volume corrections such as *OBC* GB, fail to reproduce this behavior.

Since the *GBn* model attempts to approximate a molecular surface dielectric boundary, it should be capable of reproducing the barrier in the PMF. As shown in Figures 5 and 6 this result is seen in most cases, a distinct departure from implicit solvent models that allow interstitial high dielectrics.[44] In general, the *GBn* minima are less deep, and the maxima are less high than the PB PMFs. This is probably a consequence of the CFA. The CFA underestimates the descreening contribution of nearby regions relative to more distant regions, because $\mathbf{r}^{-4}$ diminishes less rapidly than the higher order integrands of more accurate expressions.[21,39] Since the neck region is very close to the atom of interest, it seems likely that its effect is underestimated by the CFA, leading to a smaller difference between minimum and maximum. The shallow minima exhibited by the *GBn* model, most notable in the β-sheet model of Figure 5, raise concerns that secondary structure may not be stable, possibly leading to denaturation. However, this has not been observed in molecular dynamics trajectories (see the following), perhaps because the extent of destabilization is less in the protein environment than for these highly solvated model systems, or because the time scales of the simulations conducted here are not sufficient to observe these problems.

3.4. Molecular Dynamics. The primary purpose for the development of computationally efficient pairwise approximated GB is application in dynamics; the *GBn* model, implemented in AMBER, was tested by conducting 10 ns molecular dynamics trajectories of ubiquitin and thioredoxin. As expected, the *GBn* model retains the computational efficiency of the *OBC* GB model running only 8−10% more slowly. Conformational stability of trajectories is commonly assessed by computing the RMSD of alpha carbons from their crystal coordinates; plots of the RMSD for thioredoxin and ubiquitin trajectories conducted using the *GBn* and *OBC* GB models are shown in Figure 7. The *GBn* model maintains approximately the same high level of stability as *OBC* GB, with slightly higher RMSD in the thioredoxin trajectory and lower RMSD in the ubiquitin trajectory.

Performance of a GB model is affected by the set of atomic intrinsic radii used to define the dielectric boundary. Previous work has shown that for simulations conducted under the *HCT* or *OBC* GB models, structural stability is slightly increased and results are somewhat improved by increasing the intrinsic radius of hydrogens bound to nitrogen, H(N), from their Bondi radii[48] of 1.2 Å to 1.3 Å (forming the mbondi2 radius set).[26,32] As seen in Figure 7, little benefit is realized by this change when using the *GBn* model.

The *GBn* model presented here was parametrized for peptides and proteins. This parametrization of *GBn* is not recommended for use with nucleic acids, since they require different degrees of correction for overlap and CFA error than amino acids. In some of our MD simulations of DNA 10 bp duplexes at room-temperature conducted under this *GBn* parametrization we observed breaking of a substantial number of Watson-Crick bonds after a few nanoseconds (results not shown), in contrast to the corresponding explicit water simulations.

3.5. Comparison with Explicit Solvent Ensembles. To examine whether the improved PMFs seen in Figures 5 and 6 translate into improvements in the ensemble of macromolecular conformations sampled during MD, distributions of hydrogen bond lengths were compared between 10 ns ubiquitin trajectories conducted under *OBC* GB, *GBn*, and

Table 4. Maximum Value of Integral of $\mathbf{r}^{-4}$ over the Neck Region (Defined in Eqs 13−15), Tabulated for a Range of Radii for Atoms 1 and 2, Assuming a Solvent Molecule (R_w) Radius of 1.4 Å[a]

	atom 1						
atom 2	1.20	1.25	1.30	1.35	1.40	1.45	1.50
1.20	0.35281	0.36412	0.37516	0.38594	0.39645	0.40670	0.41670
1.25	0.31853	0.32889	0.33902	0.34890	0.35855	0.36797	0.37717
1.30	0.28847	0.29798	0.30728	0.31637	0.32525	0.33392	0.34240
1.35	0.26199	0.27074	0.27930	0.28768	0.29587	0.30387	0.31170
1.40	0.23859	0.24666	0.25455	0.26228	0.26985	0.27725	0.28449
1.45	0.21783	0.22528	0.23258	0.23972	0.24673	0.25358	0.26029
1.50	0.19935	0.20624	0.21300	0.21962	0.22611	0.23247	0.23870
1.55	0.18285	0.18923	0.19550	0.20165	0.20767	0.21358	0.21938
1.60	0.16807	0.17400	0.17982	0.18553	0.19114	0.19664	0.20203
1.65	0.15480	0.16031	0.16573	0.17104	0.17626	0.18139	0.18642
1.70	0.14285	0.14798	0.15303	0.15798	0.16285	0.16764	0.17233
1.75	0.13207	0.13685	0.14155	0.14618	0.15073	0.15520	0.15959
1.80	0.12231	0.12677	0.13117	0.13549	0.13975	0.14393	0.14804

	atom 1					
atom 2	1.55	1.60	1.65	1.70	1.75	1.80
1.20	0.42646	0.43598	0.44527	0.45434	0.46319	0.47183
1.25	0.38615	0.39492	0.40348	0.41185	0.42001	0.42799
1.30	0.35069	0.35878	0.36669	0.37441	0.38196	0.38934
1.35	0.31936	0.32684	0.33416	0.34131	0.3483	0.35514
1.40	0.29158	0.29851	0.30529	0.31193	0.31842	0.32477
1.45	0.26686	0.27330	0.27959	0.28575	0.29179	0.29769
1.50	0.24480	0.25078	0.25664	0.26237	0.26799	0.27349
1.55	0.22505	0.23062	0.23607	0.24141	0.24665	0.25178
1.60	0.20732	0.21251	0.21759	0.22258	0.22747	0.23226
1.65	0.19135	0.19620	0.20095	0.20561	0.21018	0.21466
1.70	0.17694	0.18147	0.18591	0.19027	0.19455	0.19875
1.75	0.16390	0.16814	0.17230	0.17638	0.18039	0.18433
1.80	0.15208	0.15605	0.15995	0.16378	0.16754	0.17124

[a] These are the values used for m_0 in eqs 6 and 7. Distances and atom radii are in angstroms.

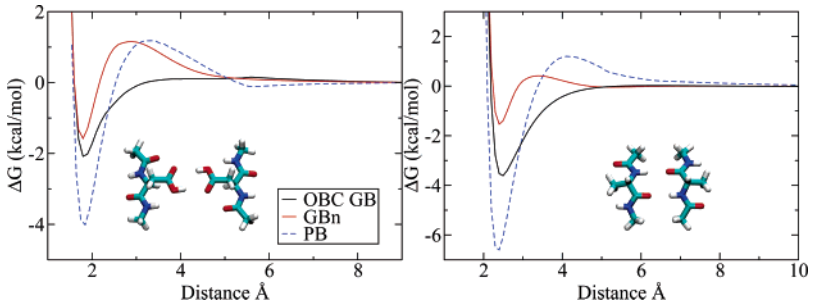


Figure 5. Potentials of mean force for hydrogen bonding systems not included in the objective function, calculated with three implicit solvent methods. Two protonated aspartic acids and two alanines (β-sheet model) are used as examples. Potential includes electrostatic and van der Waals energies.

TIP3P explicit solvation models. Figure 8 illustrates the differences in mean and standard deviation of hydrogen bond length for native backbone hydrogen bonds under the three solvation models. In nearly all cases, the *OBC* GB model yields hydrogen bonds with a higher mean length and standard deviation than in explicit solvent. As a consequence of the narrower potential wells seen in the PMFs, hydrogen bonds under the *GBn* model are generally shorter and their length distributions usually have lower standard deviations than under *OBC* GB. The *GBn* average hydrogen bond lengths are in better agreement with explicit solvent results than *OBC* GB in 24 of 32 cases, while the length distribution standard deviations are in better agreement in 23 of 32 cases. Based on these results, the null hypothesis that there is no impovement of *GBn* over *OBC* GB in reproducing the explicit solvent ensemble of hydrogen bond lengths can be rejected with $p < 0.01$ and $p \approx 0.01$, respectively. The differences between *GBn* and *OBC* GB are particularly noticeable for the shorter, more stable hydrogen bonds (left portions of the plots in Figure 8), where length distributions are presumably mostly determined by the potential between bonding partners, while distributions for longer hydrogen

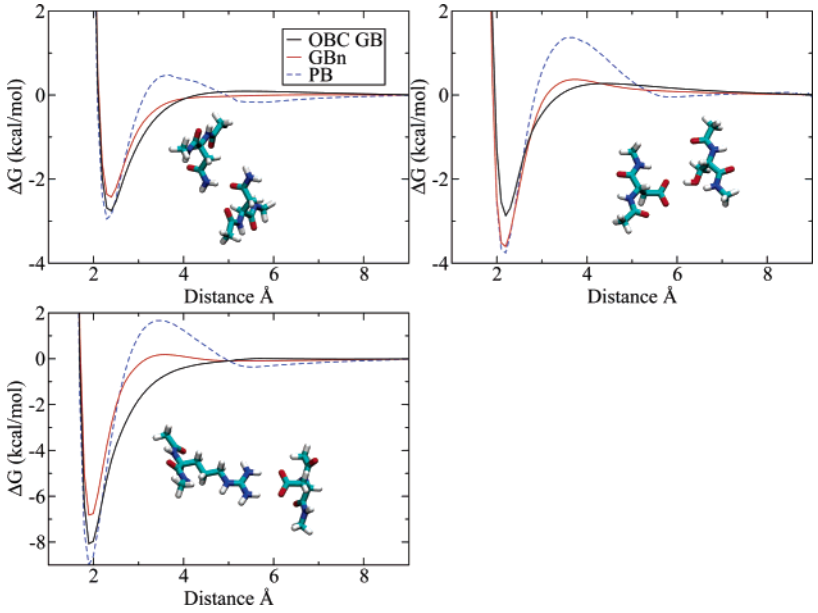


Figure 6. Potentials of mean force for hydrogen bonding and salt bridge systems included in the objective function, calculated with three implicit solvent methods. The hydrogen bonding systems are asparagine and asparagine; aspartate and serine; and arginine and aspartate. Potential includes electrostatic and van der Waals energies.

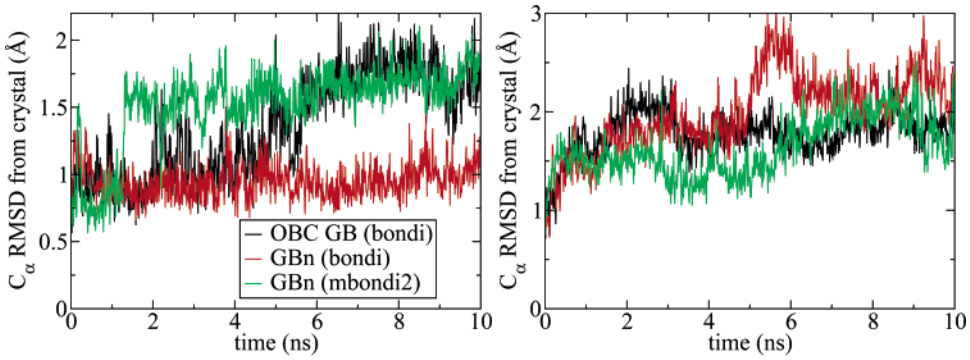


Figure 7. RMSD of alpha carbons from crystal structure over the course of 10 ns of molecular dynamics of ubiquitin (left) and thioredoxin (right).

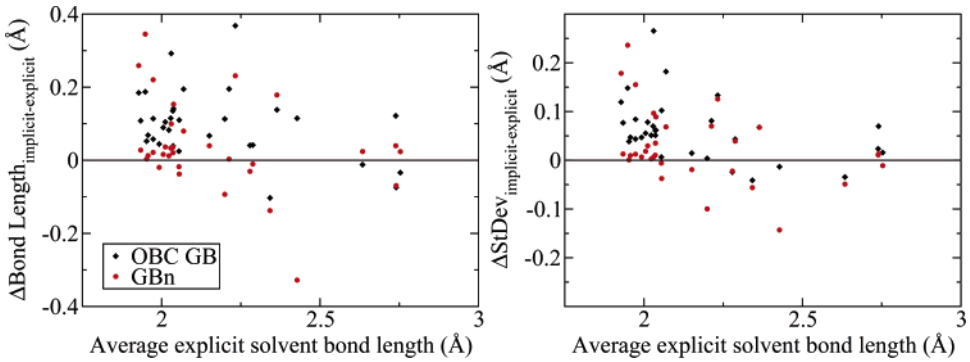


Figure 8. Ubiquitin backbone hydrogen bond length data collected over 10 ns of MD for TIP3P explicit solvent, *OBC* GB and *GBn*. Plots represent the difference between implicit and explicit solvent bond length distribution mean (left) and standard deviation (right) as a function of mean explicit solvent bond length. The zero line represents an exact match between the explicit and implicit solvent results. Hydrogen bond lengths under the *GBn* model are generally in better agreement with explicit solvent results.

bonds may be more affected by tertiary structural forces. The data in Figure 8 suggest that the free energy barrier introduced by the neck correction affects not only dynamics and kinetic properties but also average properties of the ensemble sampled by MD.

Further exploration of the effects of *GBn* on conformational ensembles were conducted using a small polypeptide system where converged sampling of the ensemble is feasible. Simmerling and co-workers[49] recently used replica exchange molecular dynamics (REMD) simulations[50,51] to

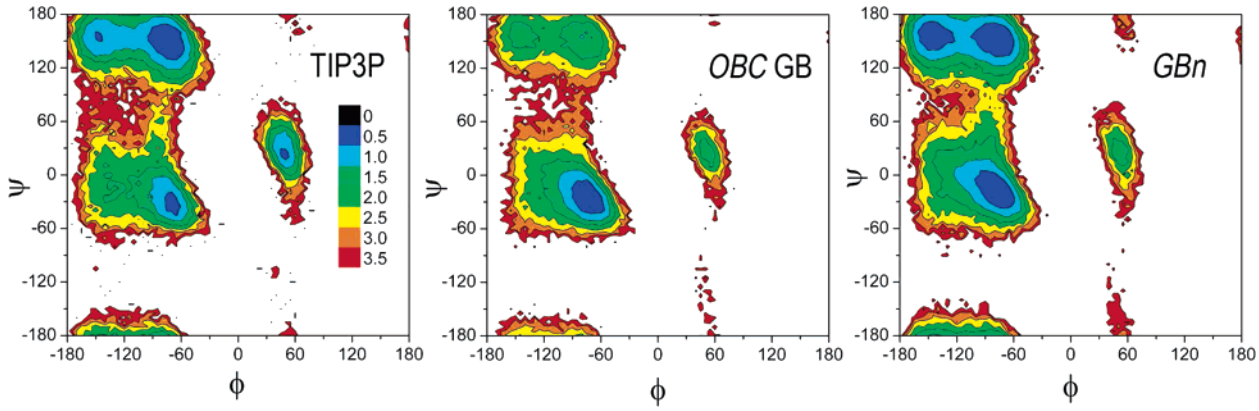


Figure 9. Free energy surfaces at 300 K for the backbone conformation of Ala5 in the Ala10 peptide calculated from 100 ns of REMD. Energies are in kcal/mol, with the lowest free energy assigned a value of 0. TIP3P and *GBn* result in similar free energies for the α, β, and polyproline II basins, while *OBC* GB shows a strong preference for α-helix.

show that the *OBC* GB and *HCT* GB models performed poorly for short polyalanine sequences. Both of these GB models demonstrated a strong bias favoring α-helical conformations as compared to simulations with TIP3P explicit solvent. These calculations (100 ns REMD) were repeated with *GBn*. Figure 9 illustrates the ϕ/ψ free energy surface for Ala5 of deca-alanine. This residue does not adopt a preferred conformation in explicit solvent, with the basins corresponding to the major secondary structure types (right- and left-handed α-helix, β-sheet, and polyproline II) nearly equal in free energy. At the same time, *OBC* GB favors right-handed α-helix by 1−1.5 kcal/mol relative to the other basins; for example, the ratio of the total α-helix to β-sheet populations is 8.67, in noticeable disagreement with the corresponding explicit solvent value of just 1.7. In the *GBn* model, this α-helical bias is no longer present, and the landscape is in much better agreement with the explicit solvent data: the same ratio of α to β populations is 1.64, in very close agreement with the explicit solvent result. Both the *OBC* GB and *GBn* models show somewhat too shallow minima for the left-handed α-helix basin with positive ϕ values as compared to explicit solvent simulations.

Since the free energy surfaces as in Figure 9 give insight primarily into local conformational preferences, more global properties of the chain were examined by calculating end-to-end distance distributions for the ensembles obtained with the different solvent models (Figure 10). As previously described,[49] the distribution is broad in explicit solvent, in concordance with the lack of specific structural preferences seen in Figure 9. At the same time, *OBC* GB yields a shifted distribution that is distinctly peaked near 10 Å due to a high population of fully α-helical conformations that are not observed in explicit solvent. In contrast, the distribution obtained using the *GBn* model is in good agreement with the explicit solvent data, providing further evidence that the neck model represents a significant improvement over the previous *OBC* GB model. These improvements suggest that the correction term introduced by the *GBn* model is a move in the right direction with respect to development of fast analytical GB models. However, due to the computational costs associated with generating explicit solvent PMFs, we have been able to provide direct comparisons for only a few

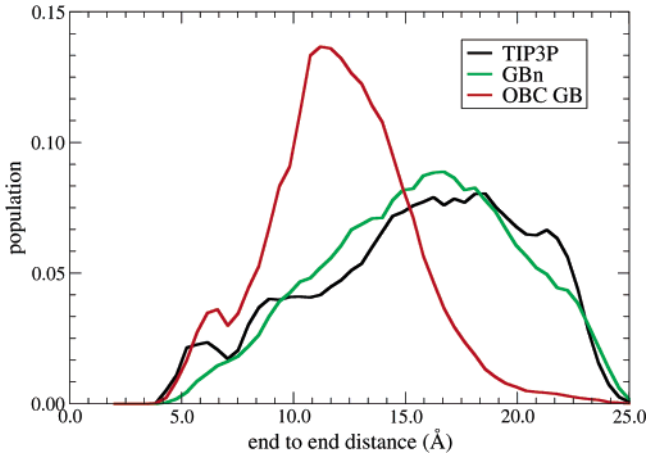


Figure 10. End-to-end distance distributions of deca-alanine at 300 K for 3 solvent models. Profiles from *GBn* and TIP3P explicit water are in good agreement, with a relatively broad distribution slightly peaked near 15−20 Å. However, *OBC* GB significantly differs from the other models, with a strong peak at 10 Å.

systems, and therefore due caution is recommended when applying the *GBn* model to systems dissimilar to those described above.

4. Methods

PB solvation energies and "perfect" radii were calculated using a modified version of APBS 0.3.2. The linearized PB model was employed along with the multiple Debye−Huckel boundary condition. Charge was discretized using the cubic B-spline method (spl2). Dielectric values were 1.0 for solute and 80.0 for solvent regions, except for "perfect" radii calculations, where solvent had dielectric 1000.[35] A Lee-Richards type dielectric boundary (mol) was used. APBS versions 0.3.2 and earlier have a flawed molecular surface algorithm that overestimates solute volume; this flaw was fixed in the APBS version used here. All calculations were performed initially on a coarse grid and then on a smaller, finer grid using the coarse grid potential as boundary conditions. Grid spacings were 0.5/0.25 Å (coarse/fine) for protein solvation and perfect radii calculations and 0.2/0.1-Å for PMF calculations.

GB effective radius, solvation energy, and MD trajectories were calculated using a prerelease version of AMBER 9.[52] MD was carried out using the AMBER ff99 force field.[53,54] Backbone torsional potentials for thioredoxin and ubiquitin were modified by frcmod.mod_phipsi.1;[29] a newer version of these modifications[55] was employed for the Ala_{10} simulations. The time step was 2 fs. Explicit solvent MD employed the TIP3 water model.[56] Implicit solvent MD, GB effective radius, and GB solvation energy calculations used the *OBC* GB or *GBn* models with no cut off. Nonpolar solvation effects were represented using a surface area term of 0.005 kcal/mol·Å^2. Bonds involving hydrogen were constrained using SHAKE.[57] Temperature was maintained at 300 K using the Berendsen weak coupling method and a time constant of 2 ps for the thioredoxin trajectory and using Langevin dynamics with a collision frequency of 1.0 ps^{-1} for the ubiquitin trajectory. The crystal structures (2TRX and UBQ) were prepared for dynamics with 100 steps of steepest descent minimization during which all atoms were harmonically restrained with a weight of 1.0 kcal/mol·Å^2, followed by a 20 ps period of equilibration during which all atoms were harmonically restrained with a weight of 0.1 kcal/mol·Å^2. The enesemble of protein-A structures was generated by temperature unfolding as previously described.[32]

Ensembles of Ala_{10}, with acetylated and amidated N- and C-termini, respectively, were generated using replica exchange molecular dynamics (REMD)[50,51] as implemented in AMBER 8.[52] Data for the TIP3P and *OBC* GB models were previously described.[49] Parameters for REMD simulations were identical to those used for MD (described above), except eight replicas were used to cover the temperature range of 270−570 K (as with the previously described *OBC* GB simulations). Exchanges were attempted every 1 ps, with the REMD simulation running for 100 000 exchange attempts (100 ns). The first 5 ns were discarded. Data convergence was monitored by calculating populations of ϕ/ψ basins corresponding to secondary structure types, which were essentially unchanged after 30 ns. Free energy surfaces were calculated using 2-dimensional histograms for backbone ϕ and ψ dihedrals, with a bin size of 5 degrees. Free energies for bin i relative to the most populated bin were calculated using $\Delta G = -RT \ln(N_i/N_0)$ where N_i and N_0 are the populations of bin i and the most populated bin, respectively. End-to-end distances for Ala_{10} were calculated between C_α atoms of Ala2 and Ala9 (omitting terminal residues) using the ptraj module of AMBER.

Illustrations of molecular geometry in Figures 5 and 6 were produced with VMD[58].

The S_{neck}, α, β, γ, and S_x parameters were optimized using the Nelder-Mead simplex algorithm[45] implemented by the SciPy library.[59] The objective function that was minimized measured agreement between PB and GB solvation free energies over a training set consisting of structures from denaturation trajectories of apo-myoglobin[46] and protein L[21] and structures representing varying degrees of separation of a salt bridge between aspartate and arginine and hydrogen bonds between two asparagine side chains and between serine and aspartate. The total value of the objective function was the sum of each system's contribution. For the structures from the denaturation trajectories, the difference between PB and GB solvation free energy was calculated for each structure, and a linear regression was performed on these data points using the structure's time value (for apo myoglobin) or number of native tertiary contacts (for protein L) as the independent variable, yielding a regression line slope, m, and intercept, b. Additionally, the root-mean-square deviation (RMSD) between PB and GB solvation free energies for each structure was calculated. Each system's contribution to the objective function was defined as $RMSD - |b/2| + |m \cdot (\#\ of\ structures)|$. This term is designed to emphasize minimizing native state bias (represented by m) and random error while not overly penalizing systematic error for a particular system. Salt bridge and hydrogen bond systems consisted of 80 configurations where the bonding partners were separated by 1 Å in the first configuration and are moved 0.1 Å further apart in each subsequent configuration (see Figure 6 for picture of orientations). PB and GB solvation free energies were calculated for each configuration, and the PB and GB solvation free energies were set to be equal at maximum separation by subtracting the energy calculated for maximal separation from that calculated for every other configuration. The objective function term for these systems was the RMSD of the adjusted errors multiplied by 10. The RMSD was increased by a factor of 10 to prevent the objective function from being dominated by the larger errors of the larger protein systems. Since the objective function has multiple local minima, 100 minimizations were performed starting from random initial points. Initial points were chosen from the following intervals of a uniform random distribution: $S_{neck} \in [0.2, 0.5]$, $\alpha \in [0.5, 1.5]$, $\beta, \gamma \in [0.5, 3.0]$, and $S_{\{C,H,N,O\}} \in [0.6, 0.95]$.

5. Conclusion

The *GBn* model, presented here, extends current pairwise GB models with an intuitively attractive property: geometry-dependent exclusion of high dielectric (representing water) from regions into which a water molecule is too large to fit. This extension is computationally efficient, slowing MD simulations by only about 10%. Implementation of the neck correction is simple, requiring only two lookup tables and (in the present implementation) approximately 30 lines of code. The *GBn* model is available in both the *sander* and *pmemd* modules of version 9 of the AMBER suite, and given its simplicity it should be straightforward to add the neck correction to any pairwise volumetric integration-based GB method. Although the correction is a pairwise approximation, it yields nonbonded PMFs with a free energy barrier to separation, a property unique to molecular surface-like dielectric boundaries. The improved agreement between explicit and implicit solvent ensembles sampled by proteins and polypeptides under *GBn* underscores the importance of calculating accurate solvation energies for high free energy configurations as well as the more stable configurations that have traditionally received more attention.

The neck GB model is the fastest model that reproduces the essential characteristics of molecular surface dielectric boundaries, but it does not correlate as well with PB results as the slower GBMV2 model of Lee et al.[38,44] One potential

source of error is the fairly simplistic treatment of neck region overlaps in the current model. Some improvement might be realized by a higher order approach to overlaps, but the largest source of error appears to be the use of the Coulomb field approximation (CFA) to define the integral used to calculate effective radii. Even with a perfect region of integration, errors due to the CFA are large, with effective radii overestimated by a factor of 2 in the worst case.[39] Despite the limitations imposed by the CFA, the current model serves as a proof of principle that a simple pairwise correction can produce an accurate approximation of molecular surface-like solvation properties. We anticipate that a pairwise GB model based on the neck correction and a non-CFA integral, currently under development, will yield substantially improved accuracy.

Acknowledgment. J.M. is supported by Burroughs Wellcome through La Jolla Interfaces in Science. C.S. is supported in part by NIH grant GM6167803 and NCSA allocation MCA02N028 for computer time. D.A.C. is supported in part by NIH grant GM57513. A.O. is supported in part by NIH grant GM076121. This investigation was conducted in part in a facility constructed with support from Research Facilities Improvement Program Grant Number C06 RR-017588-01 from the NCRR, NIH. This work has been supported in part by grants from NSF, NIH, the NSF Center for Theoretical Biological Physics, the National Biomedical Computing Resource, and Accelrys, Inc.

Appendix

The neck region can be analytically defined using only basic trigonometry, but as the derivation is somewhat tedious, the details are provided here. As shown in Figure 1, a triangle is formed by the centers of the atoms and the solvent molecule; the angles of the vertices centered at atoms 1 and 2 are defined as angle A and angle B, and their cosines can be expressed in terms of the four parameters d, R_1, R_2, and R_w, using the law of cosines, as shown in eq 8.

$$\cos A = \frac{d^2 + (R_1 + R_w)^2 - (R_2 + R_w)^2}{2d\,(R_1 + R_w)}$$

$$\cos B = \frac{d^2 - (R_1 + R_w)^2 + (R_2 + R_w)^2}{2d\,(R_2 + R_w)} \qquad (8)$$

The system is cylindrically symmetric about an axis connecting the centers of the two atoms, so it is most naturally analyzed in cylindrical coordinates. The origin is placed at the center of atom 1 with the positive z axis extending toward the center of atom 2. There are three geometric cases for the neck region, illustrated in Figure 11: (i) the atoms overlap and the neck region is ring shaped; (ii) the atoms are moderately separated forming a contiguous region; and (iii) the atoms are widely separated such that the surface of the solvent molecule intersects the z axis, forming two noncontiguous spike regions. When $d \geq R_1 + R_2 + 2R_w$ a solvent molecule can pass between the atoms, and there is no neck region.

For case (i), a second triangle can be formed between the centers of the two atoms and a point at which the surfaces

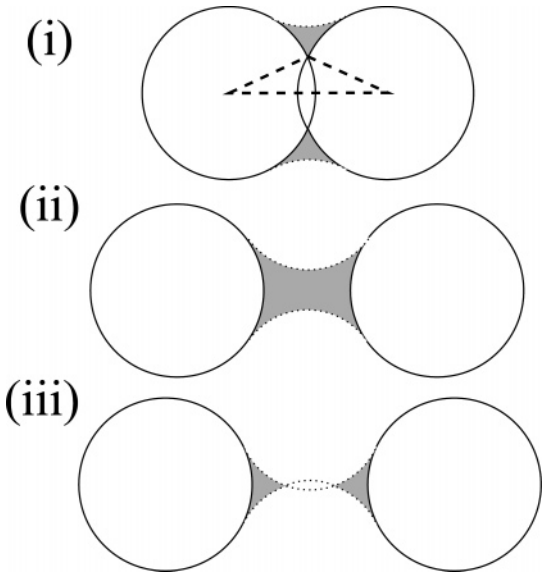


Figure 11. Three cases of neck regions (shaded) formed by atoms (solid circles) at varying separations. Dotted lines represent the surface of the solvent sphere. The leftmost vertex of the dashed triangle in (i) describes the angle A' referenced in eq 9. Although this figure shows two atoms with the same radius, neck regions may also be formed between atoms with unequal radii.

of the atoms intersect. The angle formed by the vertex of this triangle that is located at the center of atom 1 is designated A' and its cosine is defined in eq 9.

$$\cos A' = \frac{d^2 + {R_1}^2 - {R_2}^2}{2dR_1} \qquad (9)$$

Computation of a CFA term based on the neck region requires an expression for the integral of $\mathbf{r}^{-4}$ over the neck region. In the cylindrical coordinate system used here, $|\mathbf{r}| = \sqrt{r^2+z^2}$ so when the volume element is included, the integrand becomes $r(r^2 + z^2)^{-2}$. Because of the cylindrical symmetry, the limits of integration over θ are always 0 to 2π. The upper limit of the integration over r is formed by the surface of the solvent molecule (dotted line in Figures 1 and 11). Since the expression defining the r coordinate of the solvent surface as a function of z (that is, the perpendicular distance from the z axis to the dashed line in Figure 1 for a given z) is somewhat complex, the notation is clarified by defining a function, solv, representing this expression:

$$\mathrm{solv}(z, R_1, R_2, R_w, d) = (R_1 + R_w)\sqrt{1 - \cos^2 A} - \sqrt{{R_w}^2 - (z - (R_1 + R_w)\cos A)^2} \qquad (10)$$

The lower limit of integration for r is defined by the surface of atom 1, the z axis ($r = 0$), or the surface of atom 2, depending on the value of the z coordinate. In addition to defining the geometric extents of the neck region, the limits of integration over z are used to break the overall integral into pieces at the points where the r lower limit of integration changes. Thus in case (i) the integral has two contiguous pieces defined by three z limits: the coordinate at which

the solvent molecule touches atom 1, the coordinate for the intersection of the two atoms, and the coordinate where atom 2 touches the solvent molecule. Case (ii) has three contiguous pieces, with the extreme upper and lower limits defined by the locations that the solvent molecule touches the atoms, as in (i) and the two intermediate limits occurring where the lower r limit changes at the edges of atoms 1 and 2. Finally, case (iii) has two noncontiguous spike regions, each of which is composed of two parts, where the z limits are the intersection of the atom and solvent molecule, the edge of the atom, and the tip of the spike. The tips of the spikes are located at the two points where the solvent sphere intersects the z axis (see Figure 11). The z coordinate of these intersections can be obtained by setting the function in eq 10 equal to zero and solving for z, yielding

$$\mathrm{zinter}_{(-)}(R_1, R_2, R_w, d) = (R_1 + R_w)\cos A - \sqrt{R_1(R_1 + 2R_w)(-1 + \cos^2 A) + {R_w}^2\cos^2 A} \quad (11)$$

$$\mathrm{zinter}_{(+)}(R_1, R_2, R_w, d) = (R_1 + R_w)\cos A + \sqrt{R_1(R_1 + 2R_w)(-1 + \cos^2 A) + {R_w}^2\cos^2 A} \quad (12)$$

Using the preceding definitions, the integrals of $\mathbf{r}^{-4}$ over the neck region for cases (i)−(iii) are presented in eqs 13−15.

$$\text{case(i)}: \int_{\text{neck region}} \mathbf{r}^{-4} = \int_{R_1\cos A}^{R_1\cos A'} \int_0^{2\pi} \int_{\sqrt{R_1^2 - z^2}}^{\mathrm{solv}(z,R_1,R_2,R_w,d)} r(r^2 + z^2)^{-2}\, dr d\theta dz + \int_{R_1\cos A'}^{d - R_2\cos B} \int_0^{2\pi} \int_{\sqrt{R_2^2 - (d-z)^2}}^{\mathrm{solv}(z_1 R_1,R_2,R_w,d)} r\,(r^2 + z^2)^{-2}\, dr d\theta dz \quad (13)$$

$$\text{case(ii)}: \int_{\text{neck region}} \mathbf{r}^{-4} = \int_{R_1\cos A}^{R_1} \int_0^{2\pi} \int_{\sqrt{R_1^2 - z^2}}^{\mathrm{solv}(z,R_1,R_2,R_w,d)} r(r^2 + z^2)^{-2}\, dr d\theta dz + \int_{R_1}^{d - R_2} \int_0^{2\pi} \int_0^{\mathrm{solv}(z,R_1,R_2,R_w,d)} r(r^2 + z^2)^{-2}\, dr d\theta dz + \int_{d - R_2}^{d - R_2\cos B} \int_0^{2\pi} \int_{\sqrt{R_2^2 - (d-z)^2}}^{\mathrm{solv}(z,R_1,R_2,R_w,d)} r(r^2 + z^2)^{-2}\, dr d\theta dz \quad (14)$$

$$\text{case(iii)}: \int_{\text{neck region}} \mathbf{r}^{-4} = \int_{R_1\cos A}^{R_1} \int_0^{2\pi} \int_{\sqrt{R_1^2 - z^2}}^{\mathrm{solv}(z,R_1,R_2,R_w,d)} r(r^2 + z^2)^{-2}\, dr d\theta dz + \int_{R_1}^{\mathrm{zinter}_{(-)}(R_1,R_2,R_w,d)} \int_0^{2\pi} \int_0^{\mathrm{solv}(z,R_1,R_2,R_w,d)} r(r^2 + z^2)^{-2}\, dr d\theta dz + \int_{\mathrm{zinter}_{(+)}(R_1,R_2,R_w,d)}^{d - R_2} \int_0^{2\pi} \int_0^{\mathrm{solv}(z,R_1,R_2,R_w,d)} r(r^2 + z^2)^{-2}\, dr d\theta dz + \int_{\sqrt{d - R_2}}^{d - R_2\cos B} \int_0^{2\pi} \int_{\sqrt{R_2^2 - (d-z)^2}}^{\mathrm{solv}(z,R_1,R_2,R_w,d)} r(r^2 + z^2)^{-2}\, dr d\theta dz \quad (15)$$

References

(1) Gilson, M. K.; Davis, M. E.; Luty, B. A.; McCammon, J. A. *J. Phys. Chem.* **1993**, *97*, 3591−3600.

(2) Honig, B. H.; Nicholls, A. *Science* **1995**, *268*, 1144−1149.

(3) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J. M.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; McCammon, J. A. *Comput. Phys. Commun.* **1995**, *91*, 57−95.

(4) Beroza, P.; Case, D. A. *Energ. Biol. Macromol. B* **1998**, *295*, 170−189.

(5) Scarsi, M.; Apostolakis, J.; Caflisch, A. *J. Phys. Chem. A* **1997**, *101*, 8098−8106.

(6) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161−2200.

(7) Baker, N. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137−143.

(8) Im, W.; Beglov, D.; Roux, B. *Comput. Phys. Commun.* **1998**, *111*, 59−75.

(9) Luo, R.; David, L.; Gilson, M. K. *J. Comput. Chem.* **2002**, *23*, 1244−1253.

(10) Lu, Q.; Luo, R. *J. Chem. Phys.* **2003**, *119*, 11035−11047.

(11) Feig, M.; Brooks, C. L. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217−224.

(12) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.

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

(14) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824−19839.

(15) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578−1599.

(16) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005−3014.

(17) Edinger, S. R.; Cortis, C.; Shenkin, P. S.; Friesner, R. A. *J. Phys. Chem. B* **1997**, *101*, 1190−1197.

(18) Jayaram, B.; Liu, Y.; Beveridge, D. L. *J. Chem. Phys.* **1998**, *109*, 1465−1471.

(19) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983−10990.

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

(21) Lee, M. S.; Salsbury, F. R.; Brooks, C. L. *J. Chem. Phys.* **2002**, *116*, 10606−10614.

(22) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins* **2004**, *56*, 310−321.

(23) Romanov, A. N.; Jabin, S. N.; Martynov, Y. B.; Sulimov, A. V.; Grigoriev, F. V.; Sulimov, V. B. *J. Phys. Chem. A* **2004**, *108*, 9323−9327.

(24) Dominy, B. N.; Brooks, C. L. *J. Phys. Chem. B* **1999**, *103*, 3765−3773.

(25) David, L.; Luo, R.; Gilson, M. K. *J. Comput. Chem.* **2000**, *21*, 295−309.

(26) Tsui, V.; Case, D. A. *J. Am. Chem. Soc.* **2000**, *122*, 2489−2498.

(27) Calimet, N.; Schaefer, M.; Simonson, T. *Proteins* **2001**, *45*, 144−158.

(28) Spassov, V. Z.; Yan, L.; Szalma, S. *J. Phys. Chem. B* **2002**, *106*, 8726−8738.

(29) Simmerling, C.; Strockbine, B.; Roitberg, A. E. *J. Am. Chem. Soc.* **2002**, *124*, 11258−11259.

(30) Wang, T.; Wade, R. C. *Proteins* **2003**, *50*, 158−169.

(31) Nymeyer, H.; Garcia, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13934−13939.

(32) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383−394.

(33) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479−499.

(34) Lee, M. C.; Duan, Y. *Proteins* **2004**, *55*, 620−634.

(35) Sigalov, G.; Scheffel, P.; Onufriev, A. *J. Chem. Phys.* **2005**, *122*, 094511.

(36) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297−1304.

(37) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712−3720.

(38) Lee, M. S.; Feig, M.; Salsbury, F. R.; Brooks, C. L. *J. Comput. Chem.* **2003**, *24*, 1348−1356.

(39) Grycuk, T. *J. Chem. Phys.* **2003**, *119*, 4817−4826.

(40) Im, W.; Lee, M. S.; Brooks, C. L. *J. Comput. Chem.* **2003**, *24*, 1691−1702.

(41) Wojciechowski, M.; Lesyng, B. *J. Phys. Chem. B* **2004**, *108*, 18368−18376.

(42) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 265−84.

(43) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379.

(44) Swanson, J. M. J.; Mongan, J.; McCammon, J. A. *J. Phys. Chem. B* **2005**, *109*, 14769−14772.

(45) Nelder, J. A.; Mead, R. *Comput. J.* **1965**, *7*, 308−315.

(46) Onufriev, A.; Case, D. A.; Bashford, D. *J. Mol. Biol.* **2003**, *325*, 555−567.

(47) Masunov, A.; Lazaridis, T. *J. Am. Chem. Soc.* **2003**, *125*, 1722−1730.

(48) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441−451.

(49) Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2006**, *2*, 420−433.

(50) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140−150.

(51) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(52) Case, D. A.; Darden, T., III; T. E. C.; Simmerling, C.; Wang, J.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Duke, R. E.; Crowley, M.; Brozell, S.; Luo, R.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 9;* University of California, San Francisco: 2006.

(53) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(54) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049−1074.

(55) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *3*, 712−725.

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

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

(58) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics Modell.* **1996**, *14*, 33−38.

(59) Jones, E.; Oliphant, T.; Peterson, P. *SciPy: Open source scientific tools for Python;* SciPy.Org: 2001−2006.

CT600085E

JCTC Journal of Chemical Theory and Computation

Optimizing the Poisson Dielectric Boundary with Explicit Solvent Forces and Energies: Lessons Learned with Atom-Centered Dielectric Functions

Jessica M. J. Swanson,* Jason A. Wagoner, Nathan A. Baker, and J. A. McCammon

Howard Hughes Medical Institute, Center for Theoretical Biological Physics, Department of Chemistry and Biochemistry and Department of Pharmacology, University of California at San Diego, La Jolla, California 92093-0365

Received July 1, 2006 ; Revised Manuscript Received September 25, 2006

Abstract: Accurate implicit solvent models require parameters that have been optimized in a system- or atom-specific manner on the basis of experimental data or more rigorous explicit solvent simulations. Models based on the Poisson or Poisson−Boltzmann equation are particularly sensitive to the nature and location of the boundary which separates the low dielectric solute from the high dielectric solvent. Here, we present a novel method for optimizing the solute radii, which define the dielectric boundary, on the basis of forces and energies from explicit solvent simulations. We use this method to optimize radii for protein systems defined by AMBER *ff*99 partial charges and a spline-smoothed solute surface. The spline-smoothed surface is an atom-centered dielectric function that enables stable and efficient force calculations. We explore the relative performance of radii optimized with forces alone and those optimized with forces and energies. We show that our radii reproduce the explicit solvent forces and energies more accurately than four other parameter sets commonly used in conjunction with the AMBER force field, each of which has been appropriately scaled for spline-smoothed surfaces. Finally, we demonstrate that spline-smoothed surfaces show surprising accuracy for small, compact systems but may have limitations for highly solvated protein systems. The optimization method presented here is efficient and applicable to any system with explicit solvent parameters. It can be used to determine the optimal continuum parameters when experimental solvation energies are unavailable and the computational costs of explicit solvent charging free energies are prohibitive.

I. Introduction

Properly accounting for solvation effects is a long-standing and constantly evolving challenge in computational biophysics. Both the dynamic and thermodynamic properties of a solvated molecule are strongly influenced by the microscopic structure and organization of the water that surrounds it. Predicting solvation free energies and forces accurately across different chemical architectures requires a formalism that collectively accounts for electrostatic, nonelectrostatic, and specific (e.g., hydrogen bonding) solute−solvent interactions. Microscopic formalisms treat these interactions explicitly with an atomistic representation of water. Macroscopic formalisms offer a less physically accurate but more computationally efficient approach: they replace individual molecular interactions with an implicit representation of water, most often as a linearly polarizable continuum.[1] The efficiency of implicit solvent models makes them ideal for large systems and computationally expensive problems such as extensive conformational sampling or high-throughput analyses. The lack of physical detail in implicit solvent models can be compensated by parametrization, for example, fitting Born radii to solvation free energies,[2,3] which generally results in more accurate quantitative predictions for specific

* Corresponding author phone: 858-822-2771; fax: 858-534-4974; e-mail: jswanson@ucsd.edu.

10.1021/ct600216k CCC: $37.00
Published on Web 11/15/2006

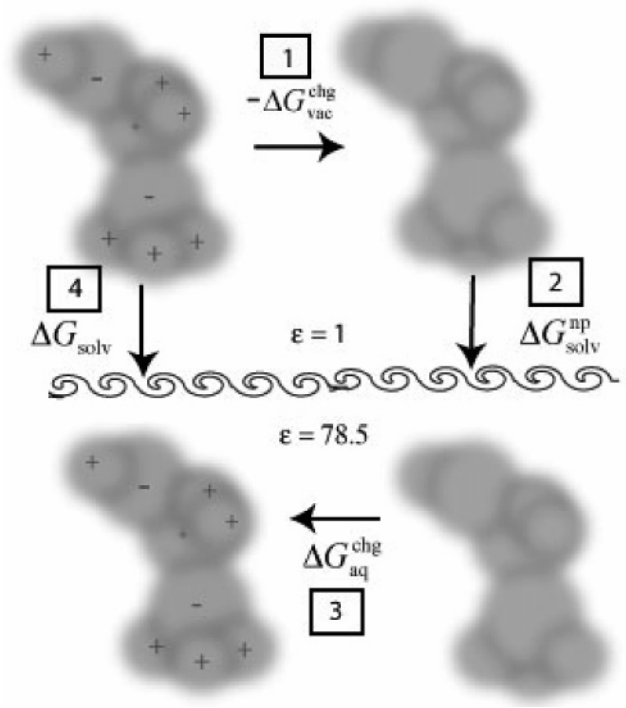


Figure 1. Solvation thermodynamic cycle in which solvation occurs in three path-dependent phases and according to which the total solvation free energy, ΔG_{solv}, is decomposed into polar, $\Delta G^{p}_{solv} = \Delta G^{chg}_{aq} - \Delta G^{chg}_{vac}$, and nonpolar, ΔG^{np}_{solv}, contributions.

solutes but questionable transferability between solutes and different solute conformations.

Most implicit solvent models divide the solvation free energy into polar and nonpolar contributions, $\Delta G_{solv} = \Delta G^{p}_{solv} + \Delta G^{np}_{solv}$. This division is rigorously based on a thermodynamic cycle (Figure 1) in which the solute charges are turned off in a vacuum (phase 1: $-\Delta G^{chg}_{vac}$), the neutral solute is solvated (phase 2: ΔG^{np}_{solv}), and the solute is recharged in the aqueous environment (phase 3: ΔG^{chg}_{aq}). The nonpolar contribution, because of hydrophobic and dispersion interactions, is commonly, though inadequately,[4,5] treated with a solvent-accessible surface area (SASA) model.[6−8] Improvements on this model include independently accounting for dispersion contributions,[4] atom-dependent SASA coefficients,[9] and more recently including an additional term which is proportional to the solvent-accessible volume.[5]

The polar contribution is the difference between charging the system in a vacuum and doing so in the solvent environment, $\Delta G^{p}_{solv} = \Delta G^{chg}_{aq} - \Delta G^{chg}_{vac}$. It can be approximated in a number of ways,[1] but we will focus on methods which solve the Poisson equation,[10−12]

$$-\nabla\cdot[\epsilon(r)\,\nabla\,\phi(r)] = 4\pi\,\rho(r)$$

for the electrostatic potential of the system, $\phi(r)$, given a position-dependent dielectric coefficient, $\epsilon(r)$, and the solute charge distribution, $\rho(r)$. Solving the electrostatic potential of a solute in a vacuum, ϕ_{vac}, and the aqueous environment, ϕ_{aq}, yields the polar solvation free energy

$$\Delta G^{p}_{solv} = -\frac{1}{2}\int \rho(r)[\phi_{aq}(r) - \phi_{vac}(r)]\,dr$$

The parameters which go into both the polar and nonpolar approximations are generally optimized with solvation energies from experiments[13] or explicit solvent simulations.[5,14] Although energies calculated with simulations are only as accurate as the underlying explicit solvent force fields, they can be quantitatively separated into polar and nonpolar contributions. The polar parameters in Poisson calculations include the solute charge distribution and a spatially varying dielectric coefficient. Energies and forces obtained from the Poisson equation are particularly sensitive to the nature and location of the boundary which separates the low dielectric solute ($\epsilon = 1-20$) from the high dielectric solvent ($\epsilon \approx 80$).

Traditionally, the dielectric boundary has been an abrupt transition from low to high dielectric values at the molecular surface, as defined by Lee and Richards[15] or Connolly[16]. Much work has gone into optimizing solute radii for molecular surfaces such that implicit solvent calculations reproduce accurate solvation energies.[13,17,18] Although molecular surfaces work well for most purposes involving static biomolecular structures, they are problematic for continuum solvent dynamics simulations because they are computationally costly and result in unstable forces which vary rapidly with changes in molecular conformation.[19] These problems can be avoided by defining a smooth dielectric transition at the solute surface with overlapping atom-centered polynomial or Gaussian functions.[19,20] The spline-smoothed surface introduced by Im et al.,[20] for example, is particularly well-suited for continuum solvent dynamics simulations because it results in numerically stable and efficient force calculations. Changing the nature of the dielectric transition, however, also changes its optimal location as defined by the solute radii. Not only do spline-smoothed surfaces have a gradual dielectric transition, they also define the low dielectric (internal) volume with a topology that is fundamentally different from the molecular surface. In particular, the spline-smoothed surface introduces interstitial spaces between atoms with significantly higher dielectric values than the remainder of the low dielectric solute volume. For these reasons, radii appropriate for an abrupt transition at the solute surface are not appropriate for spline-smoothed surfaces;[21] in fact, they result in solvation energies and forces which are overestimated by 10−40%.[22,23]

Here, we present a novel and efficient method for optimizing the Poisson dielectric boundary on the basis of explicit solvent simulations. Following the work of Wagoner and Baker,[5] we demonstrate how mean atomic solvation forces can be used alone or in addition to molecular solvation energies to optimize the solute radii for a given partial charge set and surface definition. This approach is closely related to force-matching techniques used in multiscale models[24,25] and provides atomically detailed information about the performance of implicit solvent models. It is significantly more efficient than previous optimization schemes which have relied on either experimental solvation free energies[13,17,26−28] or explicit solvent charging free energies.[18,21,23,29] Its simplicity and efficiency make it ideal for large, unusual,

or highly charged solutes for which solvation energies are unavailable and charging free energies are computationally expensive.

We use this approach to optimize a set of Poisson radii for protein systems defined by spline-smoothed surfaces and AMBER *ff*99 partial charges.[30] We compare radii that have been optimized with forces alone to those optimized with forces and energies, showing that the latter are superior when both forces and energies are important. We test our optimized radii against four other commonly used Poisson parameter sets,[13,28,30,31] each of which has been optimally scaled for spline-smoothed surfaces. Finally, we explore the effects of atom-centered surface definitions on radii optimizations and discuss their limitations in large, globular solutes. Combining the polar parameters developed here with the recently developed and complementary nonpolar parameters[5] should greatly increase the accuracy of the Poisson-based implicit solvent framework for protein systems.

II. Theory

A straightforward statistical mechanical formalism for implicit solvent models[1,32] begins with the potential energy of a solute in an aqueous medium decomposed into

$$U_{\mathrm{TOT}}(X,Y) = U(X) + U(Y) + U(X,Y) \tag{1}$$

where X and Y are the solute and solvent degrees of freedom, respectively, $U(X)$ is the intramolecular solute potential, $U(Y)$ is the solvent−solvent potential, and $U(X,Y)$ is the solute−solvent potential. The system's free energy is then

$$-\beta G = \ln\left(\int \mathrm{d}X \int \mathrm{d}Y\, \mathrm{e}^{-\beta[U(X)+U(Y)+U(X,Y)]}\right) \tag{2}$$

where $\beta = (k_{\mathrm{B}}\mathrm{T})^{-1}$ is the inverse thermal energy. Equation 2 can be used to derive a potential of mean force, $W(X)$, in the usual way:[1,33]

$$\mathrm{e}^{-\beta W(X)} = \frac{\int \mathrm{e}^{-\beta[U(X)+U(Y)+U(X,Y)]}\,\mathrm{d}Y}{\int \mathrm{e}^{-\beta[U(Y)]}\,\mathrm{d}Y} \tag{3}$$

Note that the integral of $\mathrm{e}^{-\beta W(X)}$ over the solute degrees of freedom gives $\mathrm{e}^{-\beta\Delta G_{\mathrm{solv}}}$ where ΔG_{solv} is the solute's solvation free energy. Choosing the unsolvated solute as the reference state, the solvation free energy of a particular solute conformation X is $\Delta W(X) = W(X) - U(X)$. This solvation free energy can be decomposed into polar and nonpolar terms according to a path-dependent process: phases 1−3 of the thermodynamic cycle shown in Figure 1. The coupling between polar and nonpolar solvation interactions, which has been demonstrated in several studies (see Dzubiella et al.[34,35] for discussion), is not explicitly defined in this thermodynamic cycle but implicitly included in the charging phase, $\Delta G_{\mathrm{aq}}^{\mathrm{chg}}$. Decomposition into polar and nonpolar solvation contributions begins by expressing the solute−solvent potential as a sum of polar and nonpolar contributions,[1] $U(X,Y) = U^{\mathrm{p}}(X,Y) + U^{\mathrm{np}}(X,Y)$, such that

$$\Delta W(X) = \Delta W^{\mathrm{np}}(X) + \Delta W^{\mathrm{p}}(X) \tag{4}$$

and

$$\mathrm{e}^{-\beta\Delta W^{\mathrm{np}}(X)} = \frac{\int \mathrm{e}^{-\beta[U(Y)+U^{\mathrm{np}}(X,Y)]}\,\mathrm{d}Y}{\int \mathrm{e}^{-\beta[U(Y)]}\,\mathrm{d}Y} \tag{5}$$

and

$$\mathrm{e}^{-\beta\Delta W^{\mathrm{p}}(X)} = \frac{\int \mathrm{e}^{-\beta[U(Y)+U^{\mathrm{np}}(X,Y)+U^{\mathrm{p}}(X,Y)]}\,\mathrm{d}Y}{\int \mathrm{e}^{-\beta[U(Y)+U^{\mathrm{np}}(X,Y)]}\,\mathrm{d}Y} \tag{6}$$

These contributions can be calculated in two successive stages of a free energy perturbation or thermodynamic integration[36] according to

$$\Delta W^{\mathrm{np}}(X) = \int_0^1 \mathrm{d}\lambda_1 \left\langle \frac{\partial U^{\mathrm{np}}(X,Y)}{\partial \lambda_1} \right\rangle_{(\lambda_2=0)} \tag{7}$$

and

$$\Delta W^{\mathrm{p}}(X) = \int_0^1 \mathrm{d}\lambda_2 \left\langle \frac{\partial U^{\mathrm{p}}(X,Y)}{\partial \lambda_2} \right\rangle_{(\lambda_1=1)} \tag{8}$$

where λ_1 and λ_2 are coupling parameters that scale the nonpolar and polar solvent−solute interactions from 0 (off) to 1 (on). The polar contribution is therefore a "charging free energy" but is often referred to as the polar solvation free energy. The mean forces can be obtained by differentiation of $W(X)$ with respect to the solute coordinates[1], X

$$\langle F^{\mathrm{np}}(X)\rangle = -\frac{\partial W^{\mathrm{np}}(X)}{\partial X} = -\left\langle \frac{\partial U^{\mathrm{np}}(X,Y)}{\partial \mathrm{X}} \right\rangle_{\mathrm{np}} \tag{9}$$

and

$$\langle F^{\mathrm{p}}(X)\rangle = -\frac{\partial W^{\mathrm{p}}(X)}{\partial X} = -\left\langle \frac{\partial U^{\mathrm{p}}(X,Y)}{\partial \mathrm{X}} + \frac{\partial U^{\mathrm{np}}(X,Y)}{\partial \mathrm{X}} \right\rangle_{\mathrm{p}} + \left\langle \frac{\partial U^{\mathrm{np}}(X,Y)}{\partial \mathrm{X}} \right\rangle_{\mathrm{np}} \tag{10}$$

where $\langle\cdots\rangle_{\mathrm{np}}$ and $\langle\cdots\rangle_{\mathrm{p}}$ denote the ensemble averages with nonpolar only and full (nonpolar and polar) solute−solvent interactions, respectively.

In this work, we will focus on polar contributions, $\Delta W^{\mathrm{p}}(X) \equiv \Delta G_{\mathrm{solv}}^{\mathrm{p}}$ and $\langle F^{\mathrm{p}}(X)\rangle \equiv \bar{F}^{\mathrm{p}}$. A similar treatment of nonpolar contributions, presented by Wagoner and Baker,[22] was utilized to optimize nonpolar implicit solvent parameters with explicit solvent forces.[5] The goal of the present work is to optimize polar parameters, specifically the solute radii used to define spline-smoothed dielectric boundaries in protein systems, by fitting Poisson solvation free energies and forces to explicit solvent charging free energies and mean atomic forces.

Our approach assumes that polar parameters are independent of the solute's conformation and polar−nonpolar coupling. Though this is clearly an approximation,[34,35] it may be a reasonable approximation for biomolecular systems because they are predominantly composed of similar subunits (e.g., amino and nucleic acids) with significant solute−solvent dispersion interactions. Previous efforts have shown that optimizing complementary polar and nonpolar parameters in a system- or atom-dependent manner can result in

Table 1. Force and Force−Energy-Optimized Radii

atom name[a]	residues	opt. radii[b]	opt. radii[c]
	Backbone		
C	all	2.338	2.317
O	all	1.766	1.732
N	all	2.331	2.399
CA	all except G	2.425	2.428
CA	G	2.122	2.353
CAY/CAT	ACE,NME	2.022	2.148
	Side Chains		
CB	D,E,C,H,M,F,S,T,W,Y	2.128	2.231
CB	A,R,N,Q,I,L,K,V	2.209	2.576
CG*	R,Q,I,L,K,M,T,V	2.414	2.414
CG	H,F,W,Y	2.147	2.136
CG/CD	N,Q,D	2.197	2.353
CG	E	2.195	2.341
CB/CG/CD	P	2.506	2.440
CD	R,K	2.316	2.442
CD*	I,L	2.640	2.559
CD*/CE*/CZ	H,F,W,Y	2.282	2.217
CE	M	1.940	1.972
CZ/CE	R,K	2.398	2.461
OD*/OE*	N,Q,D,E	1.729	1.756
OG*	S,T	1.956	2.101
OH	Y	1.871	2.108
NE,NH*,NZ	R,K	2.323	2.328
ND2/NE2	N,Q	2.122	2.397
ND1,NE2	H	1.927	2.040
NE1	W	1.880	1.960
SG/SD	C,M	2.406	2.337
	Hydrogens[d]		
type H	bb HN	1.244	1.388
type H	bound to N	1.228	1.302
type HO/HS	bound to O/S	0.999	0.917
type H1/HP	polar	1.905	2.019
type HC/HA	nonpolar	1.809	1.680

[a] Radius groups are distinguished by AMBER atom names for all heavy atoms and by atom type for hydrogen atoms. [b] Final radii [Å] from force optimizations. [c] Final radii [Å] from force and energy optimizations. [d] Hydrogens specified by atom type with type "H" divided into two groups: amide backbone "HN" and all other N-bound hydrogens.

fairly accurate, though sometimes system-dependent, solvation frameworks.[13,26,27] The SM*x* models, for example, can predict solvation free energies for neutral molecules with errors less than 1 kcal/mol and have shown recent success in treating charged solutes and ions with the addition of an explicit water molecule.[26,27] Thus, combining the presented polar parameters with complimentary nonpolar parameters[5] should result in a complete Poisson-based implicit solvent framework for protein systems that is accurate to the exclusion of polar−nonpolar coupling, conformational dependence, and the limitations of mean-field theories.

III. Methods

Force-based optimizations involve three steps, which are described in more detail below. First, explicit solvent forces are collected for each solute in the training set in one or more fixed conformations. Next, the continuum calculations are set up such that energies and forces are well converged, that is, independent of grid resolution, boundary conditions, and so forth. Finally, the Poisson atomic radii are optimized to reproduce the explicit solvent forces. Optimizations using both energies and forces additionally require calculating explicit solvent charging free energies with thermodynamic integration or free energy perturbation simulations as previously reported.[23]

Training Sets. To test the limitations of our methodology, we began by optimizing *each* atomic radius in three nonzwiterionic *N*-acetyl-X-*N*′-methylamide dipeptides, where X represents alanine with ϕ/ψ angles of 180°/180° (ala1), alanine with ϕ/ψ angles of −60°/−40° (ala2), and serine with ϕ/ψ angles of 180°/180° (ser). As previously described,[23] using simple PARSE, Bondi, or AMBER-vdW atom types limits the accuracy of the resulting parameters because of significant differences in the solvation structures, energies, and forces for atoms within these atom types. Therefore, we used 31 "atom types" or radius groups (Table 1) which were previously identified as having similar chemical signatures and surrounding solvation structures according to explicit solvent charge distribution functions.[23]

The training and test sets are summarized in Table 2. We used two sets of model systems in our force-based optimizations. The first set was two conformations of intestinal fatty acid binding protein (IFABP) representing highly populated conformational clusters from an a priori fully equilibrated explicit solvent simulation.[22] The second set was two short polypeptide chains: Trpcage (PDB ID 1l2y)[37] and the C-terminal fragment (residues 41−56) of protein G (PDB ID 2gb1),[38] hereafter called G-peptide. In the energy- and force-based optimizations, we additionally used the explicit solvent charging free energies of 20 *N*-acetyl-X-*N*′-methylamide dipeptides, where X represents each of the amino acids (ϕ/ψ angles of 180°/180°) and seven polyalanine peptides with common secondary structures (three β turns, two helices, and two β hairpins). Each of the polyalanine peptides was modeled from fragments of lysozyme (PDB ID 1ati) or crambin (PDB ID 1ejg), which were mutated to polyalanine and capped with neutral blocking groups. The specific peptide conformations and explicit solvent charging free energies were previously reported.[23]

Testing Sets. To determine if the optimized radii are transferable to molecules outside of the training set, we used the forces and solvation energies of the other test systems (e.g., the Trpcage and G-peptide forces for the IFABP-force-trained radii) as well as the solvation energies of the following molecules: Trpcage, G-peptide, two additional proteinlike polypeptides, five new polyalanine peptides, and 20 amino acid dipeptides in a new side chain and backbone conformation (ϕ/ψ angles of −60°/−40°).[23] We also used the forces of IFABP in six new conformations. The relative root-mean-square deviation (RMSD) values for all eight IFABP structures ranged from 1.2 to 2.2 Å.[22] All of the aforementioned systems were used to compare the optimized radii to four other continuum parameter sets: AMBER *ff*99 partial charges combined with *ff*99 van der Waals (vdW) radii,[30] Bondi radii,[31] and a set of radii recently published by Luo et al.,[28] as well as the PARSE parameters (charges

Table 2. Solutes Used to Train and Test Radii on the Basis of Explicit Solvent Forces or Energies

training sets solute[s]	quantity used	trained radii	additional test sets solute[s]	quantity used
IFABP conformations 1 & 2	forces	f-opt 1	IFABP conformations 3−8	forces
Trpcage & G-peptide	forces	f-opt 2 and fe-opt	20 dipeptides conformation 2	energies
20 dipeptides conformation 1	energies	fe-opt	polyalanine peptides 8−12	energies
polyalanine peptides 1−7	energies	fe-opt	4 proteinlike peptides	energies

and radii).[13] Each of the comparison parameter sets was scaled for use with spline-smoothed surfaces as described below.

Explicit Solvent Forces. As shown in eq 10, the polar solvation forces are calculated by subtracting the nonpolar solvation forces from the total forces ($\bar{F}^{p} = \bar{F}^{np+p} - \bar{F}^{np}$). The nonpolar forces are averaged from an ensemble generated with only nonpolar (vdW) solute−solvent interactions, while the total forces are averaged over an ensemble generated with both polar (electrostatic) and nonpolar solvent−solute interactions. Thus, the calculation of polar forces requires two simulations: one with all solute−solvent interactions and one with only nonpolar solute−solvent interactions (i.e., solute charges set to zero).

The IFABP simulations were performed, as previously reported,[5] with the AMBER *ff*99[30] force field using AMBER 7 software.[39] Eight conformations were obtained by clustering a 2.5 ns simulation which started from a nuclear magnetic resonance structure (PDB ID 1ael)[40] and was run in TIP3P water in the presence of ∼160 mM NaCl at 300 K and 1 atm of pressure. The clusters were calculated with the single linkage algorithm available in GROMACS[41] using a 1.0 Å cutoff over all main-chain atoms. The eight resulting clusters accounted for 81% of the trajectory structures. Each conformation was then constrained via belly dynamics[39] and simulated for 1.05 ns with SHAKE-enabled 2 fs time steps under isobaric−isothermal conditions. The forces for conformations 3−6 were averaged from 250 snapshots taken every 4 ps from the last nanosecond of simulation. The simulations of conformations 1 and 2 were extended an additional 4 ns, and forces were averaged from 1000 snapshots taken every 4 ps from the last 4 ns of simulation. Comparing these forces with those averaged from the last nanosecond of an independent 1.05 ns simulation resulted in an average squared residual of 0.085 56 kcal/mol Å, indicating that the forces were well-converged.

Trpcage, G-peptide, and the three dipeptides (ala1, ala2, ser) were simulated with the AMBER *ff*99 converted to CHARMM format and the CHARMM software (version 31a1).[42] The solutes were protonated and solvated in a sphere of TIP3P water molecules that extended 10.0 Å beyond the solute, resulting in three to four hydration shells. The spherical solvent boundary potential including Kirkwood's multipolar expansion reaction field was used to approximate the influence of bulk water beyond the explicit water sphere.[43] This model was previously used in the free energy perturbation simulations to calculate charging free energies[23] because it alleviates difficulties that result from perturbing charged systems with periodic boundary conditions. It was chosen here for consistency with the molecular charging free energies. It has been shown to give reliable results for proteins, nucleic acids, and small molecules.[1,18,44] All solute atoms were restrained to their original coordinates. The solvent was first energy-minimized with 50 steps of steepest descent followed by 1000 steps of the ABNR method and then equilibrated for 200 ps. Langevin dynamics were employed at a constant temperature (300 K) using SHAKE-enabled 2 fs time steps, infinite cutoffs for nonbonded interactions, and a friction constant corresponding to a relaxation time of 5 ps applied to water oxygen atoms.

The peptide forces were averaged from snapshots taken every 0.2 ps over 500 ps of pre-equilibrated simulation for a total of 2500 conformations. Comparing forces from the first 200 ps with those from the full 500 ps resulted in a root-mean-square deviation (RMSD_$\bar{F}^{p}$) of 0.187 kcal/mol/Å and an average relative error of 0.0884 kcal/mol/Å, demonstrating reasonable convergence after only 200 ps. Comparing duplicate 500 ps simulations of the dipeptides dropped RMSD_$\bar{F}^{p}$ to 0.05−0.09 kcal/mol/Å, which was considered sufficient convergence. Because the IFABP and peptide systems were used in independent optimizations, the effects of the different simulation protocols for these two systems are expected to be minor.

Implicit Solvent Forces. All implicit solvent energies and forces were calculated with the Adaptive Poisson−Boltzmann Solver (APBS; http://apbs.sf.net/) version 0.4.0.[45] APBS employs the analytical method of Im et al.[20] to evaluate polar solvation forces. In particular, the dielectric function, $\epsilon(r)$, is defined by overlapping atom-centered cubic polynomials

$$\epsilon(r) = 1 + (\epsilon_s - 1)\prod_i H_i(|r - r_i|)$$

where ϵ_s is the dielectric coefficient for the solute, H_i is the

$$H_i(r) = \begin{cases} 0 & r < R_i - w \\ \dfrac{-1}{4w^3}(r - R_i + w) + \dfrac{3}{4w^2}(r - R_i + w)^2 & R_i - w < r < R_i + w \\ 1 & r \geq R_i + w \end{cases}$$

polynomial for each atomic sphere, r is the distance from a point in the system and atom i, R_i is that atom's radius, and w is the half-width of the transition region.[20] The resulting total solvation force is composed of three terms: a reaction field force due to interaction of the solute charge with the total electric field, a dielectric boundary force due to spatial variation in the dielectric function, and an ionic boundary force for systems with ions. The last term was zero in this study because we used zero bulk ionic strength. The abruptness of the transition between the solute and solvent

values in these spline-smoothed surfaces is controlled by the spline window half-width, w, a user-defined parameter in APBS. We optimized our radii with the recommended value, $w = 0.3$ Å. We additionally scaled these radii for spline windows 0.2 Å $\leq w \leq$ 1.0 Å with both force- and force−energy-based genetic algorithm optimizations (as described below). The Poisson equation was solved at a temperature of 300 K with the multiple Debye−Hückel sphere boundary condition and cubic B-spline charge discretization. We used the experimental solvent dielectric of 78.4, which differs from the reported dielectric constants[46,47] of 91−104 for TIP3P water[48] but is consistent with typical values used with implicit solvent models. Calculations run with the PARSE parameters used a solute dielectric of 2.0 in agreement with their original development.[13] All other calculations used a solute dielectric of 1.0, which is appropriate for calculations that explicitly sample solute conformations and desirable for consistency with nonpolarizable force fields. It is important to ensure that the chosen continuum parameters, particularly the grid parameters, provide sufficient spatial resolution to ensure the convergence of energies and forces. We compared the energies and forces calculated with successively finer grids (down to 0.10 Å) and selected production calculation grid parameters that resulted in relative errors below 0.5%. Our final grids had resolutions of either 0.2 or 0.25 Å, depending on the solute, and overall dimensions which extended beyond the solute by 35% of its length in each dimension.

Genetic Algorithm Optimizations. We used a genetic algorithm to optimize the radii from various starting values (AMBER *ff*99 vdW, Bondi, and our previously published smooth boundary radii[23]) to their final values. Populations of 50 solutions were run for 10 to 100 generations. The initial populations were generated from a uniform distribution that varied from 0.1 to 0.3 Å around the starting values, depending on the optimization. The fitness of each solution was evaluated, and the next population was generated via selection, crossover, and mutation. Selections were based on the stochastic universal sampling algorithm,[49] which chooses solutions on the basis of a probability proportional to their fitness

$$p_i = \frac{f_i}{\sum_{j=1}^{N} f_j} \tag{11}$$

where p_i is the probability that solution i will be selected and f_i is its fitness defined by

$$f_i = (1 + \mathrm{RMSD\_}\bar{F}_i^{\mathrm{p}})^{-1} \tag{12}$$

where RMSD_$\bar{F}^{\mathrm{p}}$ is the standard deviation of all 3N force components in an N-atom system for the radii solution set, i. The fitness scores range from 0 (poor) to 1 (perfect). A uniform crossover process in which two solutions from the previous generation were randomly distributed to two new solutions was applied to either 80% or 90% of the population, while mutations selected from a Gaussian distribution with standard deviations which ranged from ±0.025 to ±0.1 Å were applied to anywhere from 5% to 30% of the population. The various different crossover and mutation values were tested to ensure that the final solutions were not an artifact of limiting genetic algorithm run parameters. This process was repeated until a desired fitness was reached or the maximum number of generations was exceeded.

As we will demonstrate, accurately fit solvation forces do not necessarily ensure accurate solvation energies. Thus, we ran a second set of optimizations that incorporated the solvation energies of 20 amino acid dipeptides and seven polyalanine peptides as well as the Trpcage and G-peptide solvation forces into the fitness function according to

$$f_i = [1 + 1/2(\mathrm{RMSD\_}\bar{F}_i^{\mathrm{p}} + \mathrm{AAE\_}\Delta G_{p\_i}^{\mathrm{solv}})]^{-1} \tag{13}$$

where RMSD_$\bar{F}^{\mathrm{p}}$ is defined as before and AAE_$\Delta G_{p_i}^{\mathrm{solv}}$ is the average absolute error for the 27 solutes' polar solvation energies for solution set i. Because the optimal RMSD_$\bar{F}^{\mathrm{p}}$ and AAE_$\Delta G_{p_i}^{\mathrm{solv}}$ terms are not known a priori, it is not possible to define a fitness function that would weight the forces and energies equally. Equation 13 was chosen to give a small bias to energies over forces based on previously computed AAE_$\Delta G_{p_i}^{\mathrm{solv}}$ values (∼0.9−4.0 kcal/mol) which were slightly larger than the RMSD_$\bar{F}^{\mathrm{p}}$values (∼0.7−1.5 kcal/mol/Å).

Scaling Radii for Spline-Smoothed Surfaces. Because the spline-smoothed surface brings high dielectric values closer to each atom's center than the molecular surface does, radii that work well with molecular surfaces overestimate solvation energies and forces by 10−40% when used with spline-smoothed surfaces.[22,23] A first-order correction to this overestimation is to scale each radius by a single factor, x, according to

$$R_{\mathrm{new}} = R_{\mathrm{orig}} + x \tag{14}$$

where R_{new} is the new radius and R_{orig} the original. The four parameter sets chosen to benchmark our optimized radii are all intended for use with molecular surfaces. Therefore, to make a fair comparison, we had to determine the best scaling factor, x_{best}, for each set of radii. We chose to do this using force-only and force- and energy-based optimization schemes analogous to those used in our radii optimizations described above. We used the same approach to scale *down* our spline-smoothed radii from Table 1 for molecular surfaces in order to compare the two surface topologies (molecular and spline-smoothed) directly. Finally, we scaled our radii (Table 1), which use a spline window half-width of $w = 0.3$ Å, for different spline window half-widths ranging from $w = 0.2$ to 1.0 Å. We selected this scaling function (eq 14) over that used in previous optimizations, $R_{\mathrm{new}} = x(R_{\mathrm{orig}} + w)$,[21] because it resulted in slightly higher final fitness values and converged in one to three fewer generations.

IV. Results

Dipeptide Test Case. It is important know the optimal degree of accuracy one can expect from the presented methodology, that is, the instrinsic limitations of force- and force−energy-based parametrizations. To address this question, we generated "perfect (Poisson) radii"[50] by optimizing *every* atom's radius in three simple dipeptides, alanine in

Table 3. Individually Optimized Radii

	force-optimized ala1 radii		force−energy-optimized ala1 radii	
	$\Delta\Delta G_{solv}$[a]	RMSD_F[b]	$\Delta\Delta G_{solv}$[a]	RMSD_F[b]
ala1	−1.753 (0.290)	0.147 (0.009)	0.000 (0.005)	0.244 (0.025)
ala2	−1.594 (0.336)	0.358 (0.013)	0.152 (0.073)	0.432 (0.026)
ser	−3.314 (0.247)	0.344 (0.020)	−1.633 (0.164)	0.333 (0.046)
	force-optimized ala2 radii		force−energy-optimized ala2 radii	
	$\Delta\Delta G_{solv}$[a]	RMSD_F[b]	$\Delta\Delta G_{solv}$[a]	RMSD_F[b]
ala1	−1.865 (0.359)	0.302 (0.009)	−0.054 (0.103)	0.287 (0.020)
ala2	−1.906 (0.265)	0.160 (0.003)	−0.001 (0.005)	0.322 (0.046)
ser	−3.810 (0.391)	0.412 (0.011)	−1.820 (0.168)	0.342 (0.027)
	force-optimized ser radii		force−energy-optimized ser radii	
	$\Delta\Delta G_{solv}$[a]	RMSD_F[b]	$\Delta\Delta G_{solv}$[a]	RMSD_F[b]
ala1	−0.603 (0.210)	0.291 (0.036)	1.476 (0.171)	0.422 (0.027)
ala2	−0.450 (0.209)	0.390 (0.009)	1.603 (0.148)	0.616 (0.024)
ser	−2.420 (0.233)	0.200 (0.009)	0.000 (0.002)	0.418 (0.025)
	force-optimized radii		force−energy-optimized radii	
	$\Delta\Delta G_{solv}$[c]	RMSD_F[d]	$\Delta\Delta G_{solv}$[c]	RMSD_F[d]
ala1	0.376	0.262	0.217	0.265
ala2	0.569	0.444	0.449	0.476
ser	−0.831	0.560	−0.842	0.598

[a] The *average* solvation energy deviation [kcal/mol]. [b] *Average* force RMSD [kcal/mol/Å] for three dipeptides using individually force and force−energy-optimized radii. Statistics collected from 30 independent genetic algorithm optimizations. Standard deviations are given in parentheses. [c] The solvation energy deviation [kcal/mol]. [d] Force RMSD [kcal/mol/Å] for the three dipeptides using the force and force−energy-optimized radii presented in Table 1.

two different conformations (ala1 and ala2) and serine (ser). We used both the force- and the force−energy-based approaches (eqs 11 and 12, respectively). To ensure that the results were independent of the genetic algorithm run and the parameters, the optimizations were repeated 30 times varying the initial population distribution from ±0.1 to 0.4 Å, the mutation distribution from ±0.025 to 0.1 Å, and the mutation ratio from 0.05 to 0.4 (i.e., mutations were applied to 5−40% of the solutions in the next generation). The resulting force and polar solvation energy statistics (shown in Table 3) demonstrate errors that are larger than the inherent errors in the explicit and implicit calculations. The explicit solvation energy and force errors were $\Delta\Delta G_p^{solv} \approx 0.02-0.17$ kcal/mol and RMSD_$\bar{F}^p \approx 0.05-0.09$ kcal/mol Å, respectively, while the continua were $\Delta\Delta G_p^{solv} \approx 0.03$ kcal/mol and RMSD_$\bar{F}^p \approx 0.05$ kcal/mol Å.

There are five key points to take away from this example. First, the genetic algorithm solution space is highly frustrated. The majority of optimizations generated solutions of similar fitness but unique parameters, indicating that there are multiple solutions even for simple molecules with independently fit radii. Second, the explicit solvent forces cannot be perfectly reproduced even with independently fit radii. Whether this discrepancy is caused by the limited dielectric definition (i.e., the union of spline-smoothed spheres) or limitations in the Poisson equation itself (i.e., the assumption of linear and local solvent response) is not clear. It is our expectation that all of these limitations contribute to the differences between explicit and implicit solvation forces but that the limited dielectric definition is the major source of error. This analysis suggests that the limit of accuracy for implicit solvent forces calculated with this surface is RMSD_$\bar{F}^p \gtrsim 0.15$ kcal/mol Å. Third, each solute has a unique set of optimal continuum radii. When the force-optimized radii for the serine peptide are used on alanine or vice versa, the average RMSD_$\bar{F}^p$ increases by 0.14 to 0.21 kcal/mol Å. Similarly, when the force−energy-optimized radii are switched, the average $\Delta\Delta G_p^{solv}$ decreases from ∼0.0 to between −1.820 and −0.054 kcal/mol. Because serine and alanine share many of the same atom types, these increased errors indicate that neighboring atoms *change* the optimal continuum radius for a specific atom type. Fourth, the solute's conformation also changes the optimal continuum radii. When the radii are switched for ala1 and ala2, we observe a similar, though less dramatic, increase in the average RMSD_$\bar{F}^p$ by 0.15 to 0.16 kcal/mol Å and a change in the average $\Delta\Delta G_p^{solv}$ to between −0.054 and 0.152 kcal/mol. Although the goal of optimizing radii with multiple solutes and conformations is to find a solution set which is transferable to similar solutes and to different solute conformations, it is not likely that such a solution set will be more accurate than radii that are individually optimized for these three very similar peptides. Thus, one can expect the limit of accuracy of force-optimized radii used on varying

Table 4. Summary of Force RMSDs and Solvation Energy Average Absolute Errors (AAEs) for Optimized Radii and Comparison Parameter Sets

		RMSD_F [kcal/mol/Å]		AAE_ΔG_{solv} [kcal/mol]			
parameters[a]	x_{best}	Trpcage & G-peptide[b]	IFABP[c]	di/poly-ala peptides[d]	di/poly-ala peptides[e]	protein	peptides[f]
Amber-f	0.378	0.870	1.020	4.953	4.719	20.296	7.60%
Amber-fe	0.281	1.055	1.222	1.986	1.873	14.997	6.02%
Parse-f	0.236	1.801	2.438	4.766	6.346	26.608	10.99%
Parse-fe	0.310	1.822	2.178	2.944	3.342	11.213	4.00%
Parse0-f	0.236	1.812	2.456	4.186	4.346	27.778	11.47%
Parse0-fe	0.269	1.816	2.327	3.929	3.342	14.725	6.23%
Bondi-f	0.311	1.216	1.460	3.314	2.356	28.122	11.29%
Bondi-fe	0.373	1.240	1.443	1.088	0.971	7.014	2.99%
Luo-f	0.403	0.923	1.014	5.565	5.173	15.108	5.67%
Luo-fe	0.283	1.150	1.360	1.902	1.949	26.472	10.74%
IFABP Opt-f	na	0.750	0.906	4.136	3.799	11.379	4.35%
optimized-f	na	0.656	0.971	1.035	0.832	7.4545	3.00%
optimized-fe	na	0.743	1.029	0.453	0.536	5.7164	2.04%

[a] Optimized radii and comparison parameter sets scaled by scaling factor, x_{best} [Å], that was fit with forces (-f) or forces and energies (-fe) according to $R_{new} = R_{orig} + x$. [b] Trpcage and G-peptide forces used in force and force−energy training sets. [c] IFABP forces used in IFABP force optimization training set. [d] Dipeptides and polyalanine peptides used in force−energy training set. [e] Dipeptide and polyalanine peptides *not* used in training set. [f] Four proteinlike peptides *not* used in training sets.

conformations and across unique amino acid architectures to be RMSD_$\bar{F}^p \approx 0.29$ kcal/mol Å and that of force−energy-optimized radii to be $\Delta\Delta G_p^{solv} \approx 0.05$ kcal/mol. Our optimized radii show errors just above these limits for both forces and solvation energies (Table 3).

Finally, the best-fit forces do not necessarily result in accurate solvation energies. In fact, the solvation energies are overestimated by the individually force-optimized radii by between 0.45 and 3.81 kcal/mol! This apparent contradiction to the theoretical framework presented in section II is caused by a fundamental difference in the implicit and explicit solvent potentials. The relationship presented in eq 10 *is* exact *within* a given solvent model, and it can be used to fit the implicit potential of mean force to the explicit potential of mean force, as we are attempting to do here. However, fitting the directional derivative (i.e., force) of the potential of mean force from one solvent model to the other does not guarantee that the change in free energy will be equally well-fit. In fact, the approximate, at best, agreement between implicit and explicit solvent potentials of mean force for salt bridges and hydrogen bonds[51,52] suggests that more accurate forces would necessitate less accurate energies. The results shown in Tables 3 and 4 support this hypothesis.

Protein Radii Optimizations. Similar to the dipeptides, the solution space for the force- and force−energy-based optimizations was also highly frustrated with many similarly well-fit solutions. The various different starting radii (AMBER *ff*99 vdW, Bondi, and our previously optimized radii) had no distinguishable effect on the final solutions. However, the choice of starting radii did affect the speed of the optimizations, with the latter two sets being the most efficient. Increasing the number of target values generally decreases the frustration in the solution space. Thus, using 3N atomic solvation forces as opposed to one solvation energy for an N-atom solute makes the optimization substantially more efficient and decreases the likelihood that the final solution is locally trapped far from the global minimum. The force-based optimizations reached complete convergence; that is, the same solution set was found for five consecutive generations, in a single evolution of 20−80 generations. The force−energy-based optimizations never reached complete convergence but converged to a minimum fitness deviation (ΔRMSD_$\bar{F}^p \leq 0.002$) in one to two evolutions of ∼65 generations. Both the force- and force−energy-based optimizations were considerably faster than our previous optimizations, which used solvation energies alone and took six or more evolutions to reach a minimum fitness convergence.

Ideally, both continuum forces and energies would be faithfully predicted by a single parameter set. However, the discrepancy in implicit and explicit potentials demonstrated by the dipeptides suggests that the accuracy of one or the other must be sacrificed. This was indeed the case for the protein radii as shown in Table 4. The force-optimized radii performed 5.6−11.7% better on forces, while the force−energy-optimized radii perform 35.6−56.2% better on polar solvation energies. Depending on the particular problem at hand, accuracy in individual forces or overall energies may be more important; thus, both sets are presented in Table 1. However, we recommend the force−energy-optimized radii for general applications where accuracy in both forces and energies is important. The force-optimized radii presented in column 2 of Table 1 are those from Trpcage and G-peptide optimizations. As described below, radii obtained from these two systems provide much better global accuracy than those fit from IFABP simulations.

Effects of Interstitial High Dielectrics. The two model systems used in the force-based optimizations resulted in strikingly different solution sets. The IFABP radii were 1−10% larger than the Trpcage and G-peptide radii and underestimated solvation energies by 1−26% (Table 4). Close inspection of the IFABP atomic forces that were being significantly over- and underestimated revealed the source of this discrepancy: the significant outliers were atoms found

near interstitial high dielectrics (IHDs). IHDs are regions which have a higher dielectric value (i.e., are solvent) in atom-centered dielectric functions but a low dielectric value (i.e., are solute) in molecular surfaces. They are, by definition, too small for an explicit water molecule to penetrate. The dielectric constant in these regions is the same as the solvent's ($\epsilon \approx 80$) in a van der Waals surface (the envelope surrounding unsmoothed atomic spheres) but can take on a range of values ($1 < \epsilon \leq 80$) in spline-smoothed surfaces. IHDs can be identified by comparing molecular and spline-smoothed surface topologies. As described in the Methods, it was necessary to scale down our spline-smoothed radii (Table 1) in order to define a realistic molecular surface that could be compared to our optimal spline-smoothed surfaces. The scaling factors and solvation energies for molecular surfaces are presented in the Supporting Information. Using unscaled spline-smoothed radii would place the molecular surface too far from the solute, overestimating the solute volume and, hence, overestimating the volume of IHDs. The optimal scaling factor for the force-optimized radii was $x_{best} = -0.275$ Å.

Figure 2 shows a series of slices through the dielectric map of IFABP. The magnitude of IHDs, demonstrated by comparing the molecular surface (panel A) and the spline-smoothed surface (panel B) through a single plane of intersection, is consistent through the entire structure. Both surfaces are rendered in panels C−E to show three examples of atoms with significantly overestimated continuum forces: tyr70 OH, ser109 OG, and glu85 N are next to IHDs with volumes of 41.06, 20.44, and 13.47 Å^3, respectively. This overestimation occurs regardless of the parameter set used to define each atom's radius; it is 3.6, 4.5, and 2.9 kcal/mol/Å and 2.8, 1.9, and 3.7 kcal/mol/Å for the Trpcage/G-peptide and IFABP force-optimized radii, respectively. Moreover, analyses of the explicit solvent simulations demonstrate that these IHDs have explicit solvent occupancies of 0.431, 0.33, and 0.0 in regions with dielectric values above 70 ($70 \leq \epsilon \leq 78.4$) and 0.0 in all regions less than 70 ($10 < \epsilon < 70$), indicating that they *are* mostly buried and *should* have low dielectric values.

Figure 3 shows that the percentage of IFABP atoms with large force deviations increases as the volume of IHDs within 2.0 Å of the atom's radial boundary increases. Forces from all eight IFABP conformations were qualified as "large" if they were greater than a given cutoff value; this value was allowed to vary to test sensitivity (see Figure 3). Although the plotted data were generated with the force-optimized radii from Table 1, other radii yielded very similar results. This trend lends support to the notion that the force deviations are related to the presence of IHDs. Figure 4 shows that IFABP has a larger distribution of force deviations than Trpcage and G-peptide, when defined by either the force-optimized or scaled Bondi radii. The same trend was found for all of the tested parameter sets, as indicated by consistently larger RMSD_$\bar{F}^p$ values for IFABP than Trpcage and G-peptide (Table 4). This suggests that IHDs introduce errors in the entire protein, not just in proximal atoms.

IFABP is a highly solvated protein with a significant number of solvent-exposed residues in "buried" fatty acid binding sites. It is more susceptible to the formation of IHDs than Trpcage or G-peptide, which have compact structures.

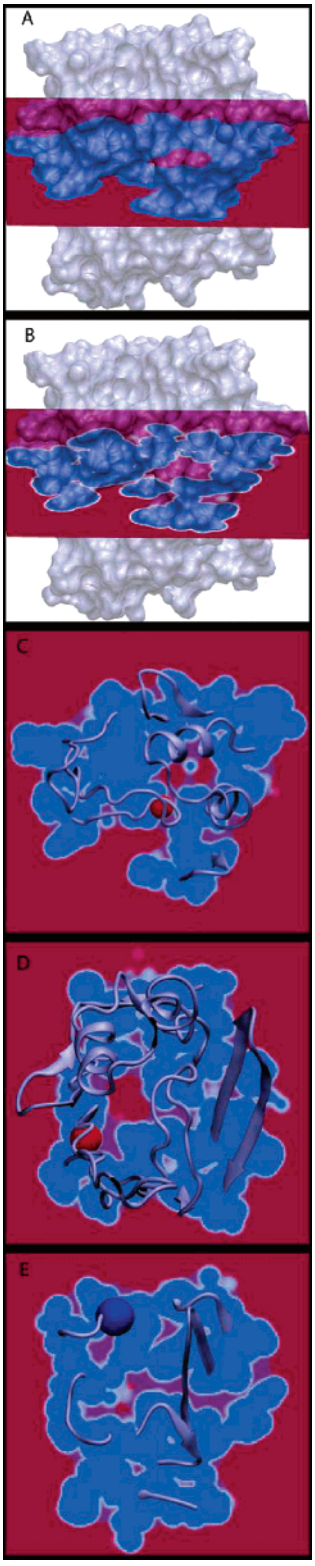


Figure 2. Dielectric maps of IFABP. Planes intersecting IFABP high dielectric ($\epsilon = 80$) regions in red, low dielectric ($\epsilon = 1$) in blue, and values between in white. Comparing the molecular surface (A) and the spline smoothed surface (B) shows the magnitude of IHDs created by the latter. A top-down view of both surfaces shows that IHDs (grey regions) are proximal to (C) tyr 70 OH, (D) ser109 OG, and (E) glu 85 N.

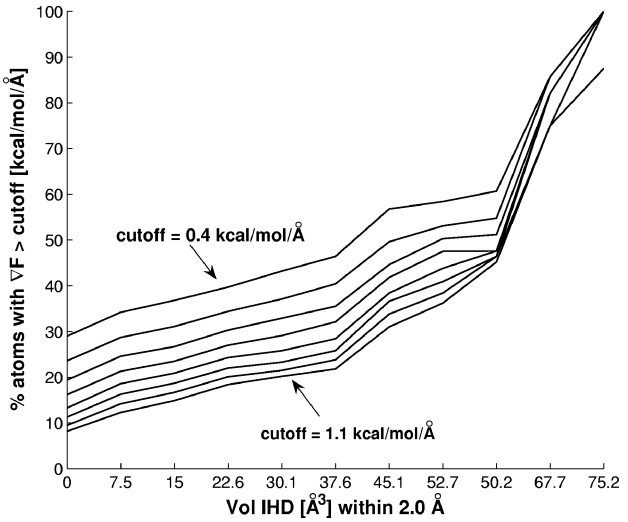


Figure 3. Percentage of IFABP atoms with large force deviations as a function of the volume of IHD ($\epsilon > 10$) within 2.0 Å of the atom's radial boundary. Deviation cutoff values ranged from 0.4 to 1.1 kcal/mol Å. All eight IFABP conformations were used with the force-optimized radii (scaled down for the molecular surface used to identify the IHDs).

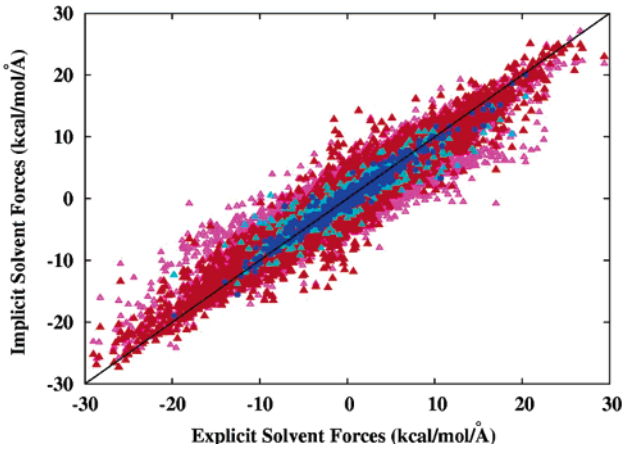


Figure 4. Comparison of explicit solvent and implicit solvent forces for IFABP and Trpcage/G-peptide showing a larger distribution of deviations for IFABP. Bondi radii are in pink (IFABP) and turquoise (Trpcage/G-peptide), while force optimized radii are in red (IFABP) and dark blue (Trpcage/G-peptide).

These findings point to a simple explanation for the different force-optimized solution sets: the IFABP-optimized radii were systematically increased to compensate for errors caused by large IHDs. The Trpcage and G-peptide-optimized radii were less affected; consequently, they perform better on polar solvation energies and forces for every tested molecule other than IFABP. They will be more widely applicable to compact systems but will not compensate for errors resulting from excessively large IHDs in globular solutes.

In retrospect, we believe that the formation of IHDs was also the source of differences between our previously presented "sharp-boundary" and spline-smoothed radii.[23] The sharp-boundary radii, fit to molecular surfaces, could not be scaled to perform as well on spline-smoothed surfaces as the spline-smoothed radii. Moreover, the spline-smoothed radii overestimated solvation energies for larger peptide systems because they were fit to small peptides. It is now clear that this overestimation was most likely caused by the spline-smoothed surfaces affecting the dielectric definition and resulting solvation forces of large solutes much more than those of small peptides.

It has been suggested that the width of the spline-smoothing window, *w*, can be increased to counteract the deleterious effects of IHDs.[19] Although it is not possible to eliminate them completely, their magnitude and dielectric value *can* be significantly decreased with larger smoothing windows. Unfortunately, this moves the dielectric boundary too far from surface atoms and creates unphysical bulges around overlapping and adjacent atoms, both of which result in underestimated solvation energies and forces. If it were possible to distinguish surface and buried atoms, then using larger radii or larger smoothing windows on buried atoms could significantly reduce the IHDs and related errors. For most globular systems, however, a clean distinction is impossible because many atoms are both exposed and buried. Thus, more involved modifications of the spline-smoothed surfaces, such as those proposed by Lu and Luo[53] and Lee et al.,[54] seem necessary for consistently accurate forces and energies. Despite these limitations, the surprisingly accurate solvation energies and forces in Table 4 suggest that spline-smoothed surfaces work quite well for compact systems.

Comparison to Other Parameter Sets. To gauge the accuracy of our optimized radii, we compared them to four other continuum parameter sets commonly used in conjunction with the AMBER force field. These included the AMBER *ff*99 charges combined with *ff*99 vdW radii,[30] Bondi radii,[31] and a set of radii recently published by Luo et al,[28] as well as PARSE charges combined with two variants of PARSE radii.[13] As previously mentioned, these radii were intended for use with molecular surfaces and so had to be scaled for use with the spline-smoothed surfaces. The scaling factors were determined with both force-only and force−energy-based genetic algorithm optimizations as described above for the radii presented in Table 1. With only one parameter to fit, these optimizations converged in one to five generations. Replica optimizations found the same scaling factors, indicating that the results were independent of the run parameters.

Two variants of the PARSE parameters were tested. They differed only in the value of the nonpolar hydrogen radius. The first set used 1 Å, as intended in the original publication, while the second used 0 Å, the standard value in PARSE parameter implementation. The first set ($R_{\mathrm{Hnp}} = 1$ Å) performed marginally better on both forces and energies, validating Sitkoff's intention for the PARSE radii to be the same as the standard Pauling vdW radii (with the exception of hydrogens which were decreased from 1.2 to 1.0 Å).[13]

The solvation energy deviations for the training set dipeptides and polyalanine peptides with different radii sets are shown in Figure 5. The deviations are not consistent across the various scaled radii, but they are consistently larger than the optimized radii deviations. A similar range of deviations was calculated for the test set peptides (see the

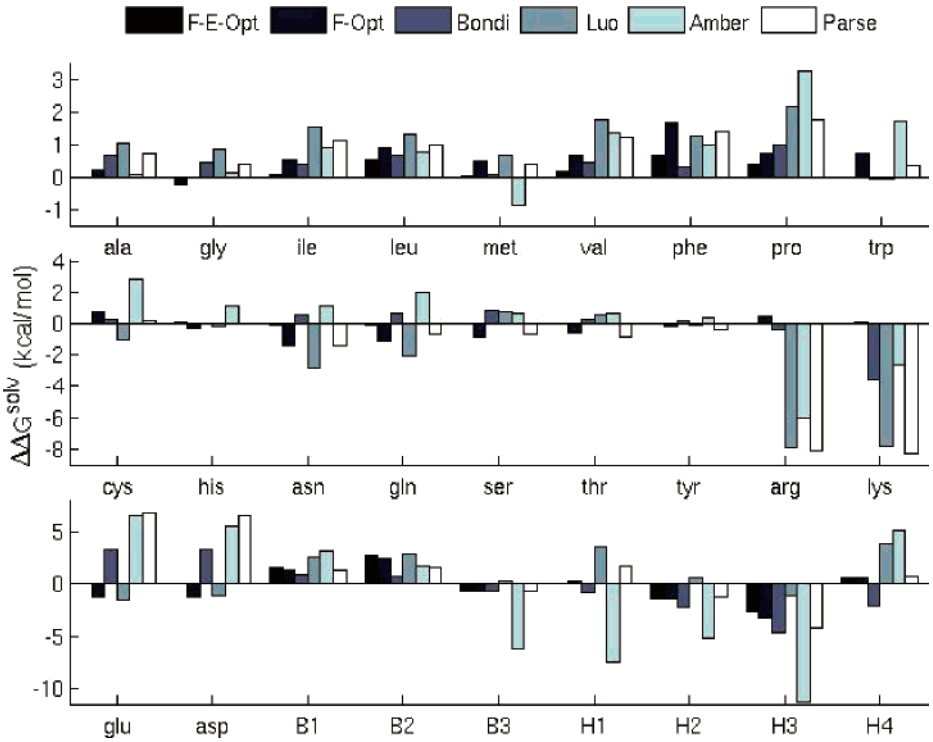


Figure 5. Solvation energy deviations for training set dipepetides (ala, gly, ... asp) and polyalanine peptides (B ≡ β-turn, H ≡ α-helix or β-hairpin) using the (F-E) force and energy optimized radii, (F) force optimized radii, Bondi, Luo, AMBER ff99, and PARSE parameters.

Supporting Information, Figure 7), but the comparison parameter sets (Bondi, Luo, AMBERff99, and PARSE) were inconsistently accurate for the two dipeptide conformations. A summary of the force and energy results is presented in Table 4. The scaling factors for the force and force−energy optimizations were significantly different for each of the parameter sets. Nevertheless, they followed several trends demonstrated by our optimized radii: the force-optimized radii performed better on forces and worse on energies than the force- and energy-optimized radii, the spread in deviations for IFABP were larger than those for Trpcage and G-peptide, and the solvation energy deviations for the amino acid dipeptides and polyalanine peptides in the test set were comparable or less than those in the training set. The only unexpected result was a larger solvation energy deviation for the proteinlike peptides with the Luo force and energy-optimized radii than the Luo force-optimized radii. Out of the comparison parameter sets, the AMBER and Luo radii performed best on forces while the Bondi radii performed best on solvation energies. As expected, our force and force−energy-optimized radii (presented in Table 1) performed better on forces and energies than all of the comparison sets.

Scaled Radii for Different Spline Window Half-Widths and Molecular Surfaces. The spline window half-width used in our optimizations ($w = 0.3$ Å) will not be optimal for every problem or be chosen by every user. It has been shown, however, that increasing or decreasing $w = 0.3$ Å radii by a window half-width-dependent scaling factor recovers accurate solvation energies.[21] As described in the Methods, we determined the best scaling factors, x_{best}, according to eq 14 for a range of spline window half-widths ($w = 0.2$, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0 Å) with force- and force-and-energy-based optimizations. The optimizations converged in one to five generations, and replica optimizations found the same scaling factors, indicating that the results were independent of the run parameters. The final results are presented in Table 5 and plotted in Figure 6. The scaling values for the force-optimized and the force-and-energy-optimized radii are well fit by quadratic polynomials, $x = -0.0518w^2 + 0.5043w - 0.1475$ and $x = -0.2078w^2 + 0.667w - 0.1823$, respectively. These equations can be used to approximate x_{best} (eq 14) for any spline window half-width but will likely yield unreliable results for widths less than 0.1 Å or larger than 1.1 Å.

Table 5. Scaling Factors [x_{best}] for Different Spline Window Half-Widths [w] for (a) Force and (b) Force−Energy-Optimized Radii

	scaling factor x_{best} [Å]	
spline half-width w [Å]	force[a]	force−energy[b]
0.2	−0.051	−0.058
0.3	0.000	0.000
0.4	0.051	0.051
0.5	0.093	0.099
0.6	0.129	0.143
0.7	0.178	0.182
0.8	0.225	0.218
0.9	0.267	0.250
1.0	0.303	0.277

If only solvation energies are desired, molecular surfaces can be used to avoid artifacts introduced by IHDs in spline-smoothed surfaces. Although our optimized radii are not ideal, they can be scaled *down* for molecular surfaces. This

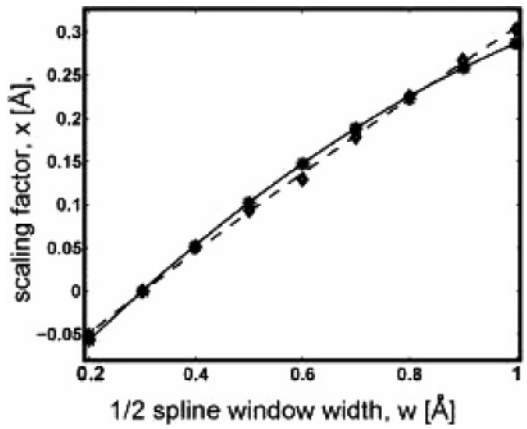


Figure 6. Scaling factors for different spline window half-widths. The force results (diamonds) are best fit by $x = -0.0518w^2 + 0.5043w - 0.1475$ (dashed). The force−energy results (stars) are best fit by $x = -0.2078w^2 + 0.667w - 0.1823$ (solid).

is analogous to scaling the comparison parameter set's molecular surface radii *up* (i.e., made larger) for spline-smoothed surfaces. As described in the Methods, we used eq 14 to scale our force-only and our force-and-energy-optimized radii (Table 1) for both abrupt and harmonically smoothed molecular surfaces. The scaling values and resulting solvation energies (see Supporting Information) indicate that these radii work well for molecular surfaces but not as well as they work for spline-smoothed surfaces.

V. Conclusions

We have presented two sets of radii optimized to define spline-smoothed dielectric boundaries in Poisson electrostatics calculations on proteins defined with AMBER *ff*99 partial charges. These optimizations were based on explicit solvent simulations and, thus, offer forces and energies consistent with explicit solvent mean polar forces and charging free energies. More importantly, we have presented a method of optimizing the Poisson dielectric boundary on the basis of atomic forces, which is substantially more efficient and widely applicable than previous optimization approaches. It can be used, for example, on highly charged, large, or unusual systems for which experimental solvation energies are unavailable and explicit solvent charging free energies are computationally challenging or prohibitive.

We have shown that optimizations based on forces alone generate radii that reproduce forces accurately but solvation energies less accurately. Optimizations based on both forces and energies, on the other hand, decrease the accuracy of forces minimally and increase solvation energy accuracy. Thus, our force−energy-optimized radii are more appropriate than our force-optimized radii when both energies and forces are important. As anticipated, our optimized radii perform better than four other parameter sets commonly used in conjunction with the AMBER force field, each of which was scaled for use with spline-smoothed dielectric boundaries. Of these comparison sets, the AMBER and Luo radii perform best on forces while the Bondi radii perform best on energies.

Finally, we have shown that atom-centered dielectric functions, such as the spline-smoothed surface used in our optimizations, form high dielectric regions in interstitial spaces, thereby limiting the accuracy of solvation forces and energies. These errors seem less pronounced in compact systems, where the IHDs fill the crevices between atoms, than in globular systems, where buried IHDs can be copious. Thus, our optimized radii, and spline-smoothed surfaces in general, are expected to work well for compact systems but will produce errors in highly solvated globular proteins. Efforts to correct atom-centered, smooth surfaces will be crucial for the continued development of Poisson-based implicit solvent models, especially in the area of continuum dynamics.

It is important, however, to appreciate the complexity of the problem at hand. The specific nature of the optimal dielectric boundary most rigorously depends on the locally defined solute−solvent potential,[34,35] a complex function of solute geometry (i.e., surface curvature) and electrostatic and nonelectrostatic solvent−solute interactions. Thus, theory and our analysis agree that an optimal continuum radius cannot be defined by atom type alone. Consequently, we anticipate that corrected smooth surfaces will reduce inaccuracies and inconsistencies but that it will take a more physically accurate implicit solvent framework which accounts for polar and nonpolar coupling to predict accurate solvation forces and energies for disparate solutes and solute conformations without system- or conformation-specific parameters.

Acknowledgment. J.M.J.S thanks John Mongan for invaluable discussion and the NSF-sponsored Center for Theoretical Biological Physics for financial support (Grant Nos. PHY-0216576 and PHY-0225630). N.A.B. and J.A.W. were supported by NIH Grant GM069702 and an Alfred P. Sloan Research Fellowship (N.A.B.). Work in the McCammon group is additionally supported by NIH, NSF, HHMI, NBCR, and Accelrys Inc.

Supporting Information Available: Additional figures of test set continuum solvation energy deviations and molecular surface results. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Roux, B.; Simonson, T. Implicit Solvent Models. *Biophys. Chem.* **1999**, *78*, 1−20.

(2) Rashin, A. A.; Honig, B. Reevaluation of the Born Model of Ion Hydration. *J. Phys. Chem.* **1985**, *89*, 5588−5593.

(3) Roux, B.; Yu, H. A.; Karplus, M. Molecular-Basis for the Born Model of Ion Solvation. *J. Phys. Chem.* **1990**, *94*, 4683−4688.

(4) Gallicchio, E.; Kubo, M. M.; Levy, R. M. Enthalpy=Entropy and Cavity Decomposition of Alkane Hydration Free Energies: Numerical Results and Implications for Theories of Hydrophobic Solvation. *J. Phys. Chem.* **2000**, *104*, 6271−6285.

(5) Wagoner, J. A.; Baker, N. A. Assessing Implicit Models for Nonpolar Mean Solvation Forces: The Importance of Dispersion and Volume Terms. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8331−8336.

(6) Tanford, C.; Kirkwood, J. G. Theory of Protein Titration Curces. I. General Equations for Impenetratable Spheres. *J. Am. Chem. Soc.* **1957**, *79*, 5333−5339.

(7) Chothia, C. Hydrophobic Bonding and Accessible Surface Area in Proteins. *Nature (London)* **1974**, *248*, 338−339.

(8) Spolar, R. S.; Ha, J. H.; Record, M. T., Jr. Hydrophobic Effect in Protein Folding and Other Noncovalent Processes Involving Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 8382−8385.

(9) Liu, T.; Ye, L.; Chen, H.; Li, J.; Wu, Z.; Zhou, R. A Combined Steepest Descent and Genetic Algorithm (SD/GA) Approach for the Optimization of Solvation Parameters. *Mol. Simul.* **2006**, *32*, 427−435.

(10) Davis, M. E.; McCammon, J. A. Electrostatics in Biomolecular Structure and Dynamics. *Chem. Rev.* **1990**, *90*, 509−521.

(11) Honig, B.; Nicholls, A. Classical Electrostatics in Biology and Chemistry. *Science* **1995**, *268*, 1144−1149.

(12) Baker, N. A. Improving Implicit Solvent Simulations: A Poisson-Centric View. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137−143.

(13) Sitkoff, D.; Sharp, K.; Honig, B. Accurate Calculation of Hydration Free Energies Using Macroscopic Solbent Models. *J. Phys. Chem.* **1994**, *98*, 1978−1988.

(14) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. On the Nonpolar Hydration Free Energy of Proteins: Surface Area and Continuum Solvent Models for the Solute−Solvent Interaction Energy. *J. Am. Chem. Soc.* **2003**, *125*, 9523−9530.

(15) Lee, B.; Richards, F. M. Interpretation of Protein Structures − Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55*, 379.

(16) Connolly, M. L. Computation of Molecular Volume. *J. Am. Chem. Soc.* **1985**, *107*, 1118−1124.

(17) Bordner, A. J.; Cavasotto, C. N.; Anagyan, R. A. Accurate Transferable Model for Water, *n*-Octanol, and *n*-Hexadecane Solcation Free Energies. *J. Phys. Chem.* **2002**, *106*, 11009−11015.

(18) Nina, M.; Simonson, T. Molecular Dynamics of the tRNA-(Ala) Acceptor Stem: Comparison between Continuum Reaction Field and Particle-Mesh Ewald Electrostatic Treatments. *J. Phys. Chem. B* **2002**, *106*, 3696−3705.

(19) Grant, J. A.; Pickup, B. T.; Nicholls, A. A Smooth Permittivity Function for Poisson−Boltzmann Solvation Methods. *J. Comput. Chem.* **2001**, *22*, 608−640.

(20) Im, W.; Beglov, D.; Roux, B. Continuum Solvation Model: Computation of Electrostatic Forces from Numerical Solutions to the Poisson−Boltzmann Equation. *Comput. Phys. Commun.* **1998**, *111*, 59−75.

(21) Nina, M.; Im, W.; Roux, B. Optimized Atomic Radii for Protein Continuum Electrostatics Solvation Forces. *Biophys. Chem.* **1999**, *78*, 89−96.

(22) Wagoner, J.; Baker, N. A. Solvation Forces on Biomolecular Structures: A Comparison of Explicit Solvent and Poisson−Boltzmann Models. *J. Comput. Chem.* **2004**, *25*, 1623−1629.

(23) Swanson, J. M. J.; Adcock, S. A.; McCammon, J. A. Optimized Radii for Poisson−Boltzmann Calculations with the AMBER Force Field. *J. Chem. Theory Comput.* **2005**, *1*, 484−493.

(24) Izvekov, S.; Parrinello, M.; Burnham, C. J.; Voth, G. A. Effective Force Fields for Condensed Phase Systems from ab Initio Molecular Dynamics Simulation: A New Method for Force-Matching. *J. Chem. Phys.* **2004**, *120*, 10896−10913.

(25) Izvekov, S.; Voth, G. A. A Multiscale Coarse-Graining Method for Biomolecular Systems. *J. Phys. Chem. B* **2005**, *109*, 2469−2473.

(26) Cramer, C. J.; Truhlar, D. G. General Parameterized Scf Model for Free-Energies of Solvation in Aqueous-Solution. *J. Am. Chem. Soc.* **1991**, *113*, 8305−8311.

(27) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. SM6: A Density Functional Theory Continuum Solvation Model for Calculating Aqueous Solvation Free Energies of Neutrals, Ions, and Solute-Water Clusters. *J. Chem. Theory Comput.* **2005**, *1*, 1133−1152.

(28) Lwin, T. Z.; Zhou, R. H.; Luo, R. Is Poisson−Boltzmann Theory Insufficient for Protein Folding Simulations? *J. Chem. Phys.* **2006**, *124*.

(29) Nina, M.; Beglov, D.; Roux, B. Atomic Radii for Continuum Electrostatics Calculations Based on Molecular Dynamics Free Energy Simulations. *J. Phys. Chem. B* **1997**, *101*, 5239−5248.

(30) Wang, J. M.; Cieplak, P.; Kollman, P. A. How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? *J. Comput. Chem.* **2000**, *21*, 1049−1074.

(31) Bondi, A. *J. Chem. Phys.* **1964**, *64*, 441.

(32) Ben-Naim, A. *Solvation Thermodynamics*; Plenum Press: New York, 1987.

(33) McQuarrie, D. A. *Statistical Mechanics*; Harper and Row: London, 1976.

(34) Dzubiella, J.; Swanson, J. M. J.; McCammon, J. A. Coupling Nonpolar and Polar Solvation Free Energies in Implicit Solvent Models. *J. Chem. Phys.* **2006**, *124*.

(35) Dzubiella, J.; Swanson, J. M. J.; McCammon, J. A. Coupling Hydrophobicity, Dispersion, and Electrostatics in Continuum Solvent Models. *Phys. Rev. Lett.* **2006**, *96*.

(36) Straatsma, T. P.; McCammon, J. A. Computational Alchemy. *Annu. Rev. Phys. Chem.* **1992**, *43*, 407−435.

(37) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. Designing a 20-Residue Protein. *Nat. Struct. Biol.* **2002**, *9*, 425−430.

(38) Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. A Novel, Highly Stable Fold of the Immunoglobulin Binding Domain of Streptococcal Protein-G. *Science* **1991**, *253*, 657−661.

(39) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. Amber, A Package of Computer-Programs for Applying Molecular Mechanics, Normal-Mode Analysis, Molecular-Dynamics and Free-Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Comput. Phys. Commun.* **1995**, *91*, 1−41.

(40) Hodsdon, M. E.; Cistola, D. P. Ligand Binding Alters the Backbone Mobility of Intestinal Fatty Acid-Binding Protein as Monitored by N-15 NMR Relaxation and H-1 Exchange. *Biochemistry* **1997**, *36*, 2278−2290.

(41) Lindahl, E.; Hess, B.; vand der Spoel, D. GROMACS 3.0: A Package for Molecular Simulation and Trajectory Analysis. *J. Mol. Model.* **2001**, *7*, 306−317.

(42) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. Charmm − A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(43) Beglov, D.; Roux, B. Finite Representation of an Infinite Bulk System − Solvent Boundary Potential for Computer-Simulations. *J. Chem. Phys.* **1994**, *100*, 9050−9063.

(44) Deng, Y. Q.; Roux, B. Hydration of Amino Acid Side Chains: Nonpolar and Electrostatic Contributions Calculated from Staged Molecular Dynamics Free Energy Simulations with Explicit Water Molecules. *J. Phys. Chem. B* **2004**, *108*, 16567−16576.

(45) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of Nanosystems: Application to Microtubules and the Ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037−10041.

(46) Price, D. J.; Brooks, C. L. A Modified TIP3P Water Potential for Simulation with Ewald Summation. *J. Chem. Phys.* **2004**, *121*, 10096−10103.

(47) Hochtl, P.; Boresch, S.; Bitomsky, W.; Steinhauser, O. Rationalization of the Dielectric Properties of Common Three-Site Water Models in Terms of Their Force Field Parameters. *J. Chem. Phys.* **1998**, *109*, 927−4937.

(48) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926−935.

(49) Baker, J. E. Reducing Bias and Inefficiency in the Selection Algorithm. In *Proceedings of the Second International Conference on Genetic Algorithms on Genetic Algorithms and their Application*; Lawrence Erlbaum Associates, Inc.: Mahwah, NJ, 1987; pp 14−21.

(50) Onufriev, A.; Case, D. A.; Bashford, D. Effective Born Radii in the Generalized Born Approximation: The Importance of Being Perfect. *J. Comput. Chem.* **2002**, *23*, 1297−1304.

(51) Masunov, A.; Lazaridis, T. Potentials of Mean Force between Ionizable Amino Acid Side Chains in Water. *J. Am. Chem. Soc.* **2003**, *125*, 1722−1730.

(52) Swanson, J. M. J.; Mongan, J.; McCammon, J. A. Limitations of Atom-Centered Dielectric Functions in Implicit Solvent Models. *J. Phys. Chem. B* **2005**, *109*, 14769−14772.

(53) Lu, Q.; Luo, R. A Poisson−Boltzmann Dynamics Method with Nonperiodic Boundary Condition. *J. Chem. Phys.* **2003**, *119*, 11035−11047.

(54) Lee, M. S.; Feig, M.; Salsbury, F. R.; Brooks, C. L. New Analytic Approximation to the Standard Molecular Volume Definition and Its Application to Generalized Born Calculations (vol 24, pg 1348, 2003). *J. Comput. Chem.* **2003**, *24*, 1348−1356.

CT600216K

JCTC Journal of Chemical Theory and Computation

Structures and Electronic States of Permethyloligosilane Radical Ions with All-Trans Form $Si_n(CH_3)_{2n+2}^{\pm}$ ($n = 2-6$): A Density Functional Theory Study

Hiroto Tachikawa*

Division of Molecular Chemistry, Graduate School of Engineering, Hokkaido University, Sapporo 060−8628, Japan

Hiroshi Kawabata

Venture Business Laboratory, Kyoto University, Kyoto 606−8501, Japan

Received May 9, 2006

Abstract: Hybrid density functional theory (DFT) calculations have been carried out for the radical cation and anion of permethyloligosilane $Si_n(CH_3)_{2n+2}^{\pm}$ ($n = 2-6$) to elucidate the electronic structures at ground and low-lying excited states, and the results were compared with the corresponding experimental values. In particular, the assignment of electronic transition appeared at near-IR and visible regions, which is strongly correlated to hole and electron conductivity, and was carried out on the basis of time-dependent DFT calculation. The structure of oligosilane was generated at 300 K by direct PM3 molecular dynamics calculations, and then the geometry was fully optimized at the DFT(B3LYP)/6-311+G(d,p) level. It was found that the hole in the radical cation and the electron in the radical anion of oligosilane are delocalized over the Si skeleton. The proton-hyperfine coupling constants calculated were in good agreement with those obtained by an electron spin resonance experiment. It was also found that the g-anisotropy of the radical anion was significantly larger than that of the radical cation. The IR bands of radical ions were assigned on the basis of theoretical calculations.

1. Introduction

Polysilane is a one-dimensional σ-conjugated polymer composed of Si−Si single bond and organic side groups.[1,2] This polymer has recently received much attention because of its potential ability as a hole and electron transport material in organic multilayer light emitting diodes, a photoresistor in lithography, and a one-dimensional semiconductor.[3−10]

To elucidate the mechanism of the electron and hole transfer processes, experimental and theoretical works for the radical cation and anion of poly- and oligosilanes have been extensively examined. Ban et al. measured transient absorption spectra of polysilanes in organic solvents after electron beam irradiation and found that the radical ions (cation and anion) of polysilanes show a strong absorption band at the UV region.[11,12] They assigned this strong peak to an electronic transition of unpaired electrons on the Si−Si skeleton. Irie et al.[13,14] measured the transient and steady-state absorption spectra of radical ions of polysilanes with aryl groups and found both UV and near-IR bands. They assigned the near-IR band to a charge resonance band between two aryl groups. Ushida et al.[15] attributed the near-IR band to charge resonance between adjacent σ-conjugated polymer segments. However, the electronic structure of radical ions of polysilanes, especially the origin of the near-IR band, is still in controversy. The low-lying excited states

* Corresponding author fax: +81 11706-7897; e-mail: hiroto@eng.hokudai.ac.jp.

10.1021/ct600163r CCC: $37.00
Published on Web 11/09/2006

in the radical ion are strongly correlated with electron (or hole) conductivity. In particular, the near-IR band of radical ions of polysilane reflects directly the ability of conductivity.

To elucidate more detailed features for the electronic states of polysilanes, an approach to use the oligomer of polysilane has been carried out by several groups. Irie et al.[16] examined chain-length dependence on the absorption spectra of the radical anion and cation of oligosilanes. The spectra observed have two peaks in the UV and near-IR regions as well as those of polysilanes, and the maxima in both spectra shifted to longer wavelengths with increasing chain length. Kumagai et al.[17,18] investigated absorption spectra of radical ions of permethyl oligosilanes. They found that the radical ions possess UV and near-IR bands, as well as those of usual polysilanes.

The magnetic properties of the radical ions have been investigated from hyperfine-coupling constants (hfcc's) and **g**-tensor components using electron spin resonance (ESR) measurements.[17,18] The ESR spectra of the radical cation of oligo- and polysilanes give a broad symmetric single line with averaged g values ranging from 2.0070 to 2.0100, which is slightly larger than that of the free electron ($g_e = 2.0023$). On the other hand, ESR spectra of the radical anion of polysilanes give an axially symmetric line shape. The radical anion of oligosilane also gives an ESR spectrum with g values in the range 2.000−2.0040. The g anisotropy decreases with increasing chain length (n).

The hfcc's of radical ions of oligosilanes have been measured by Kumagai et al.[17,18] They observed ESR spectra of radical ions of permethyloligosilanes, $Si_n(CH_3)_{2n+2}$ where n ranges from 2 to 6. For $n = 2$, the smallest radical cation, the hfcc of the proton is measured to be 0.57 mT. The hfcc's decreased with increasing n: for example, hfcc's for $n = 4$ and 6 are 0.55 and 0.19 mT ($n = 4$) and 0.39, 0.29, and 0.11 mT ($n = 6$).

The electronic states of neutral oligosilanes have been investigated mainly by ab initio calculations.[19,20] Michl and Ottosson investigated conformers of n-$Si_6(CH_3)_{14}$ using ab initio, molecular mechanics, and additive increment methods.[19a] It was found that all-transoid is the most stable conformer. Recently, the conformational dependence of $Si_4(CH_3)_{10}$ on UV absorption has been investigated using singly excited configuration interaction (SE-CI) calculations.[19b] All-transoid is the most stable as well as $n = 6$. The excitation energies of $\sigma\sigma^*$ and $\sigma\pi^*$ are not dependent on conformational change, but oscillator strengths are strongly affected. More recently, the three conformers of n-$Si_4(CH_3)_{10}$ have been investigated using the complete active space−self consistent field method, and it has been shown that the excitation energies and ionization potentials are slightly affected by the conformational change. Also, the effects of conformations in the neutral state are clearly elucidated by their calculations. Thus, information on the electronic structures of neutral oligosilanes have been accumulated from theoretical points of view. On the other hand, few theoretical works have been carried out for radical ions of oligosilanes despite their importance.

In previous papers, we investigated the structures and electronic states of the polysilane radical anion[21] and cation[22] by means of the semiempirical third-parametric configuration interaction (PM3-CI) method. The permethyl oligosilane [$Si_n(Me)_{2n+2}$] ($n = 8-20$) with the all-trans form was chosen as a model of polysilane. We showed that excess electrons and holes are fully delocalized along the skeleton of linear polysilane if the structure has the regular all-trans form. The electronic excitation energy from the ground to first exited states was gradually red-shifted as a function of the number of chain Si atoms (n). Our previous studies suggested that the excess electron (hole) is not localized in the case of the regular all-trans form in the linear polysilane but is delocalized widely along the main chain. Also, the mechanism of electron localization at finite temperatures was investigated by means of MM2 molecular dynamics and PM3 methods.[23]

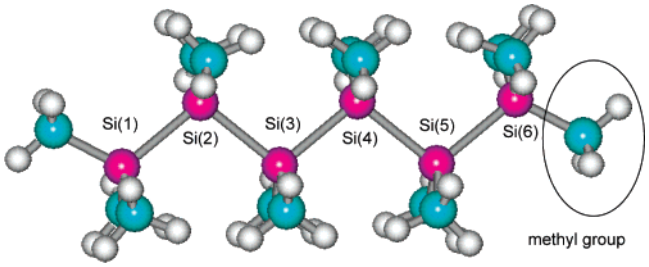


Figure 1. Schematic illustration of structure of permethyloligosilane, $Si_n(CH_3)_{2n+2}$ ($n = 6$), with all-trans form. Hydrogens in the methyl group are omitted from the figure.

More recently, Tada and Yoshimura have carried out SE-CI calculations for the radical anion of oligosilane Si_nH_{2n+2} ($n = 2-6$).[24] The ground state is composed of σ^* characteristics where an unpaired electron occupies the σ^* orbital of the Si−Si skeleton. The first electronic transition of $(Si_nH_{2n+2})^-$ ($n = 2-6$) is assigned to the Si−Si($\sigma^* \leftarrow \sigma^*$) transition. The π character appears from the third excited state. However, the radical anion $(Si_nH_{2n+2})^-$ has been not observed experimentally, and the conclusion may be therefore limited only to the $(Si_nH_{2n+2})^-$ system. Actual systems observed experimentally include alkyl groups (such as methyl and ethyl groups) in the side chain. Hence, studying of the electronic structures of methyl-substituted oligosilane is required to compare directly with the experiments.

In the present study, to interpret the experimental data obtained by Kumagai et al. from a theoretical point of view, density functional theory (DFT) calculations have been carried out for neutral permethyloligosilane and the radical cation and anion of permethyloligosilane with the all-trans form $Si_n(CH_3)_{2n+2}$ ($n = 2-6$). In particular, we focus our attention mainly on the assignment of the near-IR band and ESR spectra of the radical cation.

2. Method of Calculations

A linear oligosilane with methyl groups in the side chain, permethyl oligosilane [$Si_n(CH_3)_{2n+2}$; $n = 2-6$], was examined in the present work. The structure and atom numbers for $n = 6$ are illustrated in Figure 1. The structure of oligosilane for the neutral state was generated by direct PM3 molecular dynamics (MD) calculation[25] at 300 K. And then, the geometry was further optimized at the DFT(B3LYP)/6-311+G(d,p) level. The structures of the radical ions of oligosilane were fully optimized from the neutral structure at the B3LYP/6-311+G(d,p) level of theory. When the

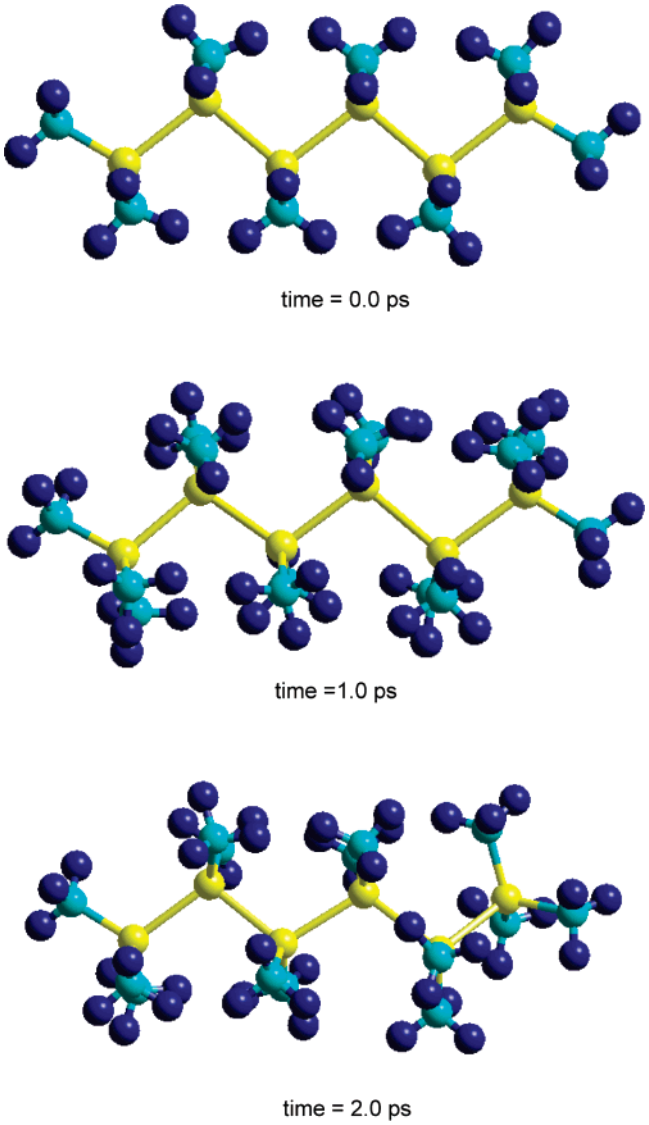


Figure 2. Snapshots of structural conformations of $Si_6(CH_3)_{14}$ obtained by direct PM3-MD calculation at 300 K.

optimized geometries were used, the excitation energies were calculated by means of time-dependent (TD)-DFT calculations at the B3LYP/6-311+G(d,p) level. Twelve excited states were solved in TD-DFT calculations. All hybrid DFT calculations were carried out using the Gaussian 03 program package.[26] Note that this level of theory gives reasonable features for several molecular device systems.[27−29]

3. Results

A. Structures of Permethyl Oligosilanes. First, the structure of oligosilane with the regular all-trans form is fully optimized by the semiempirical third-parametric molecular orbital (PM3-MO) method. Second, direct PM3 MD calculation[25] is carried out from the optimized structure to elucidate the effects of rotation of the methyl group (position of hydrogen atoms of the CH_3 group) on the structure and energetics. The mean temperature used in the MD calculation is 300 K. Results of the direct PM3-MD calculation for n = 6 are given in Figures2 and 3. The snapshots of oligosilane show that the structure of oligosilane is slightly deformed from the regular all-trans form by thermal activation. All

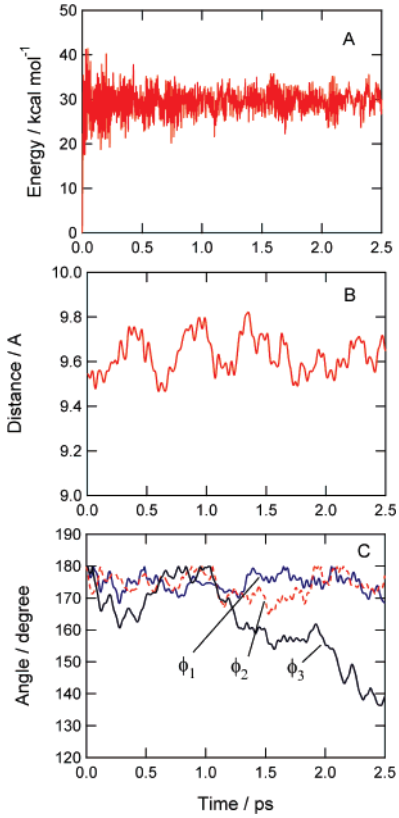


Figure 3. Results of direct PM3-MD calculation of $Si_6(CH_3)_{14}$ at 300 K. (A) Potential energy of the system, (B) the head-to-tail distance of the Si−Si main chain (in A), and (C) the dihedral angles.

dihedral angles are $\phi_1 = \phi_2 = \phi_3 = 180.0°$ at time zero, where the dihedral angles are defined as $\phi_1 = \angle$Si(1)−Si(2)−Si(3)−Si(4), $\phi_2 = \angle$Si(2)−Si(3)−Si(4)−Si(5), and $\phi_3 = \angle$Si(3)−Si(4)−Si(5)−Si(6). These angles are gradually changed as a function of time: $\phi_1 = 174.6°$, $\phi_2 = 177.3°$, and $\phi_3 = 179.5°$ at time = 1.0 ps and $\phi_1 = 174.0°$, $\phi_2 = 175.4°$, and $\phi_3 = 155.5°$ at 2.0 ps. However, the change of the main chain is not large. Instead, the rotation of hydrogen atoms of the methyl group is mainly enhanced by thermal activation.

Figure 3A−C show a potential energy of the system (n = 6), a distance between the head and tail silicon atoms in the main chain of oligosilane (defined by R), and the dihedral angles, respectively. The average of the potential energy becomes almost constant after time = 0.30 ps. The chain length (R) is gradually varied as a function of time (Figure 3B). However, the amplitude of R is only 0.32 Å, indicating that deformation of the main chain has hardly occurred at 300 K. The dihedral angles are plotted in Figure 3C. The angles are varied in the range 140−180°. However, large conformational changes, namely, rotation of the Si−Si bond, does not occur at 300 K. Only the rotation of the methyl group takes place at 300 K.

Next, a total number of 17 structures in the time range 1.0−2.5 ps are selected from the results of direct PM3-MD calculation, and then the geometries are fully optimized at the HF/3-21G(d) level. Initial and final energies of $Si_6(CH_3)_{14}$ before and after geometry optimization are summarized in Table 1. Relative energy is calculated on the basis of the

Table 1. Initial and Relaxed Energies before and after Geometry Optimization at Selected Points Obtained by Direct PM3-MD Calculation at 300 K[a]

selected point	time ps	initial total energy au	relaxed total energy au	initial energy[b] kcal/mol	final energy[b] kcal/mol
0	0.0	−2276.670 15	−2276.725 40	34.67	0.00
1	1.0	−2276.622 61	−2276.725 40	64.50	0.00
2	1.1	−2276.621 95	−2276.725 40	64.91	0.00
3	1.2	−2276.619 36	−2276.725 40	66.54	0.00
4	1.3	−2276.633 89	−2276.725 40	57.42	0.00
5	1.4	−2276.627 34	−2276.725 40	61.53	0.00
6	1.5	−2276.636 63	−2276.725 40	55.71	0.00
7	1.6	−2276.617 25	−2276.725 40	67.86	0.00
8	1.7	−2276.624 62	−2276.725 40	63.24	0.00
9	1.8	−2276.618 69	−2276.725 40	66.96	0.00
10	1.9	−2276.633 54	−2276.725 40	57.64	0.00
11	2.0	−2276.622 72	−2276.725 40	64.44	0.00
12	2.1	−2276.615 58	−2276.725 40	68.91	0.00
13	2.2	−2276.603 35	−2276.724 42	76.59	0.62
14	2.3	−2276.620 84	−2276.725 40	65.61	0.00
15	2.4	−2276.614 53	−2276.725 40	69.57	0.00
16	2.5	−2276.614 97	−2276.725 40	69.29	0.00

[a] The initial and relaxed total energies (in au) are calculated at the HF/3-21G(d) level. The geometry optimization is carried out at the HF/3-21G(d) level. [b] Zero-level corresponds to the energy level of the optimized structure at time zero.

total energy of the optimized geometry at time zero [HF/3-21G(d) level]. The initial energies are lying 35−77 kcal/mol above the zero level. After the geometry optimization, two energy minima are obtained. The geometry optimizations from the selected points give the same structure, except for one result for the selected point 13 in Table 1. This result indicated that thermal activation causes rotation of the methyl groups of the oligosilanes, but the positions of hydrogen atoms are a secondary matter. The geometry optimizations from these points give a unique structure with the all-trans form. The most stable structure is further optimized at the B3LYP/6-311+G(d,p) level.

The optimized parameters are given in Table 2. The position of the Si atom is defined in Figure 1. The geometry optimization gives no-symmetry structures for all oligosilanes ($n = 2-6$). The Si skeleton of the oligomer is not in plane, and the dihedral angles of Si−Si−Si−Si are close to 160° at the B3LYP/6-311+G(d,p) level. The Si−Si bond length is close to 3.8 Å for all neutral oligosilanes. For example, the bond lengths for $n = 3$ are calculated to be 2.3790 Å, which is significantly close to the previous theoretical value (2.3533 Å at the MP2/cc-pVTZ level)[30] and the experimental value (2.325 Å).[31]

For the radical cation, the Si−Si bond lengths are longer than those of the neutral state. This elongation is due to the fact that an electron is removed from a Si−Si σ bond of the oligomer. The structures of radical anions are also fully optimized at the same level of theory. The structures for $n = 2-4$ are hardly changed by accepting an excess electron. On the other hand, the bond lengths of $n = 5$ and 6 are largely elongated. For example, a Si−Si bond in the central position for $n = 6$, Si(3)−Si(4), is elongated from 2.3887 to 2.4426 Å.

Band gaps of neutral oligosilanes ($n = 2-6$), estimated by the energy difference between the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO), decrease gradually from 6.42 to 5.38 eV as chain lengths are increased from $n = 2$ to 6. The adiabatic ionization potential (without zero-point energy, ZPE) decreases gradually with increasing n: 7.80 eV ($n = 2$) and 6.78 eV ($n = 6$). Adiabatic electron affinities (without ZPE) are varied from −0.58 eV ($n = 2$) to −0.15 eV ($n = 6$).

B. Excitation Energies of Neutral Oligosilanes ($n = 2-6$). The low-lying excited state correlates strongly with the electron conductivity in organic semiconductors. Hence, the electronic structures for the excited states are determined in this system. The excitation energies calculated for the neutral oligosilanes with the transoid form are given in Table 3. The first electronic transition for $n = 2$ occurs at 6.15 eV with an oscillator strength $f = 0.053$. A stronger electronic transition appears at 6.30 eV with $f = 0.123$. The first electronic excitation band with the large oscillator strength is gradually red-shifted with increasing chain length (n): for example, the excitation energies are 5.38 eV ($n = 3$), 5.30 eV ($n = 4$), 4.98 eV ($n = 5$), and 4.77 eV ($n = 6$). It is also shown that the first electronic transitions are nearly degenerated in $n = 3-6$. The absorption band is assigned to the HOMO(σ) → LUMO(σ^*) transition for $n = 3-6$, but that for $n = 2$ is a $\sigma-\pi^*$ transition. For $n = 2$, the lowest excited state is calculated to be 5.73 eV ($f = 0.0$), which is assigned to a Rydberg state. This is caused by underestimation of the excitation energy due to a diffuse orbital on the hydrogen atoms. The assignment of the excitation essentially agrees with previous theoretical work by Rooklin et al.[32]

To check the reliability of the level of theory used in the present calculation, the excitation energies are compared with the experiments. The results are given in Table 4. The experimental excitation energies measured at 77 K in 2-methyltetrahydrofuran matrixes for $n = 4$, 5, and 6 are 5.28, 4.96, and 4.77 eV, respectively. The corresponding

Table 2. Optimized Geometrical Parameters for Neutral Molecule and the Radical Cation and Radical Anion of Oligosilanes (Transoid) Calculated at the B3LYP/6-311+G(d,p) Level

		Bond Length				
state	*n*	$R(Si^1-Si^2)$	$R(Si^2-Si^3)$	$R(Si^3-Si^4)$	$R(Si^4-Si^5)$	$R(Si^5-Si^6)$
neutral	2	2.3741				
	3	2.3790	2.3790			
	4	2.3801	2.3803	2.3801		
	5	2.3787	2.3853	2.3853	2.3787	
	6	2.3821	2.3862	2.3887	2.3862	2.3821
cation	2	2.7103				
	3	2.5022	2.5022			
	4	2.4772	2.4273	2.4772		
	5	2.4434	2.4343	2.4343	2.4434	
	6	2.4288	2.4269	2.4409	2.4269	2.4287
anion	2	2.3680				
	3	2.3796	2.3796			
	4	2.3789	2.3832	2.3789		
	5	2.3991	2.4458	2.4458	2.3991	
	6	2.3917	2.4217	2.4426	2.4217	2.3917

		Dihedral Angle		
state	*n*	ϕ_1	ϕ_2	ϕ_3
neutral	4	165.2		
	5	165.0	165.0	
	6	168.2	171.5	168.2
cation	4	165.2		
	5	163.9	163.9	
	6	165.1	163.3	165.1
anion	4	180.0		
	5	180.0	180.0	
	6	180.0	180.0	172.6

Table 3. Excitation Energies (in eV) and Oscillator Strengths (Given in Parentheses) of Neutral Oligosilanes with Transoid Form ($n = 2-6$) Calculated at the B3LYP/6-311+G(d,p) Level

state	$n = 2$	3	4	5	6
1^{st}	5.73 (0.0000)σR[a]	5.38 (0.0373)$\sigma\sigma^*$	5.20 (0.0005)$\sigma\pi^*$	4.98 (0.2750)$\sigma\sigma^*$	4.77 (0.6173)$\sigma\sigma^*$
2^{nd}	6.15 (0.0529)$\sigma\pi^*$	5.57 (0.0000)$\sigma\pi^*$	5.22 (0.0039)$\sigma\pi^*$	4.99 (0.000)σR	4.87 (0.0009)$\sigma\pi^*$
3^{rd}	6.15 (0.0529)$\pi\pi^*$	5.75 (0.0134)$\sigma\sigma^*$	5.30 (0.1977)$\sigma\sigma^*$	5.10 (0.0729)$\sigma\pi^*$	4.96 (0.0002)σR
4^{th}	6.30 (0.1226)$\sigma\sigma^*$	5.92 (0.0057)$\sigma\sigma^*$	5.58 (0.0523)$\sigma\sigma^*$	5.34 (0.0054)$\sigma\pi^*$	5.20 (0.0201)$\sigma\pi^*$
Rooklin[b]	6.76	5.95	5.48	5.23	4.96
Obtara[c]	6.44	5.73	5.28	4.96	4.77

[a] σ-Rydberg transition. [b] Theoretical value by Rooklin et al. cited from ref 32. [c] Experimental value by Obara and Kira cited from ref 33.

calculated excitation energies are 5.30 ($n = 4$), 4.98 ($n = 5$), and 4.77 eV ($n = 6$). The excitation energies measured in an argon matrix at 10 K are 6.63 ($n = 2$) and 5.93 eV ($n = 3$), which are also in reasonable agreement with the present calculations, 6.15 ($n = 2$) and 5.38 eV ($n = 3$), although the calculation underestimates slightly the excitation energies for shorter oligomers. The agreement with the experiments indicates that the TD-DFT(B3LYP)/6-311+G(d,p) calculation gives reasonable electronic transition energy for the oligosilane system.

C. Electronic States of Radical Cations. 1. Spin Density and g Value. Spin densities on protons of the methyl groups in the radical cation are given in Table 5. The value is averaged on each Si segment $-Si(m)(CH_3)_2-$ (where $m = 1-6$). For $n = 6$, spin densities on segment Si(m) where $m = 1-6$ are 0.138, 0.148, 0.214, 0.214, 0.148, and 0.138, respectively, indicating that an unpaired electron (or hole) is delocalized on the Si skeleton.

In general, ESR spectra of the radical cation of oligosilane show a broad single band without a hyperfine structure except for $n = 2$. Only the radical cation for $n = 2$ has a spectrum with a hyperfine structure. The hyperfine coupling constant of the hydrogen atom (H-hfcc) for $n = 2$ has been measured directly to be 0.57 mT.[18] The H-hfcc for $n = 2$ is calculated to be 0.552 mT in the present study, which is in good agreement with the ESR experiment. For $n = 4-6$, there is

Table 4. Comparison of Theoretical and Experimental Excitation Energies (E_{ex} in eV) and Wavelengths (λ_{max} in nm) at Maximum Peak Positions of Neutral Oligosilanes $Si_n(CH_3)_{2n+2}$ ($n = 2-6$)

	theoretical		experimental[a]	
n	λ_{max}	E_{ex}	λ_{max}	E_{ex}
2	202	6.15	187	6.63[a]
3	230	5.38	209	5.93[a]
4	234	5.30	235	5.28[b]
(4a,4g)[c]				(5.44,5.96)[a]
5	249	4.98	250	4.96[b]
(5aa,5ga,5gg)[c]				(5.05,5.39,5.94)[a]
6	260	4.77	260	4.77[b]

[a] In argon matrix at 10 K cited from ref 19. [b] In 2-methyltetrahydrofuran (MTHF) at 77 K cited from ref 16. [c] Notations "a" and "g" mean "anti" and "gauche", respectively.

Table 5. Proton-Hyperfine Coupling Constants (H-hfcc's) of the Radical Cation of Oligosilane (in mT) and Spin Densities on Each Subsegment of $Si(m)(Me)_2$ ($m = 1-6$) Calculated at the B3LYP/6-311+G(d,p) Level

n	state	Si(1)	Si(2)	Si(3)	Si(4)	Si(5)	Si(6)
2	hfcc	0.552	0.552				
	spin	0.50	0.50				
3	hfcc	0.362	0.531	0.362			
	spin	0.336	0.329	0.336			
4	hfcc	0.259	0.339	0.339	0.259		
	spin	0.241	0.259	0.259	0.241		
5	hfcc	0.186	0.262	0.278	0.262	0.186	
	spin	0.181	0.186	0.266	0.186	0.181	
6	hfcc	0.137	0.197	0.243	0.243	0.197	0.137
	spin	0.138	0.148	0.214	0.214	0.148	0.138

Table 6. **g**-Tensor Components of the Radical Cation of Oligosilanes Calculated at the B3LYP/6-311+G(d,p) Level

n	g_{xx}	g_{yy}	g_{zz}
2	2.0023	2.0102	2.0102
3	2.0033	2.0095	2.0173
4	2.0029	2.0094	2.0184
5	2.0028	2.0095	2.0191
6	2.0027	2.0097	2.0191

no direct experimental value because of its broad single spectrum, while simulated hfcc's were estimated: for example, H-hfcc's for $n = 6$ are estimated to be 0.11, 0.29, and 0.39 mT from the simulation. The calculated values are 0.137, 0.197, and 0.243 mT, which are in reasonable agreement with the simulated values.

Isosurfaces of spin densities for $n = 3-6$ are illustrated in Figure 4. It is shown that the spin−orbital is composed of a σ(Si−Si) orbital and an unpaired electron that is widely distributed over the oligomer.

A **g**-tensor component gives important information on the electronic structure of the paramagnetic compound at the ground state. The values of oligosilanes ($R = CH_3$) calculated at the B3LYP/6-311+G(d,p) level are given in Table 6. It is found that the anisotropy of the **g**-value for the radical cation of oligosilane is significantly small, and those are close to that of the free electron ($g_e = 2.0023$). For $n = 2$, the

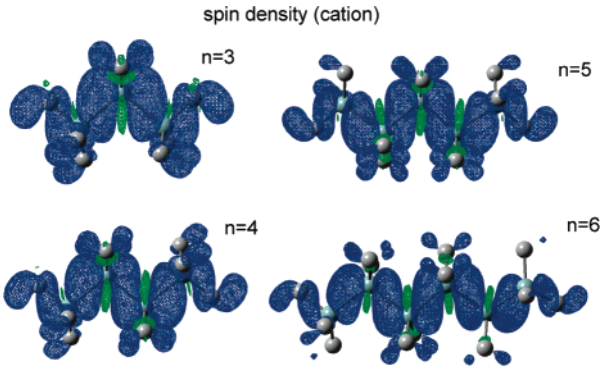


Figure 4. Illustrations of spin densities of radical cations of oligosilanes ($n = 3-6$) at the ground state.

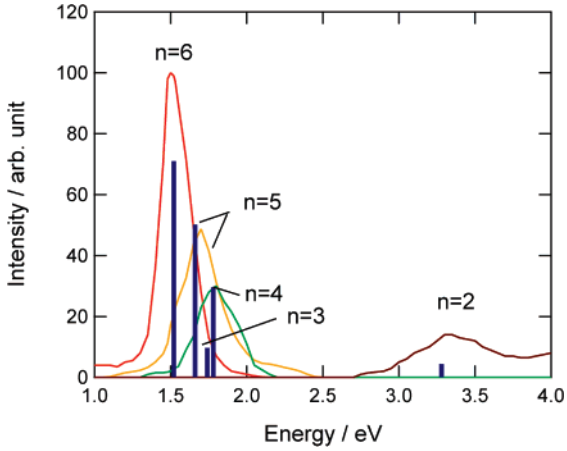


Figure 5. Calculated (stick diagram) and experimental absorption spectra (solid curve) of radical cations of oligosilane ($n = 3-6$). Experimental data are cited from ref 18.

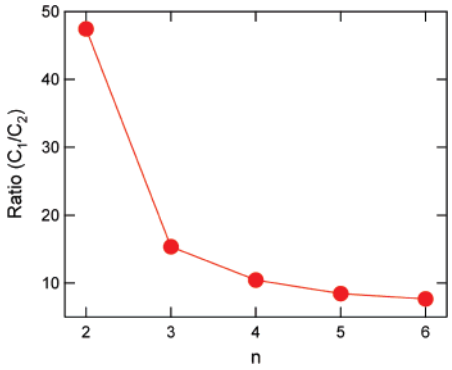


Figure 6. Ratio of coefficients (C_1/C_2) of reference functions (C_1 and C_2) as a function of *n*.

averaged *g* value is $\langle g \rangle = 2.0076$, which is in good agreement with the experimental value [$\langle g \rangle = 2.008$ measured for $Si_2(CH_3)_6^+$ at 77 K].[18] The calculation shows that the averaged *g* value increases slightly with increasing *n*, but the magnitude of the increase is small: $\langle g \rangle = 2.0076$ ($n = 2$), 2.0100 ($n = 3$), 2.0102 ($n = 4$), 2.0105 ($n = 5$), and 2.0105 ($n = 6$). These values agree well with those of the popular radical cations of poly- and oligosilanes: for example, radical cations of poly(cyclohexylmethylsilane) and poly(methylphenylsilane) have $\langle g \rangle = 2.008$, which is larger than that of the free electron (g_e).[18]

Table 7. Excitation Energies (in eV) and Oscillator Strengths (Given in Parentheses) of the Radical Cation of Oligosilanes, $[Si_n(CH_3)_{2n+2}]^+$ ($n = 2-6$), Calculated at B3LYP/6-311+G(d,p) Level[a]

state	$n = 2$	3	4	5	6
1st	3.05 (0.0000)	*1.74 (0.0524)*	*1.78 (0.1647)*	*1.66 (0.2788)*	*1.52 (0.3944)*
2nd	3.06 (0.0000)	2.75 (0.0005)	2.08 (0.0006)	2.08 (0.0001)	2.03 (0.0005)
3rd	*3.28 (0.0124)*	2.89 (0.0018)	2.86 (0.0033)	2.19 (0.0127)	2.20 (0.0084)
4th	*3.28 (0.0124)*	2.89 (0.0130)	2.88 (0.0025)	2.86 (0.0068)	2.27 (0.0004)
5th	5.07 (0.0000)	2.97 (0.0389)	2.94 (0.0451)	2.87 (0.0002)	2.86 (0.0017)
6th	5.24 (0.0000)	3.16 (0.0000)	3.00 (0.0000)	2.96 (0.0263)	2.86 (0.0023)

[a] Italic characters mean the first excitation energies with a nonzero oscillator strength.

Table 8. Proton-Hyperfine Coupling Constants (hfcc's) of the Radical Anion of Oligosilane (in mT) and Spin Densities on Each Subsegment of Si(m)(Me)$_2$ ($m = 1-6$) Calculated at the B3LYP/6-311+G(d,p) Level

n	state	Si(1)	Si(2)	Si(3)	Si(4)	Si(5)	Si(6)
2	hfcc	0.044	0.044				
	spin	0.500	0.500				
3	hfcc	0.030	0.036	0.030			
	spin	1.115	−1.229	1.115			
4	hfcc	0.027	0.026	0.026	0.027		
	spin	0.739	−0.240	−0.239	0.739		
5	hfcc	0.064	0.043	0.046	0.043	0.064	
	spin	0.187	0.206	0.213	0.206	0.187	
6	hfcc	0.049	0.031	0.041	0.041	0.031	0.049
	spin	0.141	0.155	0.204	0.204	0.155	0.141

Table 9. **g**-Tensor Components of the Radical Anion of Oligosilanes Calculated at the B3LYP/6-311+G(d,p) Level[a]

		g_{xx}	g_{yy}	g_{zz}
calcd (R = −H)	2	2.0019	2.0032	2.0051
	3	2.0021	2.0031	2.0059
	4	2.0019	2.0031	2.0068
	5	2.0019	2.0031	2.0075
	6	2.0018	2.0031	2.0081
calcd (R = $-CH_3$)	2	2.0025	2.0027	2.0027
	3	2.0026	2.0027	2.0027
	4	2.0006	2.0020	2.0063
	5	2.0006	2.0020	2.0064
	6	2.0002	2.0021	2.0064
exptl (R = $-CH_3$)[a]	4	1.9979	2.0037	2.0082
	5	1.9992	2.0032	2.0061
	6[b]	2.0032	2.0032	2.0032

[a] Experimental **g** tensors are cited from ref 17. [b] Experimental values are $g_{xx} \approx g_{yy} \approx g_{zz}$.

C.2. Excitation Energy. Figure 5 shows the first excitation energies and its oscillator strengths calculated for the radical cation of $Si_n(CH_3)_{2n+2}$ ($n = 2-6$) together with experimental absorption spectra measured by Kumagai et al.[18] Unfortunately, the absorption spectrum for $n = 3$ has not been observed yet. Hence, the experimental spectra for $n = 2$ and 4−6 are depicted in the figure. The present calculation represents excellently the absorption spectra derived from the experiment. This agreement implies that the level of theory is adequate enough to discuss the electronic states of the radical cation.

The low-lying excitation energies calculated for the radical cation are given in Table 7. The first electronic transition for $n = 2$ is calculated to lie at 3.28 eV with an oscillator strength $f = 0.0124$. The excitation energy becomes significantly lower in the case of a longer oligosilane, $n = 3$ (1.74 eV), and also the oscillator strength becomes larger ($f = 0.0524$) in $n = 3$. The excitation energy and oscillator strength for $n = 6$ are 1.522 eV and $f = 0.3944$, respectively. In larger oligomers, the excitation energy is gradually red-shifted and the oscillator strength increases with increasing n.

We found that the first excited state with a nonzero oscillator strength for the radical cation (electronic transition corresponds to the near-IR band) is composed of two electronic states, ψ_1(HOMO → SOMO) and ψ_2(SOMO → SOMO + m), where HOMO and SOMO mean the highest doubly occupied and singly occupied molecular orbitals of the radical cation, respectively, and m means an integer. The first excited state is expressed by

$$\Psi\,(1\mathrm{st}) = C_1\psi_1(\mathrm{HOMO} \rightarrow \mathrm{SOMO}) + C_2\psi_2(\mathrm{SOMO} \rightarrow \mathrm{LUMO}) + \ldots.$$

For $n = 2$, the coefficients C_1 and C_2 are calculated to be 0.997 and −0.021, respectively, indicating that the main configuration is ψ_1(HOMO → SOMO) at the first excited state. For $n = 6$, the coefficients are significantly changed to $C_1 = 0.901$ and $C_2 = -0.118$. Figure 6 shows absolute values of the ratio of C_1 and C_2 plotted as a function of n. The ratio reaches a limited value in larger n values. For $n = 6$, the weight of ψ_1(HOMO → SOMO) is calculated to be 0.81. This result means that the near-IR band in poly- and oligosilanes is assigned to an electronic transition ψ_0(SOMO) → ψ_1(HOMO → SOMO). Also, it can be expected that the UV band appearing at 250−350 nm is caused by electronic transitions from SOMO → LUMO and HOMO → LUMO + m. It can be concluded that the strong band of the radical cation is assigned to the valence transition within the Si−Si σ bond.

D. Electronic States of Radical Anions. 1. Spin Density and g Value. Spin densities of radical anions of oligosilanes calculated are given in Table 8, and isosurfaces of spin densities for $n = 5$ and 6 are illustrated in Figure 7. The unpaired electron is delocalized over the Si skeleton of oligosilane, while electron localization is not found in the range $n = 2-6$. For example, the spin densities on the Si subunits for $SiMe_3$(1), $SiMe_2$(2), $SiMe_2$(3), $SiMe_2$(4), $SiMe_2$-(5), and $SiMe_3$(6) are calculated to be 0.141, 0.155, 0.204, 0.204, 0.155, and 0.141, respectively. H-hfcc's for $n = 6$ are distributed in the rage 0.031−0.049 mT, which are significantly smaller than those of the radical cation. This is due to the fact that the unpaired electron is located on a σ*-

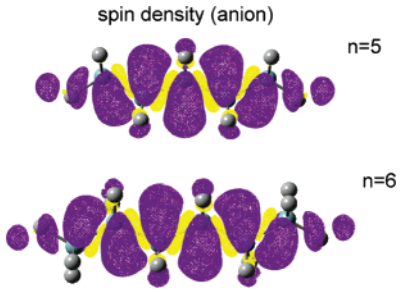


Figure 7. Illustrations of spin densities of radical anions of oligosilanes ($n = 5$ and 6) at the ground state.

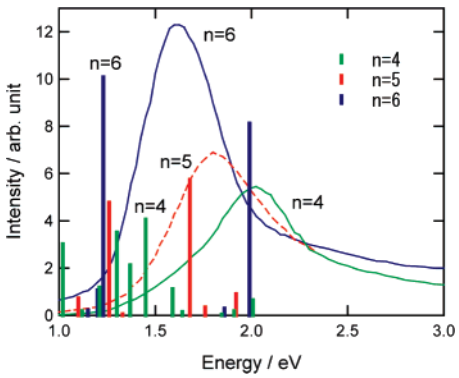


Figure 8. Calculated (stick diagram) and experimental absorption spectra (solid curve) of the radical anions of oligosilane ($n = 4-6$).

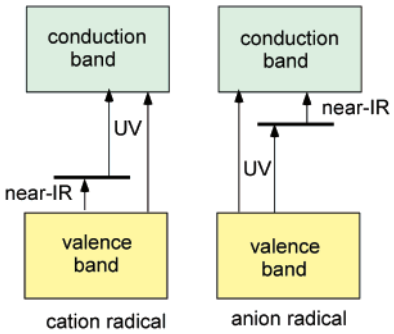


Figure 9. Schematic illustration of band structures of radical ions of oligosilanes.

(Si−Si) orbital mixed with a diffuse-type orbital in the radical anion, so that spin density in the 1s orbital of the proton is significantly diminished.

The **g**-tensor components of radical anions of oligosilanes (R = −H and $-CH_3$) are given in Table 9. The calculation shows that the component g_{xx} is slightly lower than that of the free electron ($g_e = 2.0023$), whereas g_{zz} is larger than g_e. On the other hand, g_{yy} is close to g_e. These features are in reasonable agreement with experimental values: for example, the g values for $n = 5$ are measured to be $g_{xx} = 1.9992$ ($< g_e$), $g_{yy} = 2.0032$ ($= g_e$), and $g_{zz} = 2.0061$ ($> g_e$). For R = −H and $-CH_3$, similar spectral features are obtained.

D.2. Excitation Energy. The low-lying excitation energies for the radical anion of oligosilane are given in Table 10. The first electronic transition for $n = 6$ is calculated in the range 0.69−1.20 eV with an oscillator strength $f = 0.0038-0.0803$. From the TD-DFT calculation, it is found that the near-IR band is assigned to the (HOMO − m) → SOMO (where $m = 0-10$) transition. The strong band of the radical anion is assigned to the valence transition within the Si−Si σ bond. Figure 8 shows the excitation energies and its oscillator strengths calculated for the radical anion of $Si_n(CH_3)_{2n+2}$ ($n = 4-6$) together with experimental absorption spectra. It seems that there is disagreement between theory and experiment in the radical anion: the excitation energies calculated are underestimated in all cases. This discrepancy may be caused by the contribution of two- and three-electron excitations to the excited-state wave function.

4. Discussion

A. Summary of the Present Study. In the present study, hybrid DFT calculations have been carried out for the neutral and radical ions of oligosilane, $Si_n(CH_3)_{2n+2}$ ($n = 2-6$), to elucidate the electronic structures of radical ions of oligosilane. In both radical ions, holes and excess electrons are delocalized along the Si chain at the ground state. These characteristics are in good agreement with previous theoretical[21−24] and experimental characteristics.[13,14,16−18]

For the radical cation, the unpaired electron (hole) is occupied in the σ(Si−Si) orbital, whereas the excess electron in the radical anion is occupied in the σ^*(Si−Si) orbital mixing slightly with a π state of the Si−C bond and a diffuselike orbital. The magnetic properties obtained by the present calculations, g values and hyperfine hfcc's, are in good agreement with experiments.[17,18]

The first excited state for the radical cation is composed mainly of the ψ(HOMO → SOMO) state (weight of reference is about 0.80), and the ψ(SOMO → LUMO) state is contributed to the first excited state. The calculations indicate that this state appears as a near-IR band. Also, the other low-lying excited sites of the radical cation consist of ψ(HOMO − m → SOMO) states (where m means an integer), but the excitation to these states has a small oscillator strength. It is also expected that the UV band

Table 10. Excitation Energies (in eV) and Oscillator Strengths (Given in Parentheses) of the Radical Anion of Oligosilanes, $[Si_n(CH_3)_{2n+2}]^-$ ($n = 2-6$), Calculated at the B3LYP/6-311+G(d,p) Level

state	$n = 2$	3	4	5	6
1st	0.49 (0.1630)	0.40 (0.1186)	0.30 (0.0811)	0.71 (0.0027)	0.69 (0.0803)
2nd	0.49 (0.1626)	0.47 (0.1716)	0.32 (0.1353)	0.75 (0.0994)	0.76 (0.0000)
3rd	0.54 (0.2831)	0.51 (0.2928)	0.39 (0.2084)	0.85 (0.0036)	0.86 (0.0000)
4th	1.18 (0.0000)	0.87 (0.0242)	0.72 (0.0001)	1.10 (0.0128)	1.11 (0.0011)
5th	1.18 (0.0000)	0.92 (0.0095)	0.75 (0.0010)	1.23 (0.0000)	1.15 (0.0038)
6th	1.34 (0.0000)	1.00 (0.0033)	0.79 (0.0010)	1.26 (0.0838)	1.20 (0.0188)

appearing at 250−350 nm is caused by electronic transitions from SOMO to LUMO and HOMO to LUMO + *m*.

B. Band Structure of Radical Ions of Oligosilanes. From the characteristics of the coefficients (C_1 and C_2) and the ratio (C_1/C_2), it is found that concept of the electronic states of the radical anion of oligosilane can be expanded to longer oligosilanes and polysilanes. On the basis of the present calculations, the band structures for radical ions of longer oligosilanes are expected and illustrated schematically in Figure 9. The horizontal lines drawn between valence and conduction bands mean cation (left) and anion (right) states of oligo- and polysilanes. The energy level of the cation state of oligosilane is slightly higher than the valence band because the SOMO of the cation state is mainly originated from the HOMO of neutral oligosilane. From the present calculation, it is expected that the near-IR band is assigned to the electronic transition from the valence band to the cation state, although the electronic transition from the cation state to the conduction band is slightly contaminated to the near-IR band. The UV band is attributed to the electronic transition from the valence band to the conduction band, which is energetically perturbed by the cation state.

The band structure of the radical anion of oligo- and polysilanes is also illustrated. The energy level of the anion state is lower than the bottom of the conduction band. The present calculation predicts that electronic transition from the anion state to the valence band is observed as a near-IR band. The UV band is assigned to the excitation from the valence to the anion state and that from the valence to the conduction bands.

C. Remarks. In the present study, DFT and TD-DFT (B3LYP) methods are applied to the neutral molecule and radical ions of oligosilane, $Si_n(CH_3)_{2n+2}$ ($n = 2-6$). The electronic structures, hfcc's, *g* values, and excitation energies for the neutral molecule and radical cation of oligosilanes were in good agreement with the experiments. However, this level of theory gave a poor result for excitation energies of the radical anion. To obtain a reasonable absorption spectrum of the radical anion, a more accurate wave function would be required. Despite the several assumptions introduced here, the results enable us to obtain valuable information on the electronic states of radical ions of oligosilanes.

Acknowledgment. The authors are indebted to the Computer Center at the Institute for Molecular Science (IMS) for the use of the computing facilities. Also, the authors acknowledge partial support from the president's optional budget from Kyoto University. One of the authors (H.T.) acknowledges a partial support from a Grant-in-Aid for Scientific Research (C) from the Japan Society for the Promotion of Science (JSPS).

References

(1) Miller, R. D.; Michl, J. *Chem. Rev.* **1989**, *89*, 1359.

(2) Kepler, R. G.; Zeigler, J. M.; Harrah, L. A.; Kurtz, S. R. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1987**, *35*, 2818.

(3) Suzuki, H.; Meyer, H.; Hoshino, S. *J. Appl. Phys.* **1995**, *78*, 2684.

(4) Hoshino, S.; Suzuki, H. *Appl. Phys. Lett.* **1996**, *69*, 224.

(5) Fujii, A.; Yoshimoto, K.; Yoshida, M.; Ohmori, Y.; Yoshino, K. *Jpn. J. Appl. Phys.* **1995**, *34*, L1365.

(6) Hattori, R.; Sugano, T.; Shirafuji, J.; Fujiki, T. *Jpn. J. Appl. Phys.* **1996**, *35*, L1509.

(7) Suzuki, H.; Hoshino, S.; Yuan, C. H.; Fujiki, M.; Toyoda, S.; Matsumoto, N. *IEEE J. Quantum Electron.* **1998**, *4* (1), 129.

(8) Suzuki, H.; Hoshino, S.; Yuan, C. H.; Fujiki, M.; Toyoda, S.; Matsumoto, N. *Thin Solid Films* **1998**, *331*, 64.

(9) Xu, Y.; Fujino, T.; Watase, S.; Naito, H.; Oka, K.; Dohmaru, T. *Jpn. J. Appl. Phys.* **1999**, *38*, 2609.

(10) Suzuki, H.; Hoshino, S.; Furukawa, K.; Ebata, K.; Yuan, C. H.; Bleyl, I. *Polym. Adv. Technol.* **2000**, *11*, 460.

(11) Ban, H.; Sukegawa, K.; Tagawa, S. *Macromolecules* **1987**, *20*, 1775.

(12) Ban, H.; Sukegawa, K.; Tagawa, S. *Macromolecules* **1988**, *21*, 45.

(13) Irie, S.; Oka, K.; Irie, M. *Macromolecules* **1988**, *21*, 110.

(14) Irie, S.; Irie, M. *Macromolecules* **1992**, *25*, 1766.

(15) Ushida, K.; Kira, A.; Tagawa, S.; Yoshida, Y.; Shibata, H. *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Mater.)* **1992**, *66*, 299.

(16) Irie, S.; Irie, M. *Macromolecules* **1997**, *30*, 7906.

(17) Kumagai, J.; Yoshida, H.; Koizumi, H.; Ichikawa, T. *J. Phys. Chem.* **1994**, *98*, 13117.

(18) Kumagai, J.; Yoshida, H.; Ichikawa, T. *J. Phys. Chem.* **1995**, *99*, 7965.

(19) (a) Ottosson C. H.; Michl, J. *J. Phys. Chem. A* **2000**, *104*, 3367. (b) Tsuji, T.; Michl, J.; Tamao, K. *J. Organomet. Chem.* **2003**, *685*, 9. (c) Plitt, H. S.; Downing, J. W.; Raymond, M. K.; Balaji, V.; Michl, J. *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 1653. (d) Piqueras, M. C.; Crespo, R.; Michl, J. *J. Phys. Chem. A* **2003**, *107*, 4661.

(20) Sharma, A.; Lourderaj, U.; Deepak; Sathyamurthy, N. *J. Phys. Chem. B* **2005**, *109*, 15860.

(21) Tachikawa, H. *Chem. Phys. Lett.* **1997**, *281*, 221.

(22) Tachikawa, H. *Chem. Phys. Lett.* **1997**, *265*, 455

(23) Tachikawa, H. *J. Phys. Chem. A* **1999**, *103*, 2501.

(24) Tada, T.; Yoshimura, R. *J. Phys. Chem. A* **2003**, *107*, 6091.

(25) Tachikawa, H.; Shimizu, A. *J. Phys. Chem. B* **2005**, *109*, 13255.

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

T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision B.04; Gaussian, Inc.: Wallingford, CT, 2004.

(27) Tachikawa, H.; Kawabata, H. *J. Mater. Chem.* **2003**, *13*, 1293

(28) Tachikawa, H.; Kawabata, H. *J. Phys. Chem. B* **2003**, *107*, 1113.

(29) Tachikawa, H.; Kawabata, H. *J. Phys. Chem. B* **2005**, *109*, 3139.

(30) Piqueras, M. C.; Crespo, R. *Mol. Phys.* **2006**, *104*, 1107.

(31) Almenningen, A.; Fjeldberg, T.; Eclwin, H. *J. Mol. Struct.* **1984**, *112*, 239.

(32) Rooklin, D. W.; Schepers, T.; Raymond-Johansson, M. K.; Michl, J. *Photochem. Photobiol. Sci*. **2003**, *2*, 511.

(33) Obata, K.; Kira, M. *Organometallics* **1999**, *18*, 2216.

CT600163R

JCTC Journal of Chemical Theory and Computation

Molecular Thermodynamics of Methane Solvation in *tert*-Butanol−Water Mixtures

Maeng-Eun Lee and Nico F. A. Van der Vegt*

Max Planck Institute for Polymer Research, Ackermannweg 10, D-55128 Mainz, Germany

Received July 7, 2006

Abstract: We studied solvation structure and thermodynamics of methane in mixtures of *tert*-butanol and water using computer simulations. We show that for alcohol mole fractions below 20%, methane is preferentially solvated by hydrated alcohol clusters. Because methane expels water molecules from these clusters, a large endothermic solvent reorganization enthalpy occurs. This process is responsible for the experimentally observed maximum of the heat of methane solvation close to 5% alcohol in the mixture and contributes to a positive entropy change relative to solvation in pure water. Because the structural solvent reorganization enthalpy is enthalpy−entropy compensating, the methane solvation free energy is a smoothly varying function of the alcohol/water solution composition.

1. Introduction

Binary solvent mixtures are used in many applications in chemistry and biology. They are used for example as reaction media to control the rates of chemical reactions, as selective solvents where chemically different groups of a solvated solute each preferentially interact with one of the two solvent components and as media for studying the stability of biomolecules. Mixtures of organic solvents (but also cosolutes such as inorganic salts, sugars, urea, etc.) and water are of particular interest in biochemistry. Proteins are known to lose their stability in concentrated solutions of aqueous urea or guanidinium chloride,[1] whereas they may be stabilized at low temperatures (cryopreservation)[2] in aqueous solutions of disaccharides (e.g., trehalose) or polyols (e.g., ethylene glycol).[3,4] The atomic-scale mechanisms and their relation with physical-chemical properties of the solvent system are in many cases still incompletely understood and are the topic of ongoing discussion in the literature.[5−8]

In this paper we examine the thermodynamic process of methane dissolution in mixtures of water with *tert*-butyl alcohol. In this solvent system, the observed alcohol/water macromiscibility does not hold down to molecular length scales (microimmiscibility), and self-clustering of alcohol and water molecules occurs, in particular, in the water-rich composition regions.[9,10] Available experimental data show that the heat of methane solvation in this mixture passes through a maximum at approximately 5% alcohol in going from pure water to concentrated alcohol solutions (vide infra). Interestingly, this particular alcohol concentration corresponds to another maximum, which is that of the alcohol self-clustering.[10] It is the purpose of this paper to establish an atomic-scale picture that explains the nonmonotonic behavior of the methane solvation thermodynamics. In this context it is interesting to mention the recent study of Özal and Van der Vegt[11] who, based on molecular simulations, made calculations of solvent reorganization contributions to methane solvation entropies and enthalpies in mixtures of dimethyl sulfoxide and water.

2. Computational Details

2.1. *tert*-Butanol, Water, and Methane Models. We used the flexible six-site *tert*-butanol (TBA) model developed by Lee and Van der Vegt[10] (denoted LV-model from now on) in combination with the rigid three-site SPC water model.[12] TBA−TBA and TBA−water intermolecular interactions are modeled with a sum of Lennard-Jones (LJ) terms centered on the CH_3 (united atom methyl), C (central carbon), O (hydroxyl oxygen), and OW (water oxygen) interaction sites as well as Coulombic interactions between static partial electronic charges centered on the C, O, H (hydroxyl hydrogen), OW, and HW (water hydrogen) atoms. Geometric

* Corresponding author e-mail: vdervegt@mpip-mainz.mpg.de.

10.1021/ct600226h CCC: $37.00
Published on Web 12/02/2006

Table 1. Summary of MD Simulations of *tert*-Butanol/Water Mixtures[a]

x_{TBA}	N_{TBA}	N_{H2O}	t_{MD} (ns)	ϵ_{RF}	$V^{\#}$ (nm^3)	V (nm^3)
tert-Butanol(LV-Model)[10]/Water(SPC-Model)[12]						
0	0	1000	100	78.0	31.31	30.68
0.04	74	1776	150	64.8	65.83	65.29
0.06	105	1645	150	64.0	66.68	65.26
0.10	168	1512	100	50.7	72.48	71.73
0.20	275	1100	100	36.1	77.24	76.24
0.30	336	784	100	30.0	77.72	76.70
0.50	432	432	100	18.0	83.19	82.16
0.70	470	200	100	12.7	83.08	81.97
0.90	500	55	100	11.8	84.36	83.04
1.0	500	0	100	11.8	83.10	82.13
tert-Butanol(GROMOS-Model)[13]/Water(SPC-Model)						
0.06	105	1645	100	64.0	67.61	66.69
0.1	168	1512	100	50.7	73.76	72.82
0.2	275	1100	100	36.1	78.97	77.96
tert-Butanol(OPLS-Model)[14]/Water(SPC-Model)						
0.06	105	1645	100	64.0	67.46	66.56
0.10	168	1512	100	50.7	73.54	72.71
0.20	275	1100	100	36.1	78.24	77.51
tert-Butanol(OPLS-Model)/Water(TIP4P-Model)[15]						
0.06	105	1645	100	64.0	66.74	65.76
0.1	168	1512	100	50.7	72.89	71.78
0.2	275	1100	100	36.1	78.03	76.83

[a] The systems were simulated at constant pressure (1 atm) and temperature (298 K). x_{TBA}: TBA mole fraction; N_{TBA}: number of TBA molecules; N_{H2O}: number of water molecules; t_{MD}: total MD run time; ϵ_{RF}: reaction field dielectric permittivity; V: volume of the simulation box; $V^{\#}$: volume of the simulation box including 10 methane solutes.

mean mixing rules were used to describe LJ interactions between chemically different atom types/groups. For more details on the LV force field model we refer to ref 10. To assess model dependencies, we also used the GROMOS[13] and OPLS[14] *tert*-butanol models described in our previous work.[10] The GROMOS TBA-model was combined with the SPC water model, while the OPLS TBA-model was combined with the SPC and TIP4P[15] water models. In combination with the SPC water model, the latter two force fields produce too large alcohol−alcohol and water−water aggregation.[10] Although alternative force fields could possibly alleviate this effect, we use the GROMOS and OPLS force fields to get a better understanding of the relation between solvent structure and solute interaction thermodynamics. Methane was modeled as a united atom LJ-particle with parameters $\sigma = 0.373$ nm and $\epsilon = 1.247$ kJ/mol.[16] Geometric mean mixing rules were used to describe methane−solvent interactions.

2.2. Simulation Details. All simulations were performed using the GROMACS package.[17] Alcohol/water mixtures were studied at the *tert*-butanol mole fractions summarized in Table 1. Details on system sizes and simulation times are also given in Table 1. Geometries of all molecules were kept rigid by applying constraints to the interatomic distances within the molecules, using the SHAKE algorithm[18] with a relative geometric tolerance of 10^{-4}. A twin-range cutoff scheme with 0.8 and 1.4 nm cutoff radii was applied. The nonbonded interactions in the range between these radii were updated every fifth time step. The equations of motion were integrated using the leapfrog algorithm using a time step of 2 fs. A reaction field approximation was used to account for truncation of electrostatic forces beyond the long-range cutoff (1.4 nm). The reaction field relative dielectric permittivities are listed in Table 1. Constant pressure (1 atm) and temperature simulations were performed using the Nose−Hoover thermostat[19,20] and Rahman−Parrinello barostat[21,22] with coupling times $\tau_T = 1.5$ ps and $\tau_P = 2.5$ ps. Simulations were performed at three temperatures: 278, 298, and 318 K.

2.3. Solvation Free Energies, Enthalpies, and Entropies. The solvation free energies are calculated using the Widom test particle insertion method.[23] The solvation free energy ΔG is obtained by evaluating the expression[16]

$$\Delta G = -RT \ln[\langle V e^{-\psi/RT}\rangle_{NPT}/\langle V\rangle_{NPT}] \quad (1)$$

where ψ denotes the methane binding energy with the solvent, R denotes the gas constant, T is the temperature on the Kelvin scale, and V is the volume. The methane binding energy is evaluated at 125 000 random locations in each solvent configuration stored every 1 ps by summing overall methane−solvent LJ interactions. The angular brackets $\langle\cdots\rangle_{NPT}$ denote an averaging over constant pressure and temperature configurations of the solvent. In the simulation the averaging is performed over 100 ns simulation trajectories (Table 1).

The solvation entropy $\Delta S = -(\partial\Delta G/\partial T)_P$ is calculated using a finite-difference assumption

$$\Delta S(T) = -\frac{\Delta G(T+\Delta T) - \Delta G(T-\Delta T)}{2\Delta T} \quad (2)$$

where $T = 298$ K and $\Delta T = 20$ K in our calculations. In eq 2 it is assumed that the solvation heat capacity $\Delta c_P = T(\partial\Delta S/\partial T)_P$ is independent of temperature in the interval $2\Delta T$.[24] The solvation enthalpy (298 K) is obtained from the relation

$$\Delta H = \Delta G + T\Delta S \quad (3)$$

The statistical uncertainty of ΔG (eq 1) is smaller than 0.05 kJ/mol at all mixture compositions. The statistical uncertainty of $T\Delta S$ (eq 2) and ΔH (eq 3) is below 0.8 kJ/mol at all mixture compositions. Because eq 2 is only valid if the solvation heat capacity is constant over the selected temperature range, we decided to also determine ΔH and $T\Delta S$ with a direct method. In this method, performed at $T = 298$ K, the solvation enthalpy is determined by subtracting the potential energy of the solvent and the isolated solute from the solvated solute system potential energy. One then misses a term $P\Delta V$, which in condensed, liquid phases at ambient pressures usually is as small as 1−10 J/mol and can therefore safely be neglected. It is important, however, to subtract from the obtained potential energy difference a quantity $k_B T^2 \alpha_P$,[25] where α_P is the thermal expansion coefficient of the solvent. Having obtained ΔH by this direct approach, the solvation entropy can be obtained from the standard thermodynamic formula $T\Delta S = \Delta H - \Delta G$, where ΔG is obtained from eq 1. We used the direct approach by introducing ten methane solutes. Although a concentration

of ten solutes is above the solubility limit for all solvent mixtures, methane−methane interactions were observed to be negligible. We therefore assume that the methane solvation enthalpy with either one or ten methane solutes to be equal. The isobaric thermal expansion coefficient was obtained by performing, for each solvent composition, two additional 1 ns NPT simulations in which the temperature was changed ±10 K and the volume response was monitored; i.e. $\alpha_P \simeq (\ln[V_2/V_1]/[T_2 - T_1])_P$.

To better understand the enthalpy change we decompose this quantity in solute−solvent (ΔH_{uv}) and solvent−solvent enthalpy (ΔH_{vv}) changes, i.e.

$$\Delta H = \Delta H_{uv} + \Delta H_{vv} \tag{4}$$

The solute−solvent enthalpy change ΔH_{uv} is the sum of all solute−solvent pair interactions and was obtained by test-particle insertion[16]

$$\Delta H_{uv} = \frac{\langle V\psi e^{-\psi/RT}\rangle_{NPT}}{\langle Ve^{-\psi/RT}\rangle_{NPT}} \tag{5}$$

Calculations of ΔH_{uv} (eq 5) were all based on 100 ns sampling statistics with the systems summarized in Table 1. The enthalpy associated with changes of solvent−solvent interactions—ΔH_{vv} (the solvent reorganization enthalpy)—was obtained by means of eq 4 with ΔH obtained through eqs 1−3.

2.4. Solubility Data. We obtained the experimental methane solvation free energies, enthalpies, and entropies from the solubility data reported by Wang et al.[26] These data are reported in mole fraction methane (x_{CH_4}) in the solution at equilibrium with a methane pressure P at temperature T.[26] We use the Ben-Naim definition of the solvation process and the corresponding free energy change[27]

$$\Delta G = -RT \ln(\rho_{CH_4}/\rho^{ig}_{CH_4})_{eq} \tag{6}$$

where ρ_{CH_4} and $\rho^{ig}_{CH_4}$ are the methane molar densities in the solution and ideal gas phase, respectively. Both phases are at equilibrium as denoted by the subscript eq. In terms of the variables x_{CH_4}, P, and T used in ref 26 the solvation free energy (eq 6) becomes (assuming CH_4 is present at infinite dilution in the binary solvent)

$$\Delta G = RT \ln K_H - RT \ln (RT\rho_{solv}) \tag{7}$$

where ρ_{solv} is the molar density of the pure solvent phase (i.e., the alcohol/water mixture) and $K_H = \lim_{x_{CH4}\to 0}(P/x_{CH_4})$ the Henry coefficient. Experimental information on ρ_{solv} was obtained from ref 28. We note that K_H is usually expressed in units of atmospheres in which case $RT\ \rho_{solv}$ should also be expressed in these units. Wang et al.[26] reported methane solubility data at 283.15 K, 288.15, 293.15, and 298.15 K. By applying the finite difference method (eq 2) to the free energies obtained from eq 6 at 288.15 and 298.15 K we obtain the experimental solvation entropies at 293.15 K. The experimental solvation enthalpies at 293.15 K are next obtained from eq 3 using the solvation entropies together with the experimental solvation free energies at 293.15 K. Note that the so-derived solvation enthalpies and entropies all apply at a temperature 5 K below the temperature (298 K) used in our simulations. The discrepancy due to this difference in temperature we estimate based on the experimental heat capacity change Δc_P (142 J/mol K) of solvating methane in water at 298 K.[29] Assuming this value remains constant in the 5 K temperature range results in a 0.7 kJ/mol difference in ΔH and $T\Delta S$ (i.e., the values being 0.7 kJ/mol higher at the higher temperature).

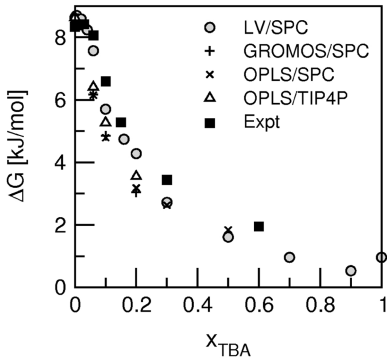


Figure 1. (a) Methane solvation free energy ΔG (eq 1) presented versus the *tert*-butanol mole fraction in aqueous *tert*-butanol solvent: squares, experimental data;[26] circles, the LV/SPC model; plusses, the GROMOS/SPC model; crosses, the OPLS/SPC model; and triangles, the OPLS/TIP4P model.

3. Results and Discussion

3.1. Free Energy of Methane Solvation. Figure 1 shows the methane solvation free energies at 298 K and 1 atm presented versus the alcohol mole fraction x_{TBA} of the solution. Transfer of the methane solute from pure water to pure *tert*-butanol causes the free energy to decrease with approximately 8 kJ/mol corresponding to a 25-fold increase of the methane solubility at 298 K. The shape of the curve ($\partial\Delta G/\partial x_{TBA} < 0$) indicates[16,27] that methane in TBA/water solution preferentially interacts with the alcohol molecules. Four TBA/water force field combinations have been used to calculate the ΔG data in Figure 1: LV/SPC, GROMOS/SPC, OPLS/SPC, and OPLS/TIP4P. The best agreement with the experimental data (squares) is observed for the LV/SPC model; the calculated data match the experimental values for x_{TBA} up to 0.06; at larger alcohol mole fractions the free energies are slightly underestimated by the model. For x_{TBA} up to 0.2, the GROMOS/SPC and OPLS/SPC models result in almost identical methane solvation free energies, which are however lower than the experimental and LV/SPC values. The ΔG values obtained based on the OPLS/TIP4P solvent model are slightly improved over those values obtained based on the corresponding OPLS/SPC model but are still systematically below the experimental and LV/SPC data. Table 3 summarizes the experimental[26] and calculated (LV/SPC model) thermodynamic data.

3.2. Preferential Solvation. Figure 2 shows methane−solvent radial distribution functions $g(r)$. The radial distribution functions (RDFs) were obtained from simulations including ten methane solutes in the systems summarized in Table 1. The left panel in Figure 2 shows the methane−TBA (methyl) RDF at TBA mole fractions 0.06, 0.1, and 0.2. The right panel shows the methane−water (oxygen)

Table 2. First Shell Coordination Numbers and Excess First Shell Coordination Numbers (in Parentheses) for Methane−TBA and Methane−Water[a]

x_{TBA}	methane−TBA	methane−water
tert-Butanol (LV-Model)/Water (SPC-Model)		
0.04	3.3(1.0)	13.7(−3.3)
0.06	5.6(2.3)	9.8(−5.7)
0.1	7.9(3.0)	5.9(−7.2)
0.2	9.5(2.1)	3.3(−5.7)
0.3	10.1(1.1)	2.2(−4.1)
0.5	10.7(−0.2)	1.2(−2.1)
0.7	11.0(−0.9)	0.6(−0.9)
0.9	11.1(−1.3)	0.2(−0.2)
tert-Butanol (GROMOS-Model)/Water(SPC-Model)		
0.06	7.0(3.8)	7.2(−9.3)
0.1	8.4(3.8)	4.7(−9.2)
0.2	9.7(2.6)	2.5(−7.0)
tert-Butanol (OPLS-Model)/Water(SPC-Model)		
0.06	6.8(3.6)	7.5(−9.0)
0.1	8.4(3.8)	4.8(−9.1)
0.2	9.7(2.5)	2.7(−6.9)
tert-Butanol (OPLS-Model)/Water(TIP4P-Model)		
0.06	7.5(4.3)	6.1(−10.7)
0.1	9.1(4.1)	4.8(−10.1)
0.2	10.6(2.9)	2.4(−7.9)

[a] The number of TBA and water molecules in the first shell was computed by integrating, respectively, the methane−CH_3 and methane−water−oxygen RDFs up to 0.53 nm.

Table 3. Thermodynamic Data for Methane Solvation (298 K, 1 atm) in TBA/Water Mixtures[a]

x_{TBA}	ΔG_{exp}	ΔG_{sim}	ΔH_{exp}	ΔH_{sim}	$T\Delta S_{exp}$	$T\Delta S_{sim}$
0.0	8.3	8.6	−11.5	−2.6	−19.8	−11.2
0.01	8.4	8.6	−8.4	−0.7	−16.6	−9.2
0.03	8.4		−7.6		−16.0	
0.04		8.2		2.9		−5.3
0.06	8.1	7.6	4.6	4.5	−3.5	−3.1
0.1	6.6	5.7	3.9	3.6	−2.7	−2.4
0.15	5.3	4.7	3.5	1.1	−1.8	−3.6
0.2		4.3		0.2		−4.1
0.3	3.4	2.7	0.7	−1.0	−2.7	−3.7
0.5		1.6		−2.9		−4.5
0.6	2.0		0.4		−1.6	
0.7		1.0		−2.8		−3.8
0.9		0.5		−3.5		−4.0
1		1.0		−1.9		−2.9

[a] The experimental data (denoted with subscript exp) were obtained from ref 26 (see section 2.4). The data calculated based on simulations (denoted with subscript sim) were obtained with the LV/SPC model[10] using test-particle insertions and finite temperature differences (eqs 1−3). Units are kJ/mol.

RDFs at the same solution compositions. Methane preferentially interacts with TBA-methyl groups. All RDFs (left panel) show a first peak at 0.4 nm followed by a second and third peak superimposed on an exponential type decay of the RDF toward unity. The limiting value (RDF = 1) is reached at approximately 1.5 nm with the LV/SPC model, while with all other models box sizes are too small to observe this limit. For the methane−water RDFs (right panel) the physical picture is reversed: a first peak is observed at 0.36 nm, followed by a water depletion region (RDF<1) up to 1.5 nm for the LV/SPC model; with all other models the limiting value RDF = 1 is not reached for distances up to 2 nm. First shell coordination numbers obtained by integrating the RDFs up to the first minimum are summarized in Table 2.

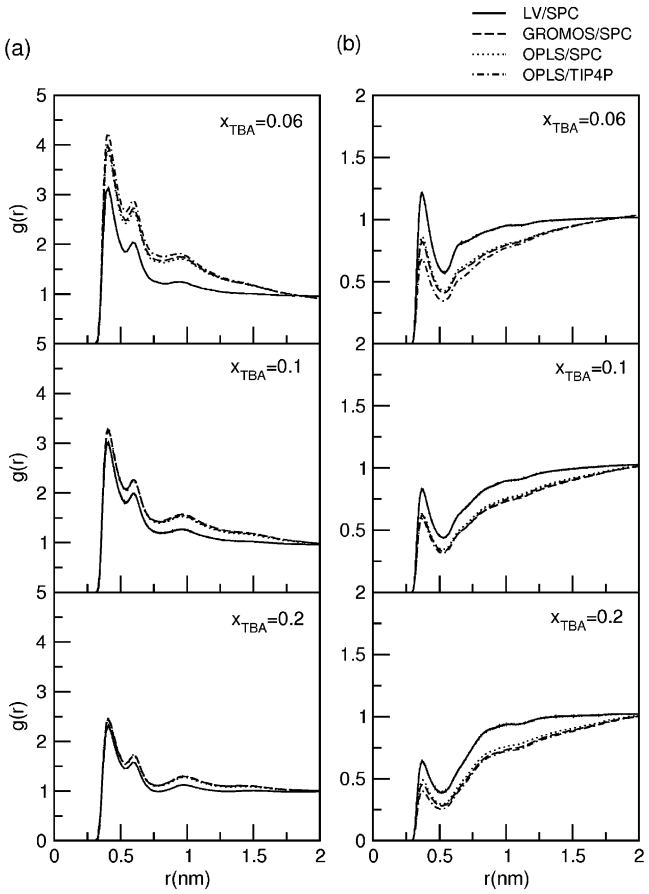


Figure 2. Solute (CH_4)−solvent radial distribution functions $g(r)$ versus the distance r at $x_{TBA} = 0.06$, $x_{TBA} = 0.1$, and $x_{TBA} = 0.2$ *tert*-butanol mole fractions: (a) methane−*tert*-butanol (methyl groups), (b) methane−water (oxygen). Solid line, LV/SPC model; dashed, the GROMOS/SPC model; dotted, the OPLS/SPC model; dashed-dotted, the OPLS/TIP4P model.

To quantify the excess amount of TBA and water vicinal to the methane solute we calculated methane−solvent excess coordination numbers N_j^{ex} defined as

$$N_j^{ex} = \rho_j \int_0^R [g_{Sj}(r) - 1] 4\pi r^2 dr \qquad (8)$$

In eq 8, ρ_j denotes the solvent component number density, $g_{Sj}(r)$ denotes the solute (CH_4)−solvent RDF, and R is an integration cutoff, which we chose to be 1 nm. Figure 3a,b shows the methane−TBA and methane−water excess coordination numbers, respectively. The CH_4−TBA excess coordination number (LV/SPC model) goes through a maximum at $x_{TBA} = 0.1$, while the CH_4−water excess coordination number goes through a minimum at this solvent composition. This behavior correlates with a maximum observed in the TBA−TBA Kirkwood-Buff integral at the same solvent composition.[10] With the GROMOS and OPLS models excess coordination numbers are found that exceed those obtained with the LV/SPC model by 50−60%.

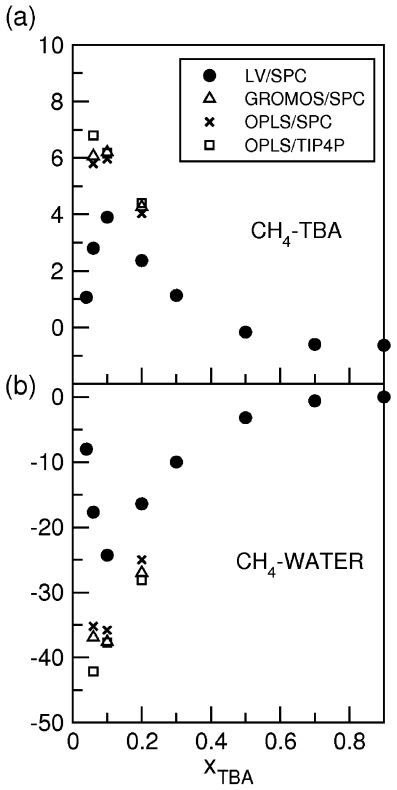


Figure 3. (a) Excess coordination numbers (eq 8 with R = 1 nm) for methane and *tert*-butanol obtained with the LV/SPC model (black circle), the GROMOS/SPC model (triangle), the OPLS/SPC model (cross), and the OPLS/TIP4P model (square). (b) Excess coordination number (eq 8 with R = 1 nm) for methane and water (oxygen).

3.3. Methane Solvation Enthalpy and Entropy. Figure 4a,b shows the methane solvation enthalpies and entropies versus the alcohol mole fraction of the solution. Interestingly, these curves are nonmonotonic in contrast to the solvation free energy (Figure 1). The experimental data (squares) indicate a rapid increase of ΔH with x_{TBA}. Methane solvation is exothermic in water but switches to endothermic between 3 and 6 mol % TBA in the solution. A maximum in ΔH is observed at $x_{TBA} = 0.06$. The methane solvation enthalpy decreases with a further increase of x_{TBA}. The methane solvation entropy rapidly increases up to $x_{TBA} \approx 0.1$ and then remains practically constant. With the LV/SPC model the experimental enthalpy curve and the transition from exo- to endothermic solvation are qualitatively reproduced. The maximum is predicted at the appropriate alcohol concentration and is almost quantitatively reproduced. At higher TBA mole fractions ($x_{TBA} > 0.06$) the solvation enthalpies are underestimated, while at higher alcohol dilution ($x_{TBA} < 0.06$) the solvation enthalpies are overestimated (see also Table 3). The same observations apply for the solvation entropies. Note that in water ($x_{TBA} = 0$) the methane solvation enthalpy and entropy are significantly overestimated. This is caused by a too large thermal expansion coefficient of the SPC water model resulting in a too large solvent reorganization enthalpy.[11,25] Based on the GROMOS and OPLS models (right panel in Figure 4) the solvation enthalpies and entropies are less well reproduced; although the downward trend of the enthalpy agrees with the experimental data, the sign is wrong, indicating that a significant, positive enthalpy contribution is missing.

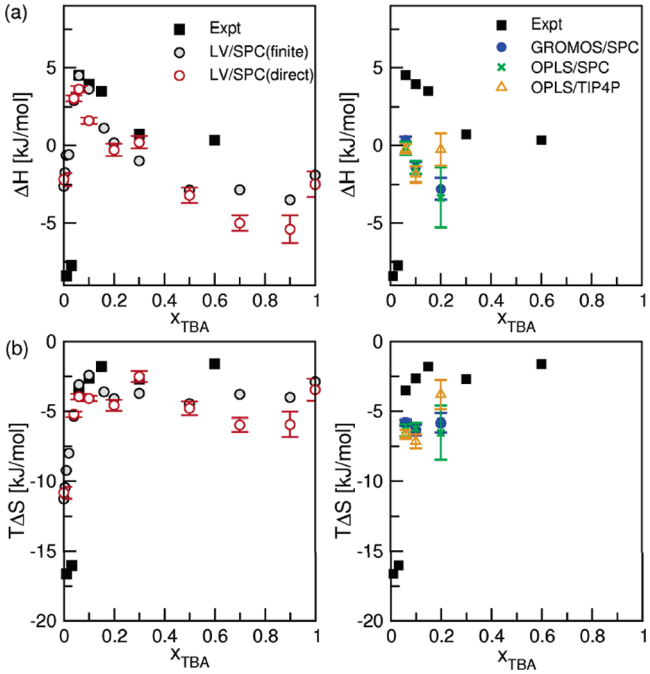


Figure 4. (a) Methane solvation enthalpies ΔH presented versus the *tert*-butanol mole fraction x_{TBA} in the solvent and (b) methane solvation entropies $T\Delta S$ versus x_{TBA}. For the LV/SPC model, solvation entropies and enthalpies obtained by finite temperature difference are shown as gray circles; the solvation enthalpies and entropies, obtained through the direct difference route, are shown as red circles. For the GROMOS and OPLS models (right panel), the solvation enthalpies and entropies were obtained through the direct difference route.

Table 4. Methane Solvation Enthalpy $(\Delta H)_{direct}$ Obtained from Taking Direct Energy Differences (cf. Section 2.2), Solvation Enthalpy $(\Delta H)_{FT}$ Obtained by Finite-Temperature (FT) Differences (Eqs 1−3), Structural Solvent Reorganization Energy (ΔH_{vv}), and Solute−Solvent Binding Energy (ΔH_{uv}) in TBA/Water (298 K, 1 atm) Obtained with the LV/SPC Model[10] [a]

x_{TBA}	$RT^2\alpha_P$	$(\Delta H)_{direct}$	$(\Delta H)_{FT}$	ΔH_{vv}	ΔH_{uv}
0.0	0.54	−2.2 ± 0.4	−2.6	11.2	−13.8
0.01			−0.7	12.7	−13.4
0.04	0.71	3.0 ± 0.2	2.9	16.2	−13.3
0.06	0.76	3.6 ± 0.2	4.5	17.7	−13.2
0.1	0.83	1.6	3.6	16.6	−13.0
0.15			1.1	14.1	−13.0
0.2	0.86	−0.3 ± 0.4	0.2	13.1	−12.9
0.3	0.88	0.2 ± 0.4	−1.0	12.0	−13.0
0.5	0.86	−3.2 ± 0.5	−2.9	10.2	−13.1
0.7	0.87	−5.0 ± 0.5	−2.8	10.2	−13.0
0.9	0.86	−5.4 ± 0.9	−3.5	9.5	−13.0
1	0.87	−2.5 ± 0.8	−1.9	11.0	−12.9

[a] Note that ΔH_{uv} depends only weakly on the TBA/water solvent composition, while ΔH_{vv} determines the compositional dependence of ΔH. Units are kJ/mol.

Table 4 and Figure 5 show the contributions of the solvent reorganization enthalpy (cf. eq 4) to the overall solvation enthalpy. Inspection of these data (LV-model) shows that

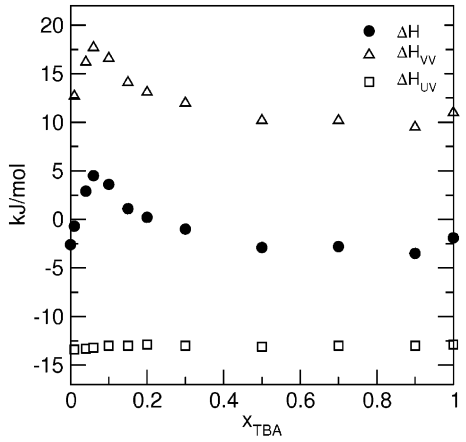


Figure 5. Methane solvation enthalpies ΔH (black circles), obtained from finite temperature differences (eqs 1−3), and contributions of solute−solvent interactions ΔH_{uv} (squares) (eq 5) and solvent reorganization enthalpy ΔH_{vv} (triangles) (eq 4) for the LV/SPC model. Note that changes of ΔH with the alcohol/water solution composition are almost exclusively determined by the enthalpy change ΔH_{vv} of structural solvent reorganization.

Table 5. Contributions of Water−Water, TBA−TBA, and TBA−Water Interactions to the Solvent Reorganization Enthalpy ΔH_{vv} at 298 K and 1 atm[a]

x_{TBA}	$\Delta H_{H_2O-H_2O}$	$\Delta H_{TBA-TBA}$	ΔH_{TBA-H_2O}
	LV/SPC		
0.04	80 ± 3	−16 ± 9	114 ± 17
0.06	−5 ± 13	−35 ± 8	219 ± 19
0.1	−41 ± 22	−11 ± 13	210 ± 31
0.2	−54 ± 42	14 ± 29	189 ± 64
0.3	13 ± 42	91 ± 35	31 ± 72
0.5	−63 ± 36	46 ± 33	151 ± 63
0.7	−21 ± 37	86 ± 36	63 ± 68
0.9	0 ± 8	131 ± 14	2 ± 18
	GROMOS/SPC		
0.06	−13 ± 19	−2 ± 11	158 ± 25
0.1	−89 ± 31	−27 ± 19	242 ± 40
0.2	−53 ± 61	33 ± 43	134 ± 95
	OPLS/SPC		
0.06	−18 ± 35	−10 ± 26	164 ± 50
0.1	−43 ± 55	−1 ± 44	171 ± 91
0.2	−138 ± 162	−63 ± 137	311 ± 272
	OPLS/TIP4P		
0.06	−63 ± 31	−33 ± 29	231 ± 48
0.1	−87 ± 60	−27 ± 44	238 ± 92
0.2	95 ± 100	165 ± 85	−119 ± 329

[a] The contributions were obtained by subtracting the potential energy components of a pure solvent box from the respective energy components of boxes including ten methane solutes. Averaging was performed over 100 ns simulations (Table 1). The error bars were obtained by block averaging. Units are kJ/mol.

the shape of the enthalpy curve is entirely determined by the solvent reorganization enthalpy ΔH_{vv}; the solute−solvent enthalpy ΔH_{uv} changes only little with the solution composition (Figure 5). Further separation of ΔH_{vv} in Table 5 shows that the largest contribution to the enthalpy maximum observed in Figures 4a and 5 results from changes in *tert*-butanol−water interactions. With the GROMOS/SPC, OPLS/SPC, and OPLS/TIP4P models changes in *tert*-butanol−water interactions result in a positive enthalpy contribution too, but these are smaller or compensated by a negative enthalpy contribution of water−water interactions resulting in a larger discrepancy with the experimental data (Figure 4a).

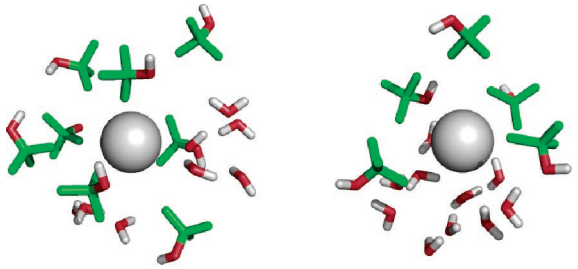


Figure 6. Simulation snapshots (LV/SPC model) showing the composition of the first solvation shell of methane (gray sphere) at $x_{TBA} = 0.1$. All solvent molecules within 0.6 nm from the central methane solute are shown. Color code: oxygen (red), hydrogen (white), and carbon/methyl (green). Alcohol molecules preferentially solvating methane lose hydration waters.

The solvation entropy (Figure 4b) follows the trend of the enthalpy (Figure 4a); however, there is no clear maximum. We note here that the solvent reorganization enthalpy occurs as a contribution $\Delta H_{vv}/T$ in the entropy,[30,31,11] hence it is not surprising that the discrepancies in ΔH, observed with the GROMOS and OPLS models, occur in the entropy changes too.

3.4. Discussion. Mixtures of *tert*-butanol and water exhibit significant alcohol clustering at low alcohol contents. The largest excess alcohol−alcohol aggregation occurs around 10 mol % alcohol in the mixture.[10] From Figures 2 and 3 it is evident that methane is solvated by these hydrated clusters. Snapshots of this situation are shown in Figure 6 and include only the solvent molecules in the first solvation shell. A key difference between the LV model, on the one hand, and the GROMOS and OPLS models, on the other hand, is that the LV model has been parametrized to reproduce the experimental alcohol and water activities in the mixture.[10] Alcohol clusters in the water-rich composition regions remain sufficiently hydrated to keep the system miscible. Water molecules are however expelled from the alcohol clusters by the introduction of methane. As shown in Table 5, this leads to a significant enthalpy increase due to disruption of alcohol−water cohesive interactions. The endothermic heat of methane dissolution around $x_{TBA} = 0.1$ (Figure 4) is therefore explained by the atomic-scale picture where methane sticks to alcohol clusters and expels hydration water.

The procedure of validating the solvation enthalpies and entropies seems to provide better insight in the accuracy of solvent models in general since the discrepancies observed in these quantities (Figure 4, right panel) are larger than in the solvation free energy (Figure 1). A similar observation was made by Weerasinghe and Smith,[32] who found that free energies of cavity formation in aqueous urea systems are independent of the extent of urea aggregation, while preferential interaction is intimately related to the solution mixture structure. As we pointed out above, the solvent reorganization enthalpy ΔH_{vv} is always exactly enthalpy−entropy compensating, and, therefore, it does not influence the free energy change.[30,31,11] The free energy change can therefore be perfectly reproduced, while its enthalpic and entropic components are

not. The solvation enthalpy, which contains a contribution ΔH_{vv}, and entropy, which contains a contribution $\Delta H_{vv}/T$, are strongly influenced by changes in solvent−solvent interactions and in that sense sensitively probe the quality of the force field describing the aqueous/organic solvent mixture.

4. Conclusions

We performed atomistic computer simulations of methane solvated in mixtures of *tert*-butanol and water. In addition to aspects of preferential methane solvation in terms of radial distribution functions and excess coordination of solvent components, we calculated the methane solvation free energy, enthalpy, and entropy based on four force field models available in the literature. Since preferential solvation of solutes by cosolvent or water molecules is likely to be coupled to aspects of cosolvent and water clustering in the solvent mixture, we used the Lee and Van der Vegt (LV) force field,[10] which was parametrized to reproduce the solvent Kirkwood-Buff integrals. We find that the enthalpy and entropy changes of methane insertion are very sensitive to changes of solvent−solvent interactions and therefore accurately probe the quality of the force fields for this solvent mixture. In particular, at high alcohol dilution, methane sticks to hydrated alcohol clusters and expels alcohol hydration water. This process is responsible for an endothermic heat of methane solvation (as opposed to exothermic solvation in pure water) and an increase of the solvation entropy.

References

(1) Tanford, C. *Adv. Protein Chem.* **1970**, *24*, 1.

(2) Carpenter, J. F.; Crowe, J. H. *Cryobiology* **1988**, *25*, 459.

(3) Xie, G. F.; Timasheff, S. N. *Biophys. Chem.* **1997**, *64*, 25.

(4) Wang, G. M.; Haymet, A. D. J. *J. Phys. Chem. B* **1998**, *102*, 5341.

(5) Lins, R. D.; Pereira, C. S.; Hünenberger, P. H. *Proteins* **2004**, *55*, 177.

(6) Shimizu, S.; Chan, H. S. *Proteins* **2002**, *49*, 560.

(7) Bennion, B. J.; Daggett, V. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *125*, 1950.

(8) Lee, M. E.; Van der Vegt, N. F. A. *J. Am. Chem. Soc.* **2006**, *128*, 4948.

(9) Matteoli, E.; Lepori, L. *J. Chem. Phys.* **1984**, *80*, 2856.

(10) Lee, M. E.; Van der Vegt, N. F. A. *J. Chem. Phys.* **2005**, *122*, 114509.

(11) Özal, T. A.; Van der Vegt, N. F. A. *J. Phys. Chem. B* **2006**, *110*, 12104.

(12) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Hermans, J. in *Intermolecular Forces*; Pullman, B. Ed.; Reidel: Dordrecht, 1981; pp 331−342.

(13) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656.

(14) Jorgensen, W. L. *J. Phys. Chem.* **1986**, *90*, 1276.

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

(16) Van der Vegt, N. F. A.; Van Gunsteren, W. F. *J. Phys. Chem. B* **2004**, *108*, 1056.

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

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

(19) Nosé, S. *Mol. Phys.* **1984**, *52*, 255.

(20) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.

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

(22) Nosé, S.; Klein, M. L. *Mol. Phys.* **1983**, *50*, 1055.

(23) Widom, B. *J. Chem. Phys.* **1963**, *39*, 2808.

(24) Smith, D. E.; Haymet, A. D. J. *J. Chem. Phys.* **1993**, *98*, 6445.

(25) Hess, B.; Van der Vegt, N. F. A. *J. Phys. Chem. B* **2006**, *110*, 17616.

(26) Wang, Y.; Han, B.; Yan, H.; Liu, R. *Thermochim. Acta* **1995**, *253*, 327.

(27) Ben-Naim, A. *Molecular Theory of Solutions*; Oxford University Press: New York, 2006.

(28) Nakanishi, K.; Kato, N.; Maruyama, M. *J. Phys. Chem.* **1967**, *71*, 814.

(29) Makhatadze, G. I.; Privalov, P. L. *J. Mol. Biol.* **1993**, *22*, 639.

(30) Yu, H. A.; Karplus, M. *J. Chem. Phys.* **1988**, *89*, 2366.

(31) Van der Vegt, N. F. A.; Trzesniak, D.; Kasumaj, B.; Van, Gunsteren, W. F. *Chem. Phys. Chem.* **2004**, *5*, 144.

(32) Weerasinghe, S.; Smith, P. E. *J. Chem. Phys.* **2003**, *118*, 5901.

CT600226H

JCTC Journal of Chemical Theory and Computation

A Distributed Computing Method for Crystal Structure Prediction of Flexible Molecules: An Application to *N*-(2-Dimethyl-4,5-dinitrophenyl) Acetamide

Victor E. Bazterra,[†,‡] Matthew Thorley,[‡] Marta B. Ferraro,[†] and Julio C. Facelli*[,‡]

Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pab. I (1428), Buenos Aires, Argentina, and Center for High Performance Computing, University of Utah, 155 South 1452 East Rm 405, Salt Lake City, Utah 84112-0190

Received June 23, 2006

Abstract: In this paper, we describe a new distributed computing framework for crystal structure prediction that is capable of performing crystal structure searches for flexible molecules within any space group and with an arbitrary number of molecules in the asymmetric unit. The distributed computing framework includes a series of tightly integrated computer programs for generating the molecule's force field, sampling possible crystal structures using a distributed parallel genetic algorithm, locally minimizing these structures and classifying, sorting, and archiving the most relevant ones. As an example, we report the results of its application to the prediction of the crystal structure of the elusive *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide, a molecule for which its crystal structure proved to be one of the most difficult cases in the last CSP2004 blind test for crystal structure prediction.

Introduction

The crystal structure prediction (CSP) of organic compounds can be described as the process of creating a list of crystal structure candidates, which are likely to be found experimentally, using only the molecular composition and connectivity as input information. The prediction of crystal structures for organic molecules is of great importance in many industries like pharmaceuticals, agrochemicals, pigments, dyes, explosives, and so forth, because many of the macroscopic properties of their products are highly dependent on their crystal structures. The existence of different crystals for the same compound, a phenomenon called polymorphism,[1,2] generally implies a significant variation of the substance's macroscopic properties, such as solubility, bioavailability, vapor pressure, crystal size and color, and so forth, depending on the predominant polymorph present in the material. Therefore, knowledge in advance of a plausible set of possible crystal structures provides important information that can be used for controlling the manufacturing process of solid organic compounds with the desired properties.[3]

Since the 1990s, a diverse group of methods have been developed for CSP.[4] The collective accomplishments of this research community have been summarized in the reports from the three blind tests for CSP conducted under the auspices of the Cambridge Crystallographic Data Center.[4−6] While the results of these workshops show gradual improvement in the predictive capability of the existing methods, there is general agreement that improvements are still necessary in both the force fields needed to better represent the molecular interactions and the optimization techniques used to search the complex energy landscape of molecular crystals.[4] This paper describes our recent developments to improve the latter aspect of CSP when taking advantage of the great progress in parallel computing realized in recent years.

Although the methods used for CSP vary, all of them follow four general procedures:[4] (1) the definition of a molecular model, which may or may not allow the simultaneous change of the molecular conformation during the

* Corresponding author e-mail: julio.facelli@utah.edu.
† Universidad de Buenos Aires.
‡ University of Utah.

10.1021/ct6002115 CCC: $37.00
Published on Web 12/16/2006

crystal structure searches, (2) an algorithm for searching plausible packing arrangements of the molecules in the crystal structure, (3) a method for ranking the structures according to their relative energies, usually defined by an empirical force field, and (4) a method for the classification and archival of the most relevant structures.

The molecular models can be classified into two groups known as *rigid and flexible.*[6] Within the rigid models, which are the most widely used today, the conformation of the molecule is kept fixed during the crystal structure search process. Usually, the molecular conformations considered are selected a priori by analyzing the molecule's conformational energy profile calculated for the isolated molecule using ab initio methods. This approach is a reasonable approximation when the intermolecular interactions are much smaller than the energy modes associated with the internal degrees of freedom of the molecule. However, this is not the case for flexible molecules for which the intermolecular interaction energies are often of the same order of magnitude than the rotational barriers around single bonds. Methods that allow the concurrent relaxation of the molecular geometry through the global search of the crystal structure are needed for many important applications in which flexible molecules should be considered.

The search for possible packing arrangements of the molecules in a crystal is generally implemented by a global optimization algorithm. Many search techniques have been used for this step, ranging from grid-based searches to stochastic methods like Monte Carlo simulated annealing and genetic algorithms (GAs). Following our previous work, here we continue using our modified GA approach,[4,7,8] which provides a convenient path to parallel implementations that can take advantage of modern computer equipment. It should be noted that our modified GA always includes a local minimization step for all of the structures considered in the GA evolution; that is, only locally minimized structures are included in the GA populations.

In principle, there are two approaches that could be used for CSP; either the search is constrained to run on a relatively small number of space groups one at a time or the search can be run totally unconstrained. The first approach leads to a significant reduction of the search space, and its usefulness lies in the fact that 92.7% of the organic crystals belong to only 18 of 230 space groups[9] and that 91.7% of the known organic crystal structures have only one molecule ($Z' = 1$) in their asymmetric unit. The disadvantage of this approach is that it is not able to find structures that may not be in the most common symmetry groups. Searches without symmetry constraints, based on the *P*1 space group with a variable number of molecules in their unit cell, have been reported, but they are extremely difficult because the significant increase in the search space greatly reduces their ability to produce a representative sampling of the possible packing arrangements.[9] In spite of the drawbacks of the unconstrained searches, in our computer programs, we have implemented both methodologies for completeness. Our preliminary findings of their relative performance of the constrained and unconstrained searches of the crystal structure of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide confirm the lack of predictive powder of unconstrained searches.

The ranking of the structures generated by the search algorithm is generally done according to their calculated energy. This criterion is based on the assumption that the crystallization process is under thermodynamic control. Although there is evidence that this is not always the case,[10,11] it is reasonable to assume that any metastable crystal structure should have a crystal energy close to the global minimum.[12] Alternative criteria for ranking crystal structures are sometimes used, for instance, taking into account the three-dimensional regularities commonly observed in the crystal structures deposited in the Cambridge Structural Database.[13] An extensive comparison of the methods currently available for CSP can be found in the Supporting Information of ref 4.

The comparison and classification of the resulting crystal structures is generally accomplished by sorting them by energy and eliminating from the list those structures that can be superimposed within a given tolerance, this can be done using programs like *CRYCOM*[14] or *COMPACK.*[15] Crystal structures with more than one molecule in the asymmetric unit may exhibit higher symmetry than required for the space group in which their search was performed; the additional symmetry elements in these structures can be found using the *ADDSYM* algorithm in *PLATON*[16] or the symmetry finder in *Cerius2.*[17]

This paper describes a distributed computer framework where several independent computer programs have been integrated to provide a comprehensive environment for CSP. Within this environment it is possible to (1) search for crystal structures within any symmetry group and with an arbitrary number of molecules and molecular types per asymmetric unit; (2) search structures using either the rigid or flexible molecule models; (3) automatically generate the molecule's force field using existing force field libraries; (4) increase the sampling power and the complexity of molecules amenable to CSP studies using the parallel and distributed computing capabilities of the system; and (5) automatically compare, sort, and archive the most relevant structures in a user database.

As an example, we show that this method can correctly predict the structure of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide, a well-recognized "problem" case in the crystal prediction literature.

Computational Methods

Modified Genetic Algorithm for Crystal Structure Prediction (MGAC). GAs are a family of search techniques rooted on the ideas of Darwinian evolution.[18] Operators analogous to crossover, mutation, and natural selection are employed to perform a search able to explore and learn the multidimensional parameter space and to determine which regions of the space provide good solutions to a problem and in which the search should be intensified. To improve their convergence, GAs are commonly coupled with local optimizations at each generation, a practice that has been followed in this work. Therefore, it is important to emphasize

that all the crystal structures reported here correspond to local minima of the potential energy surface.

Crystal Structure Model. When using GAs for the prediction of crystal structures, these structures have to be encoded in a *genome* that can be manipulated by the genetic operators as well as used to calculate the energy of the crystal structure they represent. For organic crystals, the molecular geometries are highly constrained by strong covalent bonds, leading to a considerable reduction of the number of parameters that are allowed to change during the global search. This means that it is not necessary to include the molecule's bond lengths and bond angles in the global optimization because their values in the crystal structure are always close to those in the isolated molecule. Therefore, they can be obtained by performing only local optimizations in which these parameters are allowed to change. Because the rotational barriers around single covalent bonds are comparable to the intermolecular interaction energies, their associated dihedral angles can be significantly affected by the intermolecular interactions leading to values in the crystal that are quite different than those in the isolated molecule. Therefore, these dihedral angles must be included in the global optimization to allow the exploration of conformations which may become energetically favorable for the certain packing motifs.

For each molecular species, we define a molecular frame anchored to the rigid structure of the molecule in which the positions of all the atoms can be determined. Using this molecular frame, we define the *genome* for a crystal structure with n molecules in the asymmetric unit, by specifying in the crystal frame the position of the origin of the molecular frames, $\mathbf{R}_1...\mathbf{R}_n$, and the orientation of the molecular axis relative to the crystal frame, $\Theta_1...\Theta_n$. These orientations are given by the corresponding Euler angles. For rigid molecules using these parameters, the symmetry group, and the molecular geometry, it is possible to calculate the position of all the atoms in the unit cell. For flexible molecules, as discussed above, the *genome* needs to be augmented by k scalars, $\Phi_1...\Phi_k$, that give the values of the k single-bond dihedral angles, measured in the molecular frame, that are allowed to change during the global optimization. It is very important to understand that the inclusion of these dihedral angles in the *genome* allows the GA to sample regions of the conformation space with quite different dihedral angles than those used as starting parameters. While these regions may not be energetically favorable for the isolated molecule they may be favorable for certain packing configurations. Note that regardless of which dihedral angles are included in the *genome*, all the intramolecular geometry parameters—bond lengths, bond angles, and dihedral angles—are always locally optimized in every GA generation.

The MGAC program only considers the lattice angles (α, β, γ) as independent parameters in the GA optimization. The lattice lengths (a, b, c) are treated as dependent parameters derived from the position of the molecules in the unit cell. In this respect, MGAC uses an approach similar to the one in *Polymorph Predictor*.[9] The lattice lengths are determined as follows: Given the molecular coordinates and the lattice angles that define the crystal structure, the initial estimates of the values of lattice lengths are chosen such that they define the smaller space that encloses all the molecules in the asymmetric unit. To reduce the chance of producing very short intermolecular distances between molecules in the unit cell and their neighbors, which can lead to spurious results when locally optimized, the asymmetric unit is extended to guarantee a minimal intermolecular distance (by default, 3 Å). Note that this arbitrary determination of the initial values of the lattice parameters does not have any effect on the final crystal structures as they undergo a local minimization in which all the inter- and intramolecular parameters are optimized. This last step allows that the effects of molecular interactions be included in the local refinement of the entire crystal structure.

Modified Genetic Algorithm. Except for the addition of the lattice angles, the crystal structure encoding given in the last section resembles the one used to describe molecular clusters. A very efficient GA scheme for molecular cluster optimization has been previously proposed by Niesse et al.,[19,20] and following this precedence, we have implemented in MGAC the *one-point-crossover*, *two-point-crossover*, *n-point-crossover*, *uniform-crossover*, *arithmetic-crossover*, *inversion-crossover*, *geometric-crossover*, and *gaussian mutation* genetic operators.[20] All of these operators are used in MGAC when acting on the molecular parameters. For the lattice angles, the *inversion-crossover* was removed from the operator list to preserve the crystal system, such that the resulting group of operators always transforms a triclinic structure into another triclinic and so forth.

Starting from a set of crystal structures randomly created (the initial generation or population), one can use these operators to construct a new set of crystal structure candidates for the next generation. The program verifies that the lattice angles define a linear independent set of lattice vectors in three dimensions, eliminating any spurious quasi-planar or linear structures. Any structure that does not meet this test is eliminated and replaced by a new one randomly generated. This procedure is repeated until a complete new population is generated. At each GA evolution, all the new structures (by default, the number of new structures is half of the population size) are relaxed to the corresponding local minima using the local optimization techniques available in CHARMM.[21,22] The local optimization, in which all the inter- and intramolecular degrees of freedom are allowed to change, is performed within the desired space group and produces a new set of candidate solutions that compete with the preexisting solutions for their permanence in the population. This competition is implemented by combining both sets of solutions into a larger population from which the worst candidates are eliminated until the number of structures equals the desired population size. The fitness of the solution is given by the total energy of the crystal structure evaluated after the local optimization. This defines the population from which the next generation can be obtained by repeating the procedure just described above. This evolution is repeated either for a predetermined number of generations or until the diversity of the population reaches such uniformity that the GA procedure becomes a random search.

The algorithm described above can be used to perform constrained searches in any of the 230 space groups with an

arbitrary number of molecules and molecule types on the asymmetric unit. Therefore, MGAC can be used by systematically searching solutions in different symmetry groups, normally restricting the search to the most common ones or by unconstrained searches, a methodology that has not yet been proved effective in CSP.

To reach the high sampling power required for these studies, we have used a global parallelization scheme of the genetic algorithm implemented in MGAC. In every generation, the parallelization scheme relies on the simultaneous run of the local optimization of the crystal structures belonging to the same population. The parallelization was done using our adaptive parallel genetic algorithm, *APGA*.[23] *APGA* was designed to perform efficiently on a heterogeneous cluster of computers and to provide a great degree of adaptability and performance in distributed systems. This level of parallelization allows making the execution time of MGAC approximately independent of the population size when a sufficient number of processors is available for the calculations.

The *MGAC* package is written in C++, using *BASH*[24] for scripting, *MPICH*[25] for parallel programming, *GALib*[26] for the GA implementation, and *xerces-c*[27] for parsing input and output *XML*[28] files.

Force Field Generator. An automatic force field generator, *charmmgen*, was implemented on the basis of *antechamber*.[29] Given the molecular composition and connectivity, this program calculates the molecule's force field parameters on the basis of the general amber force field, GAFF,[30] or any other force field library with a similar functional form. In this work, we have used the default parameters of GAFF,[30] which is a general force field with parameters for most organic and pharmaceutical molecules containing H, C, N, O, S, P, and halogens. It uses a simple functional form in which the energy is defined by

$$E = \sum_{\text{bonds}} k_r(r - r_{\text{eq}})^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_{\text{eq}}) + \sum_{\text{dihedrals}} \frac{\upsilon_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right]$$

where r_{eq} and θ_{eq} are equilibrium structural parameters; k_r, k_θ, and υ_n are the force constants; n is the multiplicity; and γ is the phase angle for the torsional angle parameters. The A, B, and q parameters describe the nonbonded potentials. In this work, we calculated the atomic charges, q_i, using the restrained electrostatic potential approach implemented in the RESP program.[31] This program fits the atomic charges to reproduce the electrostatic potential generated by the molecule's charge distribution, which was calculated using the *Gaussian 03*[32] package at the HF/6-31G* level. The atomic charges used for *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide were calculated at the optimized geometry (HF/6-31G* level) of the isolated molecule, but it was verified that they do not change in any appreciable way for other conformations of the molecule.

All the crystal energy calculations and local optimization are performed using CHARMM[21,22] with the GAFF[30] parameters. At least two unit cells were included in the simulation box in every direction, short range nonbonded interactions were summed up to a cutoff of 14 Å, and the electrostatic interactions were calculated using the Ewald technique.

Analysis of the Resulting Structures. Once a series of MGAC runs has been performed, it is necessary to sort and compare the crystal structures generated in these runs to find the n unique ones with the lowest energies. Because of numerical fluctuations, the set of structures generated by the MGAC runs has many similar structures with small energy differences that correspond to the same physical structure. Therefore, it is necessary to detect and remove these duplicate crystal structures from the final list. There are several well-established methodologies for comparing three-dimensional crystal structures, such as a comparison of their computed powder patterns,[33,34] and the *CRYCOM*[14] or *COMPACK*[15] programs, but because the source codes of these programs are not available, they could not be easily integrated into our computer framework. Therefore, we have implemented an alternative method that was easily integrated into our crystal prediction environment.

Our approach is similar to the one implemented in *COMPACK*,[15] based on the comparison of the structures of two finite clusters, but because MGAC preserves the atomic labels of the molecules, we avoid the expensive computational step of building a classification tree to sort the atomic labels of the molecules in the two fragments.

Our method uses two spherical clusters of n and m molecules, with $m > n$, that represent the three-dimensional structures of the two crystal structures under comparison. We call these two clusters the reference and compared fragments, respectively. The number of molecules in each cluster needs to be sufficiently large to completely characterize the environment surrounding the central molecule of each fragment, which should be one of the molecules in the asymmetric unit. By default, the values of n and m are set to 16 and 24, respectively, but it should be emphasized that the comparisons are always done using the same number of molecules in each fragment, that is, matching the reference cluster with only a fraction of the molecules in the compared cluster. Our experience shows that very small changes in crystal structures can produce a different set of molecules in the list of the n ones closest to the central molecule. This can lead to the classification of two similar crystals as different ones because a dissimilar set of molecules has been included in the two finite clusters. Hence, it is convenient to have a larger compared cluster to ensure that all the molecules in the smaller reference fragment are included in the compared cluster. Because within MGAC we preserve the atomic labeling of the individual molecules, the comparator program can easily search for the best rotation and translation to superimpose the chosen central molecules of each fragment. This is done by minimizing the root-mean-square (RMS) deviations of their atomic coordinates, excluding the hydrogen atoms, using the overlapping points algorithm develop by Kabsch.[35,36] If the resulting RMS is higher than a given threshold (by default 0.5 Å), there is no match between the selected pair of central molecules and a new set may be chosen (see below). If the RMS is lower

than the threshold, the comparator searches for the best superposition between the reference fragment and a subset of *n* molecules in the compared fragment. If the resulting RMS is lower than a given value (by default, 1.0 Å), the two crystals are considered similar. If any of these comparisons, that is, between the central molecules and/or between the fragments, do not show a match and both crystal structures have only one molecule in their asymmetric unit, they are considered different. Otherwise, the comparator program repeats the process described above, constructing a new compared cluster for a different central molecule. This process continues until all possible molecules in the asymmetric unit of the compared crystal structure have been used as central molecules of the compared cluster. If no match is found, the entire process is repeated, but applying the inversion operation to the compared fragments. Finally, if no match is found when using the inversion operation, the two crystals are considered different.

The crystal comparator, named *MOLCRY*, has been written in C++ and *Python*[37] as an extension of the Computational Crystallography Toolbox.[38] We have implemented in C++ those sections of the code that are computationally intensive, and Python code is used to interface the different parts of the code into a coherent application. Furthermore, Python was used to develop a set of Web services to automatically process and store the crystal structures produced by MGAC (see below) into a user data base.

When the crystal comparator is used, it is possible to create the list of the *n* unique crystal structures with lowest energy for a given set of MGAC runs. This is accomplished by comparing any candidate to be added to the list against a subset of the structures already in it. This subset of structures is chosen such that their crystal structures do not differ in energy by more than a given amount (by default, 16 kJ/mol) of the new candidate structure. This limits the number of comparisons needed per candidate, assuring that the time required for building the list scales linearly with its size.

Web Services for Crystal Analysis. In order to automate and standardize the process of classification and comparison of the crystal structures, we have created Web services that perform the comparison and classification of the structures. These services provide also an interface to a user database in which the crystal structures studied in our laboratory can be maintained. These services were developed in Python using representational state transfer or REST[39] architecture and the *Django* Web framework.[40] REST was implemented on top of the hypertext transfer protocol or HTTP.[41] This procedure greatly reduces the human time required for the analysis process, opening the possibility of using this framework for high-throughput CSP work. A detailed technical description of the implementation of these Web services is given as part of the Supporting Information.

Results and Discussion

The crystal structure of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide has been recently used as a test case for flexible molecules in the last Cambridge blind test, CSP2004.[4] The crystal structure of this molecule became one of the most difficult to predict by all the crystal prediction methodologies presented in the CSP2004. Actually, none of the 18 participants in the CSP2004 found its experimental structure within the first three best candidates, which was the criterion used in CSP2004 for success.

While the structure of this compound is now known,[4] all the calculations reported here have been done as if performing a blind test; that is, no information of the experimental structure was used a priori in our calculations. A series of 20 MGAC runs for each of the 14 most common space groups in organic molecules—$P1$, $P\bar{1}$, $P2_1$, $C2$, Pc, Cc, $P2_1/c$, $C2/c$, $P2_12_12_1$, $Pca2_1$, $Pna2_1$, $Pbcn$, $Pbca$, and $Pnma$—were executed for structures with one and two molecules per asymmetric unit. Each GA run produced 130 generations with 30 crystal structures each, using a crossover probability of 1.0 and a mutation probability of 0.001. This procedure gave a total of 560 MGAC runs in which approximately 1.1 million crystal structures were evaluated for this molecule. This required an approximated total of 80 000 CPU hours from the computational resources that were available at the CHPC Arches Metacluster[42] and the NCSA Teragrid cluster.[43] The first 200 different structures, sorted by energy, were extracted from the collection of all the structures (including repeated ones) generated by the combined runs. When MOLCRY running on a PC AMD Athlon 64 X2 2.2 GHz was used, this entire selection process was done in 240 min. From this list, 11 planar crystal structures were removed, and the resulting 189 structures were analyzed for missing symmetries using *PLATON*'s *ADDSYM* algorithm.[16]

In Table 1, we present the 20 crystal structures with the lowest energy (the extended list of the 189 crystal structures is provided as Supporting Information). The experimental crystal structure was found ranked third by energy. Note that according to the criterion used in the CSP blind tests this should be considered a successful prediction. The comparison between this structure and the experimental one is given in Figure 1 and Table 2, showing the excellent agreement between them. Figure 2 depicts the first three crystal structures in Table 1 and their corresponding simulated powder diffraction patterns. It is apparent that, in spite of their close energies, they are quite different structures. This makes the agreement between the third structure and the experimental one even more remarkable. Clearly, our results suggest that future experimental work in searching for these polymorphs is highly desired.

It should be noted that the experimental crystal structure was found in a search constrained to the $P2_1$ space group with $Z' = 2$, and not in the $P2_1/c$ with $Z' = 1$ as it is known experimentally. The systematic absence of the experimental structure in the search constrained to the $P2_1/c$ space group suggests that it is harder for the MGAC algorithm to find this crystal structure in its own symmetry group than in a less restricted search at $P2_1$. To try to better understand this behavior, a separated series of 20 constrained GA runs in $P2_1/c$ with $Z' = 1$, $P2_1$ with $Z' = 2$, and $P1$ with $Z' = 4$ was executed using MGAC. Note that in any of these three searches it is possible to find the experimental crystal structure. For each kind of search, we averaged the lowest energy reached by each of them at each generation. These average values and their corresponding standard deviations

Table 1. First 20 Lower-Energy Crystal Structures of *N*-(2-Dimethyl-4,5-dinitrophenyl) Acetamide[a]

ranking	*a* Å	*b* Å	*c* Å	α	*β*	*γ*	volume Å^3	space group	addsym	energy[b] kJ mol
1	4.842	14.142	15.118	90.00	102.53	90.00	1011	$P2_1/c$		0.00
2	14.865	8.097	16.749	90.00	90.00	90.00	2016	*Pbca*		0.39
3	**12.554**	**4.799**	**19.406**	**90.00**	**118.58**	**90.00**	**1027**	$P2_1$	$P2_1/c$	**1.14**
4	4.849	9.659	11.953	109.99	96.84	96.94	514	$P\bar{1}$		1.82
5	8.262	12.971	9.590	90.00	95.65	90.00	1023	$P2_1/c$		2.03
6	10.247	11.064	9.470	90.00	105.71	90.00	1034	$P2_1/c$		2.38
7	5.030	13.115	15.861	90.00	107.79	90.00	996	$P2_1/c$		2.40
8	4.915	14.318	7.319	90.00	102.79	90.00	502	$P2_1$		2.40
9	8.105	16.598	8.179	90.00	112.34	90.00	1018	$P2_1/c$		2.83
10	3.955	16.617	7.809	90.00	98.58	90.00	508	$P2_1$		3.47
11	9.037	15.783	7.481	90.00	112.19	90.00	988	$P2_1/c$		3.59
12	8.422	9.536	15.112	90.00	121.81	90.00	1031	*Pc*	$P2_1/c$	3.73
13	11.821	4.825	19.397	90.00	111.96	90.00	1026	$P2_1$	$P2_1/c$	3.76
14	7.324	14.988	9.524	90.00	99.36	90.00	1032	*Pc*	$P2_1/c$	4.10
15	8.310	8.748	14.823	94.40	95.98	100.99	1047	$P\bar{1}$		4.17
16	8.420	8.814	14.460	86.82	82.59	79.64	1046	$P\bar{1}$		4.23
17	9.408	14.723	14.990	90.00	90.00	90.00	2076	$P2_12_12_1$		4.46
18	4.773	14.640	15.048	90.00	92.83	90.00	1050	$P2_1/c$		4.58
19	9.201	7.995	14.330	90.00	100.15	90.00	1038	$P2_1$		4.59
20	4.828	14.921	14.448	90.00	95.12	90.00	1037	$P2_1$		4.67

[a] The third-ranked structure matches the experimental known structure for this molecule. [b] Relative energies with respect to the lower crystal structure energy of −309.5077 kJ/mol.

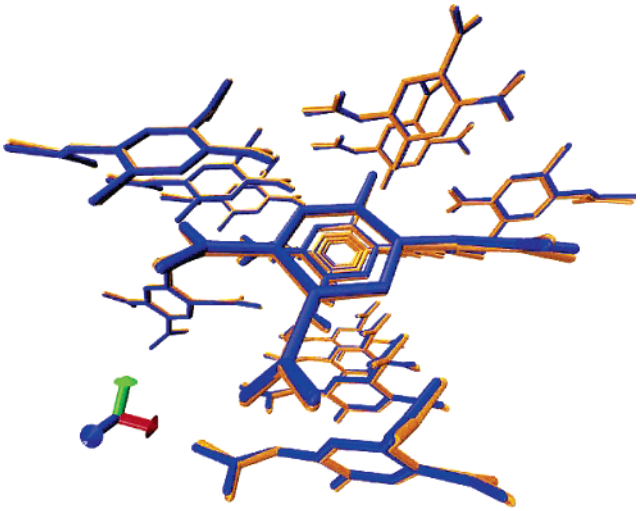


Figure 1. Superposition between experimental (blue) and predicted (orange) crystal fragments (16 molecules) of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide crystal structures. The total RMS for this superposition is 0.158 Å.

Table 2. Comparison between the Experimental and Predicted Crystal Structures of *N*-(2-Dimethyl-4,5-dinitrophenyl) Acetamide[a]

	experimental	predicted	difference
a (Å)	12.569	12.554	−0.33%
b (Å)	4.853	4.799	−1.1%
c (Å)	19.672	19.406	−1.3%
α	90.00	90.00	
β	119.95	118.58	−1.1%
γ	90.00	90.00	
vol (Å^3)	1040	1026	−1.3%
mRMS (Å)		0.076	
RMS (Å)		0.158	

[a] mRMS and RMS represent the best molecular and total root mean square.

are shown in Figure 3. It is interesting to observe that the optimizations with $Z' = 2$, while starting at a higher energy, after 70 generations find better solutions than those with $Z' = 1$. This is a remarkable behavior because it would be expected that the increase in the dimension of the configurational space for $P2_1$ would make it harder to find low-energy structures. Clearly this is what happens in the case of the search in *P*1, where the increased difficulty of finding low-energy structures is clearly noticeable by its slow convergence. MGAC is more efficient, at least for this molecule, when there are two molecules per asymmetric unit, and we speculate that this might indicate that the hypervolume of the attractive basin of the minimum corresponding to the experimental structure is much smaller for the search constrained to $P2_1/c$ and therefore less probable to find in the $P2_1/c$ configurational space than in $P2_1$.

The successful prediction of the experimental crystal structure of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide presented in this paper clearly contrasts with the results presented at the CSP2004. First, none of 18 participants found the experimental structure between the "top three" predictions, and only four participants found it in their extended list (up to 135 structures). Moreover, the best predicted structure exhibits a larger deviation from the experimental structure (RMS of 0.5 Å)[4] than the best structure reported here (RMS of 0.158 Å). This level of agreement shows that both the search methodology and the force field used in this study perform well for this molecule and most likely in similar compounds.

The energy range of the 189 structures collected in the final list is only ∼9.6 kJ/mol (Figure 4). This means that there are ∼20 crystal structures per kJ/mol, but as shown in Figure 4, their distribution as a function of energy is not

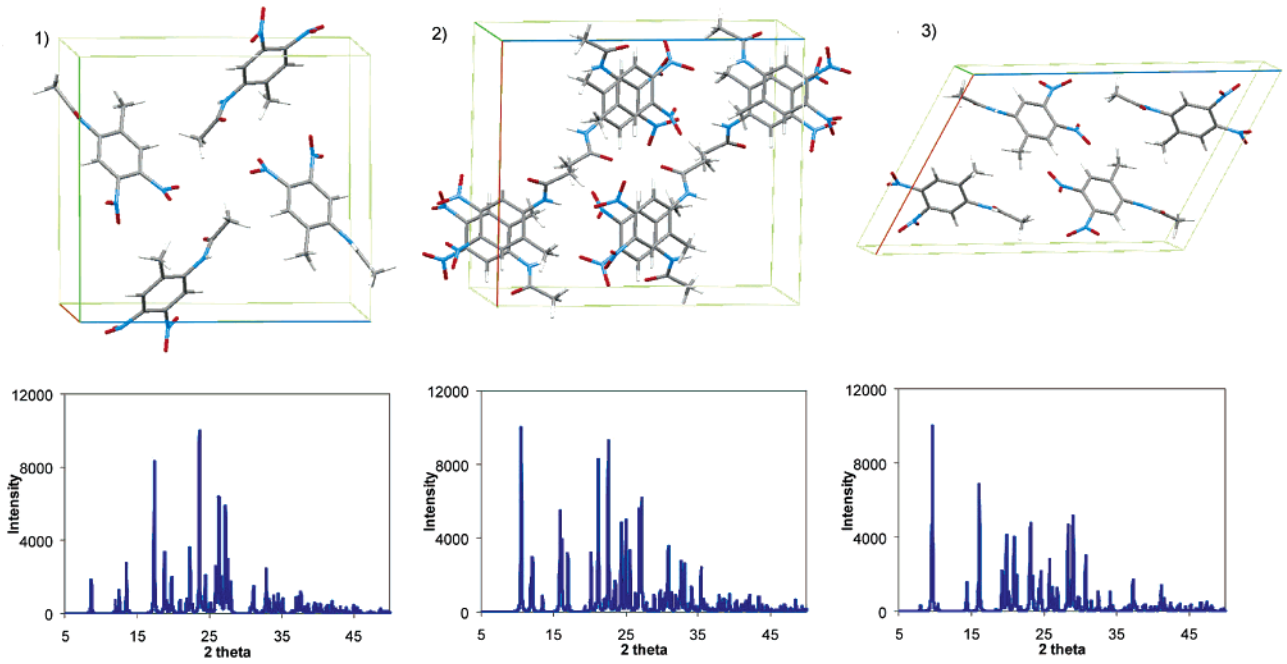


Figure 2. Comparison of the structures and powder diffraction spectra of the three lowest-energy structures of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide found in this study.

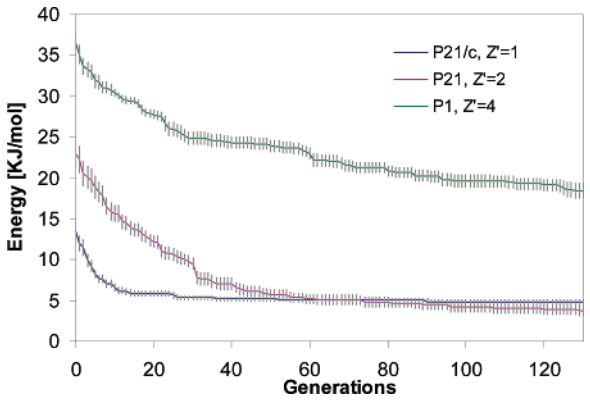


Figure 3. Average of the lowest energy per generation for 20 runs in three settings: $P2_1/c$ ($Z' = 1$), $P2_1$ ($Z' = 2$), and $P1$ ($Z' = 4$).

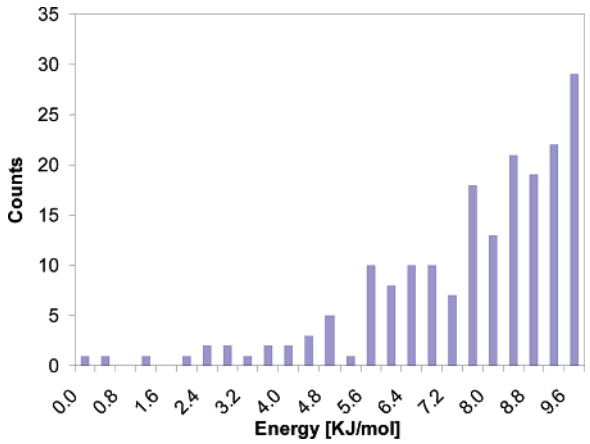


Figure 4. Energy histogram of the first 189 different crystal structures for the *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide molecule.

homogeneous. Even in this narrow range of energies, it is easy to appreciate the fast growth of the number of crystal structures close in energy to the global minimum. This shows that there are a large number of configurations with significant low energies to be considered as acceptable candidates for a minimum. Unfortunately, this can contribute to a rapid stagnation of the GA population and suggest that techniques like multiple independent runs (as used in this work), multiple concurrent populations, or random immigrants are necessary to obtain the desired convergence of the GA.

Conclusions

This paper reports the implementation and testing of a new distributed computing environment for CSP based on our previous work on GA. The environment allows the search of crystal structures for either rigid of flexible molecules without any restriction in the symmetry group or number of molecules in the asymmetric unit. The computational environment allows for the automatization of many processes, like the generation of the molecule's force field, execution of multiple GA runs, and the comparison and archival of relevant structures necessary for CSP studies.

As an example, we have shown that the method can predict (using the CSP2004 criterion) the crystal structure of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide, which none of the methods presented at CSP2004 were able to predict correctly. Moreover, our predicted structure agrees much better with the experimental one than any of those found in the extended lists of the CSP2004 blind test.

Another important conclusion that can be drawn from the difficulties encountered in predicting the crystal structure of this molecule is that the prediction of the crystal structures of flexible molecules requires advance search procedures with extensive sampling capabilities as much as it requires

greater accuracy in the modeling of the intra- and intermolecular energies.[44,45] Using the computer framework described here, we are performing more extensive tests of the MGAC method including a larger set of molecules and different force fields. The MGAC software will be available, under open-source licensing agreements, later in 2007.

Acknowledgment. The Center for High Performance Computing provided computer resources in the Arches cluster for this project. The Arches cluster was partially funded by a grant from the NIH-National Center for Research Resource (# 1S10RR017214-01). Additional computer resources were provided by NSF Teragrid MCA05T018 grant. The software for this work used the GAlib genetic algorithm package, written by Matthew Wall at the Massachussets Institute of Technology. MBF greatly acknowledge financial support from Universidad de Buenos Aires (UBACYTX035) and form the Argentinean CONICET (PIP5119/05). We thank two anonymous referees for providing critical comments on how to improve the presentation of our results.

Supporting Information Available: We provide a table with the data from the best 189 unique crystal structures of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide found in this study and the details of the implementation of Web services for crystal structure analysis and archival. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Threlfall, T. L. Analysis of Organic Polymorphs A Review. *Analyst* **1995**, *120*, 2435.

(2) Dunitz, J. D.; Bernstein, J. Disapparing Polymorphs. *Acc. Chem. Res.* **1995**, *28*, 193.

(3) Erk, P.; Hengelsberg, H.; Haddow, M. F.; Gelder, R. v. The Innovative Momentum of Crystal Engineering. *CrystEngComm* **2004**, *6*, 474.

(4) Day, G. M.; Motherwell, W. D. S.; Ammon, H. L.; Boerrigter, S. X. M.; Valle, R. G. D.; Venuti, E.; Dzyabchenko, A.; Dunitz, J. D.; Schweizer, B.; Eijck, B. P. v.; Erk, P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Leusen, F. J. J.; Liang, C.; Pantelides, C. C.; Karamertzanis, P. G.; Price, S. L.; Lewis, T. C.; Nowell, H.; Torrisi, A.; Scheraga, H. A.; Arnautova, Y. A.; Schmidt, M. U.; Verwer, P. A Third Blind Test of Crystal Structure Prediction. *Acta Crystallogr., Sect. B* **2005**, *61*, 511.

(5) Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Dzyabchenko, A.; Erk, P.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Lommerse, J. P. M.; Mooij, W. T. M.; Price, S. L.; Scheraga, H.; Schweizer, B.; Schmidt, M. U.; Eijck, B. P. v.; Verwer, P.; Williams, D. E. Crystal Structure Prediction of Small Organic Molecules: A Second Blind Test. *Acta Crystallogr., Sect. B* **2002**, *58*, 647.

(6) Lommerse, J. P. M.; Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Mooij, W. T. M.; Price, S. L.; Schweizer, B.; Schmidt, M. U.; Eijck, B. P. v.; Verwer, P.; Williams, D. E. A Test of Crystal Structure Prediction of Small Organic Molecules. *Acta allogr., Sect. B* **2000**, *56*, 697.

(7) Bazterra, V. E.; Ferraro, M. B.; Facelli, J. C. Modified Genetic Algorithm to Model Crystal Structures. I. Benzene, Naphthalene and Anthracene. *J. Chem. Phys.* **2002**, *116*, 5984.

(8) Bazterra, V. E.; Ferraro, M. B.; Facelli, J. C. Modified Genetic Algorithm to Model Crystal Structures: III. Determination of Crystal Structures Allowing Simultaneous Molecular Geometry Relaxation. *Int. J. Quantum Chem.* **2004**, *96*, 312.

(9) Gdanitz, R. J. Ab Initio Prediction of Possible Molecular Crystal Structures. In *Theoretical Aspects and Computer Modeling of the Molecular Solid State*; Gavezzotti, A., Ed.; John Wiley and Sons: New York, 1997; p 185.

(10) Gavezzotti, A. Ten Years of Experience in Polymorph Prediction: What Next? *CrystEngComm* **2002**, *4*, 343.

(11) Gavezzotti, A. Are Crystals Structures Predictable? *Acc. Chem. Res.* **1994**, *27*, 309.

(12) Day, G. M.; Chisholm, J.; Shan, N.; Motherwell, W. D. S.; Jones, W. An Assessment of Lattice Energy Minimization for the Prediction of Molecular Organic Crystal Structures. *Cryst. Growth Des.* **2004**, *4*, 1327.

(13) Apostolakis, J.; Hofmann, D. W. M.; Lengauer, T. Derivation of a Scoring Function for Crystal Structure Prediction. *Acta Crystallogr., Sect. A* **2001**, *57*, 442.

(14) Dzyabchenko, A. V. Method of Crystal-Structure Similarity Searching. *Acta Crystallogr., Sect. B* **1994**, *50*, 414.

(15) Chisholm, J. A.; Motherwell, W. D. S. COMPACK: A Program for Identifying Crystal Structure Similarity Using Distances. *J. Appl. Crystallogr.* **2005**, *38*, 228.

(16) *PLATON, A Multipurpose Crystallographic Tool*; Utrecht University: Utrecht, The Netherlands, 2005.

(17) *Cerius2*; Accelrys: San Diego, CA, 1997.

(18) Golberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: New York, 1989.

(19) Niesse, J. A.; Mayne, H. R. Global Optimization of Atomic and Molecular Clusters Using the Space-Fixed Modified Genetic Algorithm Method. *J. Comput. Chem.* **1997**, *18*, 1233.

(20) White, R. P.; Niesse, J. A.; Mayne, H. R. A Study of Genetic Algorithm Approaches to Global Geometry Optimization of Aromatic Hydrocarbon Microclusters. *J. Chem. Phys.* **1998**, *108*, 2208.

(21) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187.

(22) MacKerell, A. D.; Brooks, J., B.; Brooks, C. L., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Schreiner, P. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P., Schaefer, H. F., III, Ed.; John Wiley & Sons: Chichester, U. K., 1998; p 271.

(23) Bazterra, V. E.; Cuma, M.; Ferraro, M. B.; Facelli, J. C. A General Framework to Understand Parallel Performance in Heterogeneous Clusters: Analysis of a New Adaptive Parallel Genetic Algorithm. *J. Parallel Distrib. Comput.* **2005**, *65*, 48.

(24) GNU BASH. http://www.gnu.org/software/bash/ (accessed Oct 3, 2006).

(25) MPICH. http://www-unix.mcs.anl.gov/mpi/mpich (accessed Oct 3, 2006).

(26) Wall, M. GAlib. http://lancet.mit.edu/ga/ (accessed Oct 10, 2006).

(27) The Apache Project: Xerces-C++ Parser. http://xml.apache.org/xerces-c/ (accessed Oct 3, 2006).

(28) W3C Extensible Markup Language (XML). http://www.w3.org/XML/ (accessed Oct 3, 2006).

(29) Wang, J. Antechamber. http://amber.scripps.edu/antechamber/antechamber.html (accessed Oct 3, 2006).

(30) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157.

(31) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97*, 10269.

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

(33) Gelder, R. d.; Wehrens, R.; Hageman, J. A. A Generalized Expression for the Similarity of Spectra: Application to Powder Diffraction Pattern Classification. *J. Comput. Chem.* **2001**, *22*, 273.

(34) Karfunkel, H. R.; Rohde, B.; Leusen, F. J. J.; Gdanitz, R. J.; Rihs, G. Continuous Similarity Measure between Non-overlapping X-ray Powder Diagrams of Different Crystal Modifications. *J. Comput. Chem.* **1993**, *14*, 1125.

(35) Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr., Sect. A* **1976**, *32*, 922.

(36) Kabsch, W. A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr., Sect. A* **1978**, *34*, 827.

(37) Python Programming Language. http://www.python.org/ (accessed Oct 3, 2006).

(38) Computational Crystallography Toolbox CCTBX. http://cctbx.sourceforge.net/ (accessed Oct 3, 2006).

(39) Fielding, R. T. REST. http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm (accessed Oct 3, 2006).

(40) Django. http://www.djangoproject.com/ (accessed Oct 3, 2006).

(41) W3C: HTTP. http://www.w3.org/Protocols/ (accessed Oct 3, 2006).

(42) Arches Metacluster. http://www.chpc.utah.edu/docs/manuals/user_guides/arches/ (accessed Oct 3, 2006).

(43) NCSA TeraGrid IA-64 Linux Cluster. http://www.ncsa.uiuc.edu/UserInfo/Resources/Hardware/TGIA64LinuxCluster/ (accessed Oct 3, 2006).

(44) Eijck, B. P. v.; Mooij, W. T. M.; Kroon, J. Ab Initio Crystal Structure Predictions for Flexible Hydrogen-Bonded Molecules. Part II. Accurate Energy Minimization. *J. Comput. Chem.* **2001**, *22*, 805.

(45) Ouvrard, C.; Price, S. L. Toward Crystal Structure Prediction for Conformationally Flexible Molecules: The Headaches Illustrated by Aspirin. *Cryst. Growth Des.* **2004**, *4*, 1119.

CT6002115

JCTC Journal of Chemical Theory and Computation

Transferability of Orthogonal and Nonorthogonal Tight-Binding Models for Aluminum Clusters and Nanoparticles

Ahren W. Jasper, Nathan E. Schultz, and Donald G. Truhlar*

Department of Chemistry and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455-0431

Received August 10, 2006

Abstract: Several semiempirical tight-binding models are parametrized and tested for aluminum clusters and nanoparticles using a data set of 808 accurate Al_N ($N = 2-177$) energies and geometries. The effects of including overlap when solving the secular equation and of incorporating many-body (i.e., nonpairwise) terms in the repulsion and electronic matrix elements are studied. Pairwise orthogonal tight-binding (TB) models are found to be more accurate and their parametrizations more transferable (for particles of different sizes) than both pairwise and many-body nonorthogonal tight-binding models. Many-body terms do not significantly improve the accuracy or transferability of orthogonal TB, whereas some improvement in the nonorthogonal models is observed when many-body terms are included in the electronic Hamiltonian matrix elements.

I. Introduction

Atomistic simulations of large systems require methods for computing electronic energies and their gradients that are orders of magnitude more efficient than most ab initio and density functional theory (DFT) methods. Simple analytic potential energy functions (e.g., Lennard-Jones,[1] embedded atom,[2] etc.) are efficient, but they are not always accurate and may not be valid for uses other than those for which they are parametrized. For example, we showed previously[3,4] that analytic functions fit to either bulk data or diatomic data for pure aluminum perform poorly for particles of intermediate size, including clusters and nanoparticles. Although we were able to obtain[4] analytic potential energy functions that are accurate for small clusters, nanoparticles, and bulk aluminum using reasonably simple functional forms and an efficient fitting strategy that includes both small clusters and the bulk, the problem of extending these fits to heteronuclear systems remains unsolved and would likely require modified functional forms, such as charge-transfer terms,[5] variable atom types,[6] and so forth.

Semiempirical molecular orbital or crystal orbital methods include extended Hückel theory,[7] tight binding[8] (which is simply a more flexible form of extended Hückel theory or another name for extended Hückel theory), and neglect of differential overlap theories (like AM1[9,10] and others which have recently been reviewed[11]). They offer a theoretically attractive approach to modeling reactive systems because they are computationally affordable for many systems, while they include an orbital-based Hamiltonian, a diagonalization step, and the Pauli principle, three features that give rise to directional bonding and valence saturation. Some such methods, for example, Hoffmann's extended Hückel method,[7] include orbital overlap both in parametrizing the Hamiltonian and in the secular equation, while other methods, both in physics and in chemistry, include orbital overlap (or a function with comparable dependence on interatomic distance) in the Hamiltonian, but not in the secular equation. (The original Hückel method, in which the Hamiltonian matrix elements were constants, is no longer widely used.) Methods that neglect overlap in the secular equation are usually labeled "orthogonal", whereas those that retain it in the secular equation are labeled "nonorthogonal", and we follow this usage here. Molecular orbital methods provide a natural energetic description of bond breaking and forming and many-body effects.

Another classification that may be made is based on whether a method is designed to include self-consistent (i.e., iterative) steps, which may be necessary for accurately

* Corresponding author. Fax: (612) 624-7007. E-mail: truhlar@umn.edu.

10.1021/ct600261s CCC: $37.00
Published on Web 12/08/2006

modeling charge transfer and polar bonds. Although the motivation is to eventually use orbital methods to model Al heteronuclear chemistry, the focus of the present work is limited to pure aluminum clusters and nanoparticles where charge transfer and bond polarity may be expected to be less important than other contributions to the total energy, and we therefore restrict our attention in the present article to noniterative methods without an explicit treatment of charge interactions.

Popular orbital-based semiempirical methods were tested recently[12] for small aluminum clusters, where it was found that none of the semiempirical methods tested was accurate enough for quantitative work. The most accurate semiempirical method tested in ref 12 (AM1,[9,10] which involves iterating to self-consistency) had an average error of ~0.3 eV/atom. Subsequently, several parametrizations based on the Wolfsberg−Helmholtz[13] (WH) tight-binding (TB) model were obtained[14] using a database of aluminum cluster Al_N ($N = 2-13$) energies with an average error as small as 0.03 eV/atom. Here, we build on that work and explicitly consider the transferability of TB parametrizations using an expanded data set, containing systems as large as 177 atoms. We also consider both orthogonal and nonorthogonal TB models, and we discuss the relationship between orthogonality and many-body effects.

The question of orthogonal versus nonorthogonal formulations also arises in semiempirical methods that include a self-consistent-field step. For example, AM1[9,10] and PDDG/PM3[15] both set the overlap matrix equal to unity in the secular equation, whereas SCC-DFTB[16,17] and NO-MNDO[18] include the overlap matrix in the secular equation, thereby increasing the cost. Nonorthogonal formulations also complicate the gradient calculation,[19] and they make it more difficult to achieve linear scaling of computational times for large systems, although linear-scaling nonorthogonal formulations have been presented for both non-self-consistent[20,21] and charge-self-consistent[22] tight-binding methods. A recent[23] study found that the mean unsigned error for the heat of reaction of 34 diverse isomerization reactions is 7.1, 5.0, and 2.8 kcal/mol (0.31, 0.22, and 0.12 eV) for AM1, SCC-DFTB, and PDDG/PM3, respectively. Similar results were found for 622 heats of formation, with the nonorthogonal SCC-DFTB intermediate in accuracy between AM1 and PDDG/PM3[23] and with NO-MNDO better than MNDO for HCO compounds but not for HCN compounds.[18] Thus, orthogonal formulations may be either less or more accurate than nonorthogonal ones, depending on the parametrization. The present study is designed to test the role of nonorthogonality in tight-binding theory by parametrizing nonorthogonal and orthogonal formulations in the same way using the same training data.

Section II briefly introduces the TB models that are considered in this paper. In section III, the database of accurate energies is described, and the TB models are parametrized and tested in section IV. Section V discusses the results, and section VI is a summary.

II. Theory

A many-electron molecular wave function may be approximated as a product of one-electron wave functions ψ_j that satisfy

$$\hat{H}\psi_j = \epsilon_j\psi_j \tag{1}$$

where

$$\hat{H} = \hat{T} + \hat{V}_{\mathrm{NE}} + \hat{V}_2 \tag{2}$$

$\hat{T}$ is the kinetic energy operator; $\hat{V}_{\mathrm{NE}}$ is the Coulomb operator for the attraction of the electron to all of the nuclei, and $\hat{V}_2$ is the sum of the two-electron operators for the Coulomb repulsion, exchange, and correlation. We treat the valence electrons explicitly and combine the core electrons with the nuclei as the total core.

To solve eq 1, the one-electron molecular orbitals ψ_j are expanded in an atomic orbital basis set φ_μ^k, where i labels the individual atomic orbitals centered on atom k

$$\psi_j = \sum_i c_{ij}\varphi_i \tag{3}$$

where $i \equiv (\mu,k)$. The optimal expansion coefficients are obtained by solving the secular equation

$$\mathbf{Hc} = \epsilon\mathbf{Sc} \tag{4}$$

where ϵ is a diagonal matrix with diagonal elements ϵ_j

$$H_{\mu\mu'}^{kk'} = \langle\varphi_\mu^k|\hat{H}|\varphi_{\mu'}^{k'}\rangle \tag{5}$$

and

$$S_{\mu\mu'}^{kk'} = \langle\varphi_\mu^k|\varphi_{\mu'}^{k'}\rangle \tag{6}$$

The sum of the orbital energies (the eigenvalues of eq 4) weighted by the appropriate orbital occupancies n_j (with $n_j = 0$, 1, or 2, where j labels the eigenvalues and eigenvectors) is called the valence energy or the band energy, given by

$$E_{\mathrm{Val}} = \sum_j n_j\epsilon_j \tag{7}$$

Energy levels are typically filled to minimize eq 7, which is called the aufbau principle. However, in some cases, it is useful to minimize a more general electronic energy given by

$$E_{\mathrm{Elec}} = E_{\mathrm{Val}} + \delta_{n_j2}U_{\mathrm{Pen}} \tag{8}$$

where U_{Pen} is a penalty energy[24] for pairing electrons. We consider several values of the penalty energy (including zero). The use of a penalty energy is operationally equivalent to having different Hamiltonians for spin-up and spin-down electrons, where the Hamiltonians differ only in that U_{Pen} is added to the diagonal for spin-down electrons.

The total energy E is

$$E = E_{\mathrm{Elec}} + E_{\mathrm{Core}} - E_{\mathrm{DC}} \tag{9}$$

where E_{Core} is the core−core repulsion and E_{DC} corrects for the double counting of the two-electron interactions $\hat{V}$. The double-counting term and the core−core repulsion are often

grouped together and represented by an effective pairwise repulsive term

$$E = E_{\mathrm{Elec}} + V_{\mathrm{Rep}} \tag{10}$$

The pairwise form for V_{Rep} has been justified by Foulkes and Haydock.[25] We use a three-parameter repulsion with the form

$$V_{\mathrm{Rep}} = \sum_{k<k'} A \exp(-BR_{kk'})/R_{kk'}^{C} \tag{11}$$

where $R_{kk'}$ is the distance between atoms k and k'; A, B, and C are parameters; and the summation runs over all unique pairs of atoms.

In semiempirical tight-binding models, several additional approximations are made to make the above procedure computationally efficient. First, a minimal basis set is employed (specifically, for Al, one 3s and three 3p orbitals per atom). Next, the matrix elements $H_{\mu\mu'}^{kk'}$ are taken as empirical parameters or simple functions of empirical parameters. In this article, we consider the WH[13] model where

$$H_{\mu\mu'}^{kk} = -\delta_{\mu\mu'} U_{\mu}^{k} \tag{12}$$

U_{μ}^{k} is the valence state ionization potential for an electron

$$H_{\mu\mu'}^{kk'} = K\frac{H_{\mu\mu}^{kk} + H_{\mu'\mu'}^{k'k'}}{2} S_{\mu\mu'}^{kk'} \quad (k \neq k') \tag{13}$$

in atomic orbital i on atom k, and K is an empirical parameter, which Hoffmann optimized[7] to 1.75 but which we take to be a fitting parameter. We consider two other models for the off-diagonal elements: a parameter-free model derived by Cusachs and Cusachs[26] (CC)

$$H_{\mu\mu'}^{kk'} = (2 - |S_{\mu\mu'}^{kk'}|)\frac{H_{\mu\mu}^{kk} + H_{\mu\mu}^{k'k'}}{2} S_{\mu\mu'}^{k'k'} \tag{14}$$

and a form proposed by Lathiotakis et al.[27] for metals, which we denote LAMC.

In the present study, the valence state ionization potentials U_i^k used in eq 12 are fixed at the atomic experimental[28] values (10.62 and 5.986 eV for the 3s and 3p orbitals, respectively). Then, along with the three parameters in the repulsion, the WH model has one adjustable parameter K; the CC model has no adjustable parameters, and the LAMC model has five adjustable parameters for Al. Following Slater and Koster,[8] the matrix elements were evaluated in a symmetry-adapted local coordinate system when computing eq 14.

In all of the above formulations, $H_{\mu\mu'}^{kk'}$ depends only on atoms α and β, whereas in a more accurate calculation, the one-electron contributions also involve three-center terms, and the two-electron contributions involve three- and four-center terms. Slater and Koster[8] suggested the two-center simplification, and it is widely used. The models discussed above are usually called nonorthogonal tight-binding (NTB) to emphasize the retention of the overlap matrix **S**.

One may choose to go beyond pairwise tight binding by writing $H_{\mu\mu'}^{kk'}$ or V_{Rep} as a function of the geometry of the entire system (or the local geometry), and we call such formulations many-body tight-binding (MBTB) models.[14] (Some workers[29−33] call them environmental-dependent tight binding.) Ho and co-workers[30−32] suggested a many-body term to model screening. The screening many-body term was tested previously for WH tight binding for aluminum clusters,[14] and it was shown to be more accurate than two other many body-terms that were also tested. In the present article, we consider three implementations of the screening (S) many-body function, applied (as discussed elsewhere[14]) to screening the repulsion (SR), screening the ionization potentials (SIP), or screening the off-diagonal elements (SOD). The screening many-body function has three fitting parameters, and the SIP implementation has two additional parameters, as discussed elsewhere.[14] The many-body terms also contain a two-parameter cutoff function, as discussed in ref 3. One may call a model that retains overlap and adds many-body terms "nonorthogonal many-body tight-binding" (NMBTB).

Finally, one may also choose to neglect the overlap when solving eq 4, and we denote these models as TB and many-body tight binding; one could also say orthogonal tight-binding and orthogonal many-body tight-binding, but the usual notation in the literature is to assume orthogonality when "nonorthogonal" is not specified. There has been work[34] showing that the effect of including overlap (as in NTB) is similar to including a many-body term in the off-diagonal matrix elements of a TB model. This leads to the interpretation that NTB is more accurate than TB because it effectively goes beyond the pairwise approximation. However, both TB and NTB involve a diagonalization step, and therefore, both schemes include many-body effects. Furthermore, if one chooses to view eq 10 as a fitting scheme, then V_{Rep} not only includes the double-counting correction and the core−core repulsion, but it also empirically corrects for all of the other approximations made when applying TB to real systems. As a result, although it is true that the overlap matrix is very different from the unit matrix at chemical atom−atom separations, in light of the many other significant approximations in TB and NTB theory, it is not clear how important it is to include overlap in the secular equation. The question of whether one achieves a better *balance* of approximations when one retains or neglects overlap is one motivation for the current work.

III. Calculations

The tight-binding models are tested using the previously presented pure aluminum nanoparticle data set.[4] This data set consists of 808 energies and geometries for Al_N, with $N = 2-177$, including 127 data points for particles with diameters greater than 1 nm. Symmetric structures such as icosahedral, face-centered cubic (FCC), hexagonal close-packed, body-centered cubic, and simple cubic are included for several atomic volumes. Nonsymmetric structures are represented as well, including disordered structures and structures with over- and undercoordinated interior atoms, that is, atoms with coordination numbers greater than and less than the bulk value of 12. The data set is divided into 11 groups containing particles with sizes $N = 2$, 3, 4, 7,

Table 1. Mean Unsigned Errors (eV) for Several Tight-Binding (TB) Parameterizations[a]

method	fitting data	U_{Pen}, eV	K	ϵ_{Dim}	ϵ_{Clus}	ϵ_{Nan}	ϵ
WH	2	0	0.38	0.03	0.07	0.06	0.06
	2−13	0	0.41	0.03	0.06	0.06	0.06
	2−177	0	0.41	0.03	0.06	0.05	0.05
	2	0.02	0.39	0.02	0.07	0.06	0.06
	2−13	0.05	0.41	0.03	0.05	0.06	0.06
	2−177	0.04	0.41	0.03	0.06	0.05	0.05
	2	3	0.51	0.04	0.42	0.76	0.57
	2−13	3	0.42	0.10	0.10	0.09	0.09
	2−177	3	0.44	0.11	0.11	0.05	0.08
WH+SIP	2−13	0	0.40	0.03	0.06	0.06	0.06
WH+SR	2−13	0	0.43	0.04	0.05	0.05	0.05
WH+SOD	2−13	0	0.41	0.03	0.05	0.06	0.06
LAMC	2	0		0.03	1.89	3.81	2.77
	2−13	0		0.09	0.07	0.12	0.10

[a] All errors are averages over all 11 groups of data, even though the fits to the $N = 2$ data and the $N = 2-13$ data use only one or five, respectively, of the groups for fitting.

Table 2. Mean Unsigned Errors (eV) for Several Nonorthogonal Tight-Binding (NTB) Parametrizations

method	fitting data	U_{Pen}, eV	K	ϵ_{Dim}	ϵ_{Clus}	ϵ_{Nan}	ϵ
WH	2	0	1.85	0.08	0.63	1.20	0.89
	2−13	0	1.91	0.27	0.31	0.49	0.40
	2−177	0	1.99	0.38	0.41	0.24	0.31
	2	4.54	1.52	0.01	1.14	2.42	1.73
	2−13	0.24	1.91	0.24	0.30	0.44	0.37
	2−177	0.20	1.99	0.35	0.41	0.21	0.30
WH+SIP	2−13	0	1.94	0.15	0.20	0.40	0.30
WH+SR	2−13	0	1.89	0.24	0.31	0.44	0.37
WH+SOD	2−13	0	1.86	0.14	0.25	0.23	0.23
WHXR	2	0	1.61	0.28	0.70	1.21	0.94
	2−13	0	1.77	0.34	0.54	0.70	0.61
	2−177	0	1.90	0.49	0.70	0.35	0.49
CC	2	0		0.13	0.52	1.10	0.80
	2−13	0		0.16	0.49	1.10	0.79
	2−177	0		0.18	0.49	1.08	0.79
CCXR	none	0		0.19	0.50	1.08	0.79

Table 3. Mean Unsigned Errors (eV) for Several Previously Presented Parameterizations for Pure Aluminum

method (ref)[a]	fitting data	U_{Pen}, eV	K	ϵ_{Dim}	ϵ_{Clus}	ϵ_{Nan}	ϵ
ER (3)	2			0.01	0.82	2.75	1.80
Gol (36)	bulk			0.12	0.12	0.13	0.12
NP-A (4)	2−177			0.01	0.08	0.03	0.05
NP-B (4)	2−177			0.09	0.08	0.05	0.06
WH (14)	2−13[b]	0.07	0.39	0.03	0.06	0.06	0.06
OWH (14)	2−13[b]	0.07	*c*	0.06	0.08	0.07	0.07
EWH (14)	2−13[b]	0.07	*c*	0.05	0.06	0.09	0.08
MBTBS (14)	2−13[b]	0.07	*c*	0.02	0.05	0.08	0.06
MBTBCN (14)	2−13[b]	0.07	*c*	0.03	0.07	0.09	0.08
MBTBBA (14)	2−13[b]	0.07	*c*	0.03	0.05	0.06	0.06

[a] Reference number in parentheses. [b] A subset of the current data set for Al_N, $N = 2-13$, was used for these fits. [c] These models have multiple values of K, depending on the local orbital symmetry.

9−13, 14−19, 20−43, 50−55, 56−79, 80−87, and 89−177, and the number of data points in each group is 44, 402, 79, 42, 72, 42, 46, 23, 27, 15, and 16, respectively.

Mean unsigned errors per atom for each data group are computed as discussed elsewhere.[4] As we are interested in the transferability of the TB parametrizations for different particle sizes, we further average the errors and report the mean unsigned error ϵ_{Dim} for the dimer data, the average ϵ_{Clus} of the mean unsigned errors for the four groups containing aluminum clusters with sizes between 3 and 13 atoms, and the average ϵ_{Nan} of the mean unsigned errors for the six groups of nanoclusters and nanoparticles containing between 14 and 177 atoms. The overall mean unsigned error ϵ is obtained by averaging all 11 data groups, which is equivalent to

$$\epsilon = (\epsilon_{Dim} + 4\epsilon_{Clus} + 6\epsilon_{Nan})/11 \quad (15)$$

The available adjustable parameters were optimized by minimizing ϵ_{Dim}, $(\epsilon_{Dim} + 4\epsilon_{Clus})/5$, or ϵ using a genetic algorithm[35] for two values of the penalty energy U_{Pen}, 0 or 3 eV. Fits were also obtained by treating the penalty energy as an optimized fitting parameter.

IV. Results

Once the optimal parameters were obtained, the resulting fits were then tested against the full data set, and the results are summarized in Tables 1 and 2. For comparison, Table 3 contains errors for several fits that have previously appeared in the literature, including four analytic potential energy functions: ER,[3] a pairwise form fit to the aluminum dimer; Gol,[36] an embedded-atom form fit to bulk data; and NP-A and NP-B,[4] the two most accurate analytic potential energy functions from ref 2 fit to the same data set that is used here. Also shown in Table 3 are six tight-binding fits obtained previously[14] using a subset of the current data set with Al_N, $N = 2-13$, as well as a data set of ionization potentials and bulk properties. The OWH and EWH models are TB models that may be considered as extensions of the WH model, and the MBTBS, MBTBCN, and MBTBBA are examples of MBTB, in particular WH models with screening (S) and two other types (CN and BA) of many-body terms.[14]

V. Discussion

First, we consider the TB and MBTB results in Table 1. For zero penalty energy and when the penalty energy is optimized (resulting in values of 0.02 to 0.05 eV), the parametrizations obtained by fitting *only* to dimer data give very good overall errors of 0.06 eV/atom. When parameters are obtained by fitting to the full data set of 808 geometries, the overall errors do not improve significantly. Furthermore, these errors (∼0.05 eV/atom) are close to the error for the most accurate of the orthogonal fits and analytic potential energy functions presented previously[4,14] (see Table 3). This is perhaps a surprising result considering the long list of approximations involved in the simple, four-parameter, orthogonal WH TB model and that the training set involves no information about many-body effects.

Fits obtained with a penalty energy of 3 eV, which is the value used by Wang and Mak,[24] have larger errors than those

for smaller penalty energies. Furthermore, the transferability of the resulting fits is not as good as is observed for smaller penalty energies; that is, the errors are significantly larger for data not included in the fits when a subset of data is used as fitting data.

Including many-body terms (as in the SIP, SR, and SOD models) does not significantly improve the overall error, and this is likely due to the already excellent performance with accurate pairwise interactions.

The optimal value of the Wolfsberg-Hemlholtz parameter K is around 0.4 for the more accurate TB fits, which is very different from the value of 1.75 used by Hoffmann.[7] The value of 1.75 was obtained in the context of NTB, where Hoffmann gave an argument[7] that K should be greater than unity. In the context of orthogonal TB there is no such restriction, and the present values of $\sim$0.4 are therefore reasonable.

We also tested the LAMC functional form for the TB model. As shown in Table 1, the dimer LAMC fit does not show the excellent transferability of the WH model, although by fitting to data up to Al_{13}, good transferability is obtained to the regime of larger clusters and nanoparticles.

Next, we consider the NTB models summarized in Table 2. The overall errors for the pairwise NTB fits are in the range 0.3−1.7 eV/atom. The parametrizations obtained by fitting to the dimer data perform poorly for larger clusters. For example, the WH NTB model with zero penalty energy fits the dimer data with an error of 0.08 eV/atom, which is only slightly worse than the best TB model in Table 1. But when this parametrization is used to model clusters and nanoparticles, the mean unsigned error increases by factors of 8 and 16, respectively, resulting in qualitatively incorrect fits.

Many-body terms were incorporated in the nonorthogonal formalism, and their parameters were fit to the dimer and cluster data; after which, the resulting fits were tested against the full data set. Table 2 indicates that the presence of the many-body terms in the repulsion (WH+SR) does not significantly improve the nonorthogonal model. Adding terms in the electronic part of the calculation, however, is more useful, and the errors for the Al_N data with $N = 2-13$ decrease by 24% and 37% for the SOD and SIP models, respectively, compared to the pairwise NTB fits. When the NMBTB models are tested for transferability against the larger cluster and nanoparticle data, the SOD model is the most accurate. However, even the best NMBTB method (WH+SOD) has an error that is 4−5 times greater than that of the simplest TB method.

Tight binding is sometimes implemented without explicit repulsion, and theoretical arguments may be given in support of this choice.[7,37] (Other workers find it more natural to retain explicit core repulsion.[38]) One may wonder if including an empirical repulsion during the fitting procedure is perhaps responsible for the lack of transferability of the NTB parametrizations. (Note that removing the repulsion cannot make the fitted errors decrease, as the resulting model is less flexible. We are simply testing the effect that the repulsion has on the transferability of the NTB parametrizations.) We fit three NTB WH models without repulsion (denoted "XR"), and the results are shown in Table 2. We find that the trends are similar for NTB with and without repulsion, and that, in fact, NTB with repulsion has slightly better transferability.

The CC TB model has no parameters in the electronic part of the calculation, and three CC NTB fits were obtained. The CC model without repulsion (CCXR) has no adjustable parameters, and its errors are also presented in Table 2. For the CC model, the repulsion does not significantly improve the fitting error or the transferability. The overall errors for the CC methods are larger than those for the best WH NTB fits.

Comparing the best TB to NTB parametrizations shows that the TB models are both more accurate and more transferable than the NTB ones. Several theoretical analyses of the effect of including the overlap in TB have appeared in the literature,[7,25,26,30−33,39−43] many of which focus on the bulk band structure. We do not attempt another such analysis here, but we do note that, whenever the effect of including overlap has been discussed, there is an implicit assumption or explicit working hypothesis, sometimes based on experience, that including overlap should make the model more realistic. This belief also underlies discussions of other forms of semiempirical molecular orbital theory; for example, the statement that overlap introduces many-body effects was used to justify its inclusion in a self-consistent-charge density-functional tight-binding scheme.[16,17] However, we know of no previous systematic test as extensive as the present one.

Although it would be hard to establish firm guidelines for the inclusion of overlap, we can illustrate the qualitative difference in the TB and NTB WH dimer, and this is done in Figures 1 and 2. Figure 1b shows that the electronic eigenvalues have a qualitatively different shape for NTB and TB, even for the dimer curve. For the TB model, at $R = 2.8$ Å, the 3s and 3p energy bands are split by 1.9 and 1.6 eV, respectively. For the NTB model, the bands split by 3.5 and 8.6 eV, respectively, and several of the energy levels tend toward positive numbers at small atom−atom separations. One can explain this qualitative difference as resulting from terms such as $(1 - S^{kk'}_{\mu\mu'})$ in the denominator of the NTB energy expressions, whereas these terms are absent in the TB model. For the aluminum dimer, the energy levels that tend to large positive numbers are not occupied (for zero penalty energy), the filled energy levels have similar shapes for the TB and NTB models, and both models are able to fit the aluminum dimer data with good accuracy.

Figure 2 shows the energetics for an FCC Al_{13} cluster as a function of the distance of all 12 surface atoms from the central atom for the same two fits as were shown in Figure 1. The TB total energy is fairly accurate, with the minimum occurring very close to the accurate minimum. The NTB total energy is qualitatively incorrect. Figure 2b shows the qualitative behavior of the eigenvalues, which is similar to the behavior of the eigenvalues for the dimer. In the case of Al_{13}, however, some of the repulsive electronic energy levels for NTB are occupied, leading to an NTB total energy that is too high for Al_{13}. It is interesting to note that the electronic energy E_{Elec} for the NTB method is in qualitative agreement with the accurate total energy E.

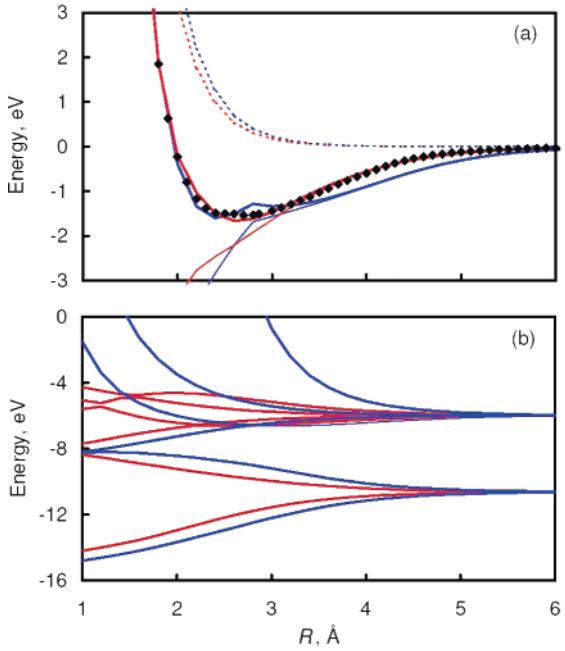


Figure 1. (a) Repulsive (thin dashed) and electronic (thin solid) components of the total energy (thick solid) for the aluminum dimer as a function of the nearest-neighbor distance R for TB (red) and NTB (blue) models. The accurate (PBEh/MG3 from ref 4) total energies are shown as black diamonds. (b) Electronic energy levels ϵ_j for TB (red) and NTB (blue) models of the aluminum dimer. Some of the levels are degenerate. For zero penalty energy, half the levels are filled.

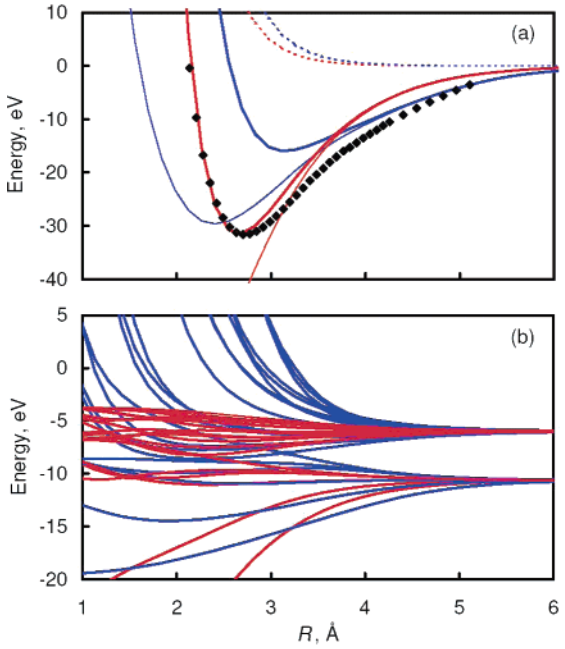


Figure 2. (a) Repulsive (thin dashed) and electronic (thin solid) components of the total energy (thick solid) for a face-centered cubic cluster of Al_{13} as a function of the nearest-neighbor distance R for TB (red) and NTB (blue). The accurate (PBEh/MG3 from ref 4) total energies are shown as black diamonds. (b) Electronic energy levels for TB (red) and NTB (blue) for the aluminum dimer. Some of the levels are degenerate. For zero penalty energy, half the levels are filled.

To gain further insight, we compare tight binding with DFT. The level of theory is unrestricted PBEh/MEC,[44,45] and the electronic (band) energy of DFT is defined as

$$E_{\mathrm{Elec}} = \sum_{\gamma=\alpha,\beta} \sum_{j} n_j^{\gamma} \epsilon_j^{\gamma} - E_0 \qquad (16)$$

where γ denotes spin, ϵ_j^{γ} is the DFT orbital eigenvalue, n_j^{γ} is the orbital occupancy (0 or 1), and E_0 sets the zero of energy. (Note that PBEh was formerly called PBE0.) We define the correction term to the electronic energy as

$$V_{\mathrm{Rest}} = E - E_{\mathrm{Elec}} - E_0 \qquad (17)$$

where E is the DFT total energy. The correction term includes core−core repulsion and the double counting correction for the electronic Coulomb interactions. We set E_0 such that V_{Rest} is zero for N infinitely separated atoms, that is, $E_0 = N(E^1 - E_{\mathrm{Elec}}{}^1)$, where E^1 and $E_{\mathrm{Elec}}{}^1$ are E and E_{Elec} for an isolated Al atom, respectively, and N is the number of atoms in the system.

Figure 3 shows V_{Rest} for PBEh/MEC and V_{Rep} for TB and NTB, scaled (as explained in the caption) by the reciprocal of the number N of atoms. Near equilibrium, V_{Rest} is approximately 3.7 and 2.5 times larger than V_{Rep} for TB and NTB, respectively. Some of the structure in the DFT Al_{13} curve is due to the lowest-energy solution changing its multiplicity as a function of nearest-neighbor separation.

Figure 4 shows the $4N$ lowest orbital energies (averaged over α and β spin for the case with an odd number of electrons and for distances where the dimer is a triplet) for

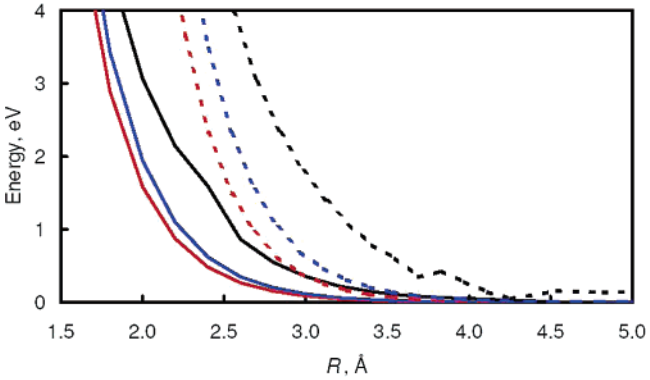


Figure 3. V_{Rep}/N for TB (red) and NTB (blue) and V_{Rest}/N for DFT (black) for aluminum dimer (solid) and FCC Al_{13} (dashed) as a function of nearest-neighbor distance R. The DFT method used for this figure is PBEh/MEC.

DFT (in particular, PBEh/MEC) calculations on Al_2 and for FCC Al_{13}. (Notes: (i) because MEC denotes the usage of an effective core potential, there are no core orbitals in a PBEh/MEC calculation, and this is equivalent to the $4N$ lowest-energy valence orbitals in an all-electron calculation; (ii) in counting to $4N$, degenerate orbitals are counted a number of times equal to their degeneracy, but we still have eight curves at some internuclear distances R for Al_2 because when the triplet is lower than the singlet for Al_2 we use the triplet solution, which breaks the degeneracy of the π orbitals in the unrestricted self-consistent-field formalism used here.) Quantitative comparisons between DFT and tight binding are not possible because of the different approximations

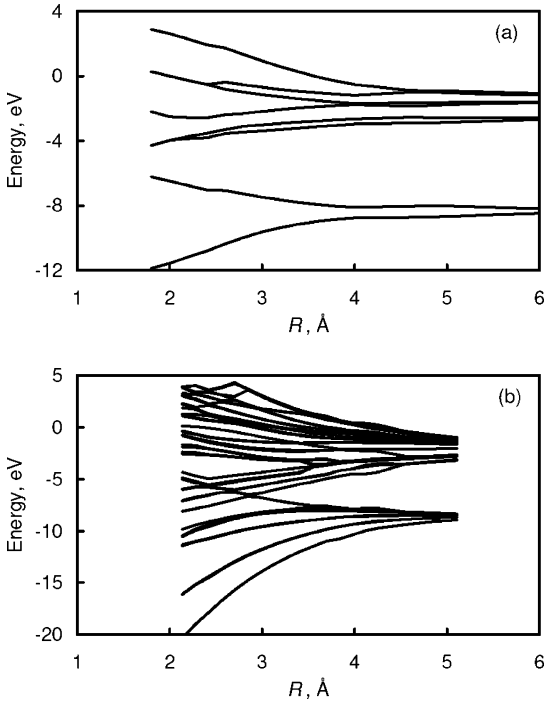


Figure 4. Electronic energy levels for DFT for (a) the aluminum dimer and (b) a face-centered cubic cluster of Al_{13} as a function of the nearest-neighbor distance R. The DFT method used for this figure is PBEh/MEC.

involved in the two theories. For example, for infinitely separated atoms, the TB and NTB energy levels give (by construction) the experimental ionization potentials, and the pairwise repulsion (again, by construction) goes to zero. In the DFT calculation, however, the energy levels for separated atoms include double counting of the Coulomb terms, and therefore, the correction term to the sum of the occupied orbital energies is not zero. We have taken account of this by our choice of the zero of energy in Figure 3, but there is no simple way to correct the individual levels shown in Figure 4. Also, the unoccupied orbitals in tight binding have a different interpretation than those in DFT. Despite these differences, we note that the qualitative shapes of the DFT energy levels in Figure 4 agree better with the orthogonal TB energy levels than the NTB energy levels, as shown in Figures 1b and 2b, although the DFT energy levels have a greater width (i.e., a larger second moment) than those for the TB method.

A final issue that we will look at is how well TB can predict the geometries of small Al clusters. The geometries of small Al_4 and Al_5 are planar,[46−48] and Al_6 is the smallest cluster to have a nonplanar ground state.[46] The ability to correctly predict the correct ground-state geometries of Al_4 and Al_5 with an analytical interatomic potential has proven to be very challenging. The embedded-atom-type models of Ercolessi and Adams[49,50] and Voter and Chen[51,52] predict that Al_4 and Al_5 are three-dimensional. We note that both of these interatomic potentials included the dimer in their fitting strategy. Pettersson et al.[53] have developed analytic interatomic potentials that correctly predict Al_4 and Al_5 to be planar, but the interatomic potential is unphysical for larger clusters. For example, the Pettersson et al.[53] interatomic

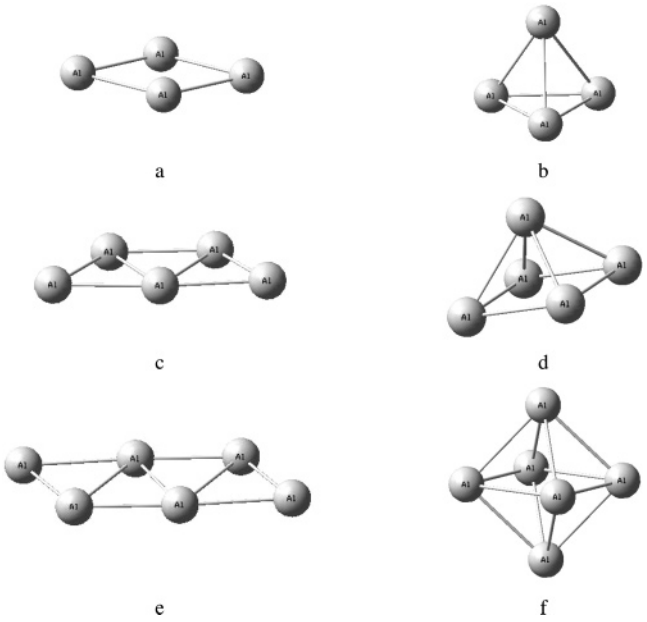


Figure 5. DFT geometries for planar and nonplanar Al_N ($N = 4-6$). The DFT method used for this figure is PBEh/MG3.

Table 4. ΔE (eV) for the Planar to Nonplanar Isomerizations with DFT, TB, and the NP-A and NP-B Interatomic Potentials

method	fitting data	Al_4	Al_5	Al_6
PBEh/MG3		0.24	0.29	−1.06
TB	2	0.18	−0.30	−0.87
	2−177	0.05	−0.26	−0.76
NP-A		−0.59	−1.18	−2.04
NP-B		−0.55	−0.89	−1.89

potential also predicts Al_{13} to be planar and has a 0.7 eV/atom error for the bulk cohesive energy.[53]

We have optimized planar and nonplanar structures for Al_4, Al_5, and Al_6 with the PBEh[44] density functional and the MG3 basis set.[54] The PBEh/MG3 structures are shown in Figure 5. In Table 4 we list the PBEh/MG3, TB, NP-A, and NP-B values for ΔE, and we note that a negative ΔE means that the nonplanar structure is energetically more favorable than the planar structure. The geometries for the TB, NP-A, and NP-B calculations were consistently optimized, and the PBEh/MG3 geometries were the starting points for the TB, NP-A, and NP-B geometry optimizations.

The NP-A and NP-B interatomic potentials both predict a three-dimensional structure for all three cluster sizes (Al_N with $N = 4-6$). The TB method that was fit to the full Al database, with $N = 2-177$, is qualitatively correct for Al_4, and the TB methods are much closer to the PBEh/MG3 values than to the results obtained with the interatomic potentials. These results are encouraging because the simple TB calculations are about three times more accurate than the best analytic functions.

VI. Conclusions

It is well-known that, in simplified methods, overlap-sensitive errors of a similar kind but opposite sign may cancel, and it is desirable to design approximate methods with this in

mind.[59] We have presented several parametrizations for various tight-binding models, and we have investigated the effect of including the overlap in the secular equation and of including many-body terms. We found that one of the simplest models (pairwise orthogonal TB fit to the dimer) is almost as accurate for the entire data set (consisting of 808 geometries from Al_2 to Al_{177}) as the most accurate of the TB fits (orthogonal TB WH+SR). As a general finding, the orthogonal TB model is more accurate than the nonorthogonal model. Many-body terms do not improve the fits in the orthogonal formulations, although they do result in some improvement in the NTB fits when applied to the off-diagonal Hamiltonian matrix elements.

Although the present tests of model types are extensive and systematic, and the results are clear-cut, the tests are restricted to a single element (the only element for which a well-validated data set of nanoparticle energies is available), and we cannot claim that it will always be more accurate to neglect overlap in semiempirical molecular orbital theory or even that it will always be more consistent to neglect it in the special case of tight binding. In fact, although there has been considerable success with orthogonal models, some workers have found, contrary to the results obtained here, that including overlap can reduce the number of parameters and make the parametrizations more transferable.[55,56] But the present results serve as a strong warning that one should not assume that removing approximations (like neglect of overlap) automatically and consistently improves approximate theories. Because including overlap in tight-binding calculations often considerably complicates them and raises the cost,[55,57,58] we recommend that researchers check carefully how much, if any, improvement is afforded by including overlap when they select a model for large-scale computations on nanoparticles, materials, or heterogeneous catalysis.

Acknowledgment. This work is supported in part by the Defense−University Research Initiative in Nanotechnology (DURINT) of the U.S. Army Research Laboratory and the U.S. Army Research Office under agreement number DAAD190110503 and also by the Office of Naval Research under award number N00014-05-1-0538. Computational resources were provided by the Minnesota Supercomputing Institute and by a Grand Challenge Grant at Pacific Northwest National Laboratories.

Supporting Information Available: Tables of the optimized parameters for several of the TB and NTB models. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Lennard-Jones, J. E. *Trans. Faraday Soc.* **1932**, *28*, 334.

(2) Daw, M. S.; Baskes, M. I. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1984**, *29*, 6443.

(3) Jasper, A. W.; Staszewski, P.; Staszewska, G.; Schultz, N. E.; Truhlar, D. G. *J. Phys. Chem. B* **2004**, *108*, 8996.

(4) Jasper, A. W.; Schultz, N. E.; Truhlar, D. G. *J. Phys. Chem. B* **2005**, *109*, 3915.

(5) Streitz, F. H.; Mintmire, J. W. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1994**, *50*, 11996.

(6) Schröder, K.-P.; Sauer, J. *J. Phys. Chem.* **1996**, *100*, 11043.

(7) Hoffmann, R. *J. Chem. Phys.* **1963**, *39*, 1397; **1964**, *40*, 2047, 2474, 2480, 2745.

(8) Slater, J. C.; Koster, G. F. *Phys. Rev.* **1954**, *94*, 1498.

(9) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

(10) Dewar, M. J. S.; Holder, A. J. *Organometallics* **1990**, *9*, 508.

(11) Bredow, T.; Jug, K. *Theor. Chem. Acc.* **2005**, *113*, 1.

(12) Schultz, N. E.; Staszewska, G.; Staszewski, P.; Truhlar, D. G. *J. Phys. Chem. B* **2004**, *108*, 4850.

(13) Wolfsberg, M.; Helmholtz, L. *J. Chem. Phys.* **1952**, *20*, 837.

(14) Staszewska, G.; Staszewski, P.; Schultz, N. E.; Truhlar, D. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2004**, *71*, 45423.

(15) Sattelmeyer, K. W.; Jørgensen, W. L. Preprint.

(16) Elstner, M.; Porezag, D.; Jungnickel, G.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *58*, 7260.

(17) Elstner, M.; Porezag, D.; Jungnickel, G.; Frauenheim, T.; Suhai, S.; Seifert, G. *Mater. Res. Soc. Symp. Proc.* **1998**, *491*, 131.

(18) Sattelmeyer, K. W.; Tubert-Brohman, I.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2006**, *2*, 413.

(19) Menon, M.; Richter, E.; Subbaswamy, K. R. *J. Chem. Phys.* **1996**, *104*, 5876.

(20) Jayanthi, C. S.; Wu, S. Y.; Cocks, J.; Luo, N. S.; Xie, Z. L.; Menon, M.; Yang, G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *58*, 3799.

(21) Bernstein, N. *Europhys. Lett.* **2001**, *55*, 52.

(22) Sternberg, M.; Galli, G.; Frauenheim, T. *Comput. Phys. Commun.* **1999**, *118*, 200.

(23) Repasky, M. P.; Chandresekhar, J.; Jørgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601.

(24) Wang, Y.; Mak, C. H. *Chem. Phys. Lett.* **1995**, *235*, 37.

(25) Foulkes, W. M. C.; Haydock, R. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1989**, *39*, 12520.

(26) Cusachs, L. C.; Cusachs, B. B. *J. Phys. Chem.* **1967**, *71*, 1060.

(27) Lathiotakis, N. N.; Andriotis, A. N.; Menon, M.; Connolly, J. *J. Chem. Phys.* **1996**, *104*, 992.

(28) Lide, D. R. *Handbook of Chemistry and Physics*, 78th ed.; CCR Press: Boca Raton, FL, 1997. Moore, C. E. *Atomic Energy Levels as Derived from the Analyses of Optical Spectra*; Circular of the National Bureau of Standards 467, U.S. Department of Commerce: Washington, DC, 1949; Vol. 1.

(29) Dorantes-Davila, J.; Pastor, G. M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1995**, *52*, 11837.

(30) Tang, M. S.; Wang, C. Z.; Chan, C. T.; Ho, K. M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *53*, 979.

(31) Haas, H.; Wang, C. Z.; Fähnle, M.; Elsässer, C.; Ho, K. M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *57*, 1461.

(32) Wang, C. Z.; Pan, B. C.; Tang, M. S.; Haas, H.; Sigalas, M.; Lee, G. D.; Ho, K. M. *Mater. Res. Soc. Symp. Proc.* **1998**, *491*, 211.

(33) Nguyen, D.; Pettifor, D. G.; Vitek, V. *Phys. Rev. Lett.* **2000**, *85*, 4136.

(34) Nguyen-Mahn, D.; Pettifor, D. G.; Vitek, V. *Phys. Rev. Lett.* **2000**, *85*, 4136.

(35) Microgenetic algorithm calculations were carried out using the program Fortran ga, version 1.6.4, as described in: Carroll, D. L. *AIAA J.* **1996**, *34*, 338.

(36) Gollisch, H. *Surf. Sci.* **1986**, *166*, 87.

(37) Papaconstantoupoulos, D. A.; Mehl, M. J. *J. Phys.: Condens. Matter* **2003**, *15*, R413.

(38) Pople, J. A.; Santry, D. P.; Segal, G. A. *J. Chem. Phys.* **1965**, *43*, 5129.

(39) Del Re, G.; Ladik, J.; Carpentieri, M. *Acta Phys. Acad. Sci. Hung.* **1968**, *24*, 391.

(40) Del Re, G.; Molář, M.; Cyrot-Lackmann, F. *J. Phys. (Paris)* **1985**, *46*, 927.

(41) Skinner, A. J.; Pettifor, D. G. *J. Phys.: Condens. Matter* **1991**, *3*, 2029.

(42) Pettifor, D. G.; Aoki, M. In *Equilibrium Structure and Properties of Surfaces and Interfaces*; Gonis, A., Stocks, G. M., Eds.; Plenum: New York, 1992; pp 123−137.

(43) Canel, L. M.; Carlsson, A. E.; Fedders, P. A. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1993**, *48*, 10739. McKinnon, B. A.; Choy, T. C. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1995**, *52*, 14531. Yang, S. H.; Mehl, M. J.; Papaconstantopoulos, D. A. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *57*, R2013. Mehl, M. J.; Papaconstantopoulos, D. A. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2000**, *61*, 4894. Frøseth, A. G.; Holmestad, R.; Derlet, P M.; Marthinsen, K. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2003**, *68*, 12105. Rittenhouse, S. T.; Johnson, B. L. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2005**, *71*, 35118.

(44) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865; **1997**, *78*, 1396 (E).

(45) Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 41.

(46) Bauschlicher, C. W., Jr.; Pettersson, L. G. M. *J. Chem. Phys.* **1987**, *87*, 2198.

(47) Geske, G.; Boldyrev, A. I.; Li, X.; Wang, L.-S. *J. Chem. Phys.* **2000**, *113*, 5130.

(48) Zhan, C.-G.; Zheng, F.; Dixon, D. A. *J. Am. Chem. Soc.* **2004**, *124*, 14795.

(49) Ercolessi, F.; Adams, J. *Europhys. Lett.* **1993**, *26*, 583.

(50) Doye, J. P. K. *J. Phys. Chem.* **2003**, *119*, 1136.

(51) Voter, A. F.; Chen, S. P. *Mater. Res. Soc. Symp. Proc.* **1987**, *82*, 175.

(52) Sebetci, A.; Güvenç, Z. B. *Modell. Simul. Mater. Sci. Eng.* **2005**, *13*, 683.

(53) Pettersson, L. G. M.; Bauschlicher, C. W., Jr.; Halicioglu, T. *J. Chem. Phys.* **1987**, *87*, 2205.

(54) Fast, P. L.; Sánchez, M. L.; Truhlar, D. G. *Chem. Phys. Lett.* **1999**, *306*, 407.

(55) Dorantes-Dávila; Pastor, G. M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1994**, *51*, 16627.

(56) Menon, M.; Subbaswamy, K. R. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1997**, *55*, 9231.

(57) Jayanthi, C. S.; Wu, S. Y.; Cocks, J.; Luo, N. S.; Xie, X. L.; Menon, M.; Yang, G. *56. Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *57*, 3799.

(58) Steinberg, M.; Galli, G.; Frauenheim, T. *Comput. Phys. Commun.* **1999**, *118*, 200.

(59) Pople, J. A.; Segal, G. A. *J. Chem. Phys.* **1966**, *44*, 3289.

CT600261S

JCTC Journal of Chemical Theory and Computation

Theoretical Investigation of Excited States of Large Polyene Cations as Model Systems for Lightly Doped Polyacetylene

Ulrike Salzner*

Department of Chemistry, Bilkent University, 06800 Bilkent, Ankara, Turkey

Received July 6, 2006

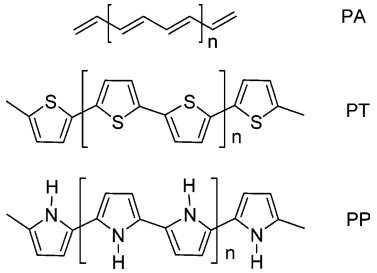
Abstract: Electronic excitations of polyene cations with chain lengths of up to 101 CH units were investigated as model systems for lightly doped polyacetylene (PA). Since high level ab initio calculations such as complete active space perturbation theory (CASPT2) are limited to systems with about 14 CH units, the performances of time-dependent Hartree−Fock (TDHF) and time-dependent density functional theory (TDDFT) were evaluated. It turned out that TDDFT excitations energies are much more accurate for polyene cations than for neutral polyenes. The difference between TDHF and TDDFT excitation energies for the first allowed excited state of $C_{49}H_{51}^{+}$ is only 0.30 eV with pure DFT and 0.21 eV with a hybrid functional. For open-shell systems, pure DFT is found to be superior to DFT-hybrid functionals because it does not suffer from spin-contamination. Pure TDDFT excitation energies and oscillator strengths for small open-shell polyene cations compare well with high level ab initio results. Excitation energies are found to be almost independent of the geometry, i.e., the size of the defect. Localization of the defect, however, shifts oscillator strengths from the HOMO−LUMO transition to higher lying excited states of the same symmetry. Lightly doped PA is predicted to exhibit several strong absorptions below 1 eV.

Introduction

Polyenes are model compounds for polyacetylene (PA) that can be used to investigate intrachain contributions to optical and electronic properties of conducting polymers. PA (Scheme 1) is a semiconductor that increases its conductivity to 10^5 S/cm when it is oxidized or p-doped.[1] The changes induced during doping can be monitored by in situ UV-spectroscopy. Undoped PA has a strong absorption peak at 1.9 eV, which is attributed to an intrachain $\pi-\pi^*$-transition. Upon doping with iodine or arsenic pentafluoride, a new broad band between 0.65 and 0.75 eV appears, while the intraband transition at 1.9 eV decreases in intensity.[2] There is no shift of the $\pi-\pi^*$-transition during doping.[3] Effects of donor doping with Na are indistinguishable form those of acceptor doping.[4]

Heterocyclic conjugated organic polymers such as polythiophene (PT) and polypyrrole (PP) (Scheme 1) behave somewhat different. Doping with perchlorate leads to decrease of the intensity of the $\pi-\pi^*$-transition at 2.7 eV of neutral PT and a shift of the interband transition to higher energy. At the same time two new features at 0.7−0.9 and 1.5−1.8 eV are produced.[5] Electrochemically produced PP is obtained in its oxidized form and shows two absorption bands at 1.0 and 2.7 eV. During reduction, a band appears at 3.6 eV that moves to lower energy. The neutral polymer has its maximum absorbance at 3.2 eV.[6] The shifting of the

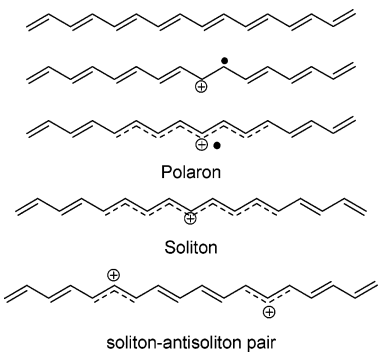
Scheme 1



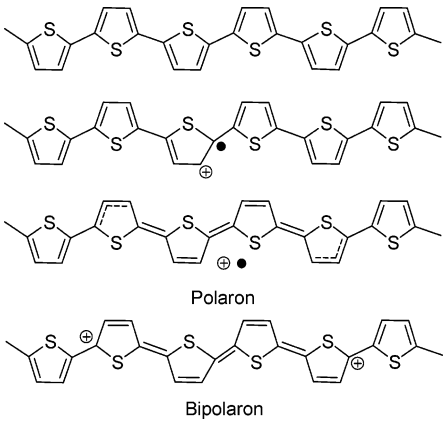
* Corresponding author e-mail: salzner@fen.bilkent.edu.tr.

10.1021/ct600223f CCC: $37.00
Published on Web 12/01/2006

Scheme 2



Scheme 3



maximum absorption peak was rationalized by considering that the material is polydisperse and that shorter chains are reduced first.[6]

Optical properties of conducting polymers have been explained in terms of self-localized defects called solitons, polarons, and bipolarons.[5,7−9] Solitons form only on polymers with degenerate ground states, such as PA. Degenerate ground state means that there is no energetic difference between the two forms, in which positions of single and double bonds are switched. Upon doping, solitons and soliton−antisoliton pairs are generated that are supposed to give rise to one interband transition in the optical spectrum (Scheme 2). Polymers with nondegenerate ground states such as PT and PP become quinoid when the positions of single and double bonds are exchanged (Scheme 3). Since quinoid forms are less stable for the neutral polymer, PT and PP do not form solitons but polarons and bipolarons (Scheme 3). Polarons are believed to be associated with three interband transitions, bipolarons with two.[5,7−9] Since PT and PP develop two sub-band transitions upon doping and have very low ESR signals in the conducting state, they are assumed to form spinless bipolarons.

The existence of bipolarons has been questioned in experimental studies on polyene[10] and on oligothiophene cations.[11−14] Spectroscopic investigations in dilute solution show that dimers form at low temperature and that cations show two but dications only one absorption in the UV and near IR region. This is confirmed by high level ab initio studies of excited states of polyene[15,16] and oligothiophene cations.[16−18] Theoretical calculations, mostly employing semiempirical and Hartree−Fock methods, have confirmed the presence of self-localization in the presence and in the absence of counterions.[19−34] At higher levels of theory, the size of the defects in the absence of counterions tends to increase[35] and bipolarons become unstable with respect to two polarons.[36−38] As an alternative to bipolaron formation, π-dimers of radical cations have been shown to be stable when solvent effects are included in the theoretical treatment.[39,40]

Since experimental and theoretical studies raise some doubt about the nature of the sub-band transitions in doped conducting polymers and since high level theoretical investigations were carried out only up to decapentaene and terthiophene cations, it seems worthwhile to have a fresh look at optical properties of conducting polymers employing ab initio methods and density functional theory (DFT). The aim is to investigate effects of doping on the geometry and electronic structure and to analyze the influence of geometry changes on UV spectra of long oligomers with degenerate and with nondegenerate ground states. Finally the question why PA shows one sub-band feature in the UV spectrum while most other conducting polymers show two is addressed. Toward this aim detailed investigations of positively and negatively charged polyenes, thiophene and pyrrole oligomers in the absence and in the presence of counterions were initiated. Self-localization is investigated on longer oligomers than previously, and the effect of localization on excitation spectra is studied. In the present investigation an appropriate theoretical level for calculating excited states of long conjugated cations is searched for. It will be shown that time-dependent density functional theory (TDDFT), which was shown to be promising for open-shell systems,[41] does not deteriorate for cations in the long chain limit as it does for neutral polyenes.[42−47] Therefore polyene cations with chain lengths up to $C_{101}H_{103}^{+}$ could be treated. It turned out that acetylene, thiophene, and pyrrole oligomer cations give rise to two low-energy transitions. The high-energy feature is strong; the low-energy feature is weak. The difference between PA and PT or PP is that only for PA the energy difference between highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) is almost identical for α- and β-electrons in open-shell monocations. As a result, the intensity of the low-energy feature of PA is so small that the peak is invisible, and the spectrum shows only one sub-band transition. In this investigation acetylene oligomers are discussed in detail. Oligothiophenes, oligopyrroles, dications, and n-doped species will be considered in forthcoming publications.

Methods

Polyene cations were investigated with DFT and with ab initio methods. Only one kind of basis set, Stevens-Basch-Krauss pseudopotentials in combination with polarized split valence basis set (CEP-31G*),[48−50] was used throughout. This basis set has been tested extensively and has been shown to yield reliable results for geometries and excited states of neutral polyenes as well as other conjugated oligomers.[43,51−54] As cations and excited states of cations have compact electron densities,[15] this basis set is sufficient for all properties of interest in the present investigation.

Table 1. Bond Lengths of $C_{13}H_{15}$ and $C_{14}H_{16}$ Starting at the Chain Ends at Different Levels of Theory

CASSCF		pure DFT		DFT-hybrid		MP2	
$C_{13}H_{15}$	$C_{14}H_{16}$	$C_{13}H_{15}$	$C_{14}H_{16}$	$C_{13}H_{15}$	$C_{14}H_{16}$	$C_{13}H_{15}$	$C_{14}H_{16}$
1.373	1.370	1.387	1.382	1.371	1.363	1.356	1.377
1.469	1.475	1.455	1.461	1.449	1.460	1.464	1.469
1.385	1.375	1.406	1.396	1.391	1.373	1.382	1.385
1.451	1.471	1.438	1.450	1.429	1.451	1.435	1.459
1.403	1.376	1.417	1.400	1.405	1.376	1.402	1.388
1.428	1.469	1.427	1.447	1.416	1.448	1.417	1.457
	1.376		1.401		1.377		1.389

For ab initio calculations, Møller-Plesset perturbation theory[55] up to second order (MP2) and the complete-active-space self-consistent-field (CASSCF) method[56−61] were employed. The size-limit for CASSCF for us is 14 electrons and 14 orbitals for closed-shell systems and 13 electrons and 13 orbitals for open-shell species. For the largest even-numbered open-shell cation $C_{14}H_{16}^{+}$ that could be treated, it was therefore necessary to remove the highest unoccupied π-orbital from the active space. Comparison with $C_{12}H_{14}^{+}$ that could be treated fully shows that the error introduced by using the limited active space is small. With MP2 our size limit is around $C_{45}H_{47}^{+}$ for closed-shell species. With DFT it is possible to optimize species with well over 100 CH units, so that all structures and defects are converged with respect to chain length.

For DFT calculations, two types of functionals were compared. For closed-shell systems, Becke's three-parameter hybrid functional[62] with 30% of Hartree−Fock exchange[53] combined with Perdew's 86 correlation functional (P86)[63] works well. Because of problems with spin-contamination, excitation energies of open-shell species are better modeled without Hartree−Fock exchange. For these calculations, Becke's gradient corrected functional[64] combined with P86 was used. The two functionals will be referred to as B3P86-30% or DFT-hybrid and BP86 or pure DFT. Excited-state calculations were carried out at the time-dependent Hartree−Fock (TDHF) and (TDDFT) levels.[42,65−67] TDHF and TDB3P86-30% excited-state calculations were done on the same geometries, optimized at B3P86-30%/CEP-31G* since there is a dependence of excitation energies on bond length alternation (BLA). Since HF overestimates BLA,[68] excitation energies are overestimated when HF geometries are used.[44] Pure-DFT calculations on cations are based on BP86 and on B3P86-30% geometries. All calculations were carried out with Gaussian 03 revisions C0.2[69] and D0.1.[70]

Results and Discussion

Geometries. *Neutral Polyenes*. To compare the performance of CASSCF, MP2, and DFT for open- and closed-shell systems, $C_{13}H_{15}$ and $C_{14}H_{16}$ were optimized in their neutral states (Table 1). For the CASSCF calculation all π-orbitals and π-electrons are included in the active space. CASSCF includes only nondynamic correlation, MP2 includes mainly dynamic correlation, and DFT accounts for some of both. For closed-shell systems, MP2 should be the best of the three methods, since it is known to give excellent geometries for closed-shell ground states.[71] For open-shell species, MP2 is not reliable because it is based on the spin-contaminated HF wave function. Pure DFT is known to underestimate bond lengths alternation (BLA)[68,72−74] but does not suffer from spin-contamination. Hybrid functionals are extremely popular since many of the shortcomings of pure DFT can be ameliorated by adding a certain amount of HF exchange. The most used hybrid functional is B3[62] which contains 20% of HF exchange. However, B3 in connection with Lee−Yang−Parr (LYP)[75] correlation functional was found to generate excessively delocalized defects on PA chains.[35] It has also been shown previously that increasing the amount of HF exchange improves the performance of DFT-hybrid with respect to band gaps.[52,53,68] Therefore 30% of HF exchange will be used throughout.

With DFT single and double bond lengths are both about 0.01 Å shorter than with MP2. Double bond lengths are actually similar to those at CASSCF. Pure DFT produces identical single bond lengths compared to DFT-hybrid but longer double bonds. There are no experimental data for $C_{13}H_{15}$ and $C_{14}H_{16}$. For hexatriene and octatetraene the experimental central double lengths are 1.368 Å and between 1.327 and 1.354 Å according to Choi et al.[68] That the double bond in octatetraene is shorter than in hexatriene is counterintuitive and indicates that there might be substantial experimental uncertainty. The DFT-hybrid/CEP-31G* values are 1.368 and 1.371 Å. The experimental single bond values are 1.457 for hexatriene and between 1.435 and 1.451 Å for octatetraene.[68] The DFT-hybrid/CEP-31G* values are 1.463 and 1.455 Å. Thus the present DFT-hybrid geometries are in reasonable agreement with experiment, theoretical bond lengths tending to be a bit longer.

Comparing results at CASSCF and MP2 levels for closed-shell $C_{14}H_{16}$ shows that BLA is overestimated at CASSCF (0.093 Å) compared to MP2 (0.068 Å). The difference arises in equal amounts from shorter single and longer double bonds. With DFT-hybrid, BLA (0.071 Å) is similar to that at MP2. For long polyenes, BLA converges to 0.058 Å with MP2, to 0.062 Å with DFT-hybrid, and to 0.023 with pure DFT. The experimental value for PA is about 0.08 Å.[76,77]

Comparing open-shell $C_{13}H_{15}$ with closed-shell $C_{14}H_{16}$ at the CASSCF level indicates that radicals have longer double and shorter single bonds than closed-shell polyenes. The difference for $C_{13}H_{15}$ and $C_{14}H_{16}$ is 0.04 Å for the terminal double bond. With the DFT-hybrid functional, the radical has a 0.08 Å longer terminal double bond. In contrast, the terminal double bond contracts by 0.21 Å at the MP2 level. The difference between CASSCF and DFT-hybrid is probably due to the more delocalized defect with DFT. The strong contraction of the double bond at MP2 is unexpected. Considering the high spin contamination of the MP2 wave function for $C_{13}H_{15}$ with an expectation value of the $<S^2>$ operator of 2.35 instead of 0.75, the MP2 result seems unreliable.

Odd-Numbered Polyene Cations. Odd-numbered polyene cations were investigated in the absence and in the presence of one Cl_3^- counterion. Cl_3^- was chosen as a model for iodine doping, which involves I_3^- and I_5^- ions.[78] Cl_3^- also forms during geometry optimization when three separate Cl

Table 2. Bond Lengths of $C_{13}H_{15}^+$ and $C_{14}H_{16}^+$ Starting at the Chain Ends at Different Levels of Theory

CASSCF		pure DFT		DFT-hybrid		MP2	
$C_{13}H_{15}^+$	$C_{14}H_{16}^+$	$C_{13}H_{15}^+$	$C_{14}H_{16}^+$	$C_{13}H_{15}^+$	$C_{14}H_{16}^+$	$C_{13}H_{15}^+$	$C_{14}H_{16}^+$
1.372	1.374	1.386	1.389	1.367	1.374	1.382	1.355
1.464	1.462	1.453	1.449	1.448	1.441	1.455	1.456
1.385	1.389	1.406	1.411	1.387	1.397	1.399	1.382
1.440	1.437	1.435	1.431	1.427	1.420	1.433	1.422
1.398	1.403	1.417	1.420	1.401	1.407	1.410	1.403
1.417	1.415	1.425	1.425	1.413	1.411	1.420	1.415
	1.406		1.423		1.410		1.418

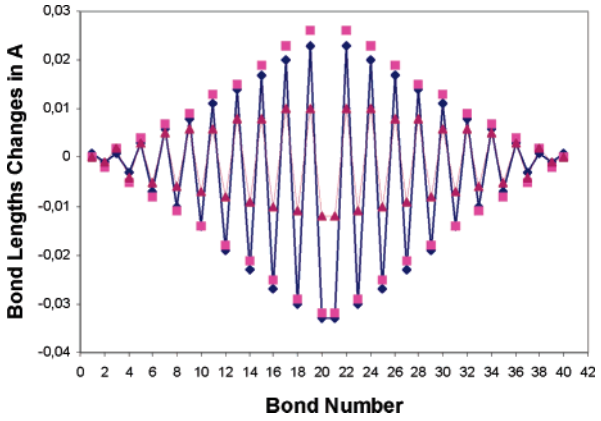


Figure 1. Changes in bond lengths of $C_{41}H_{43}^+$ compared to neutral $C_{40}H_{42}$ at MP2 (diamonds), DFT-hybrid (squares), and pure DFT (triangles) levels of theory.

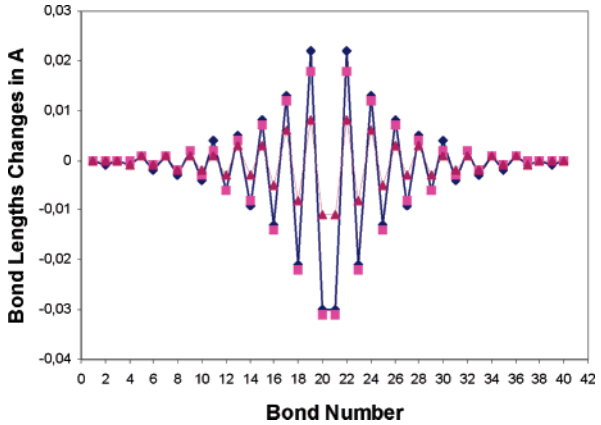


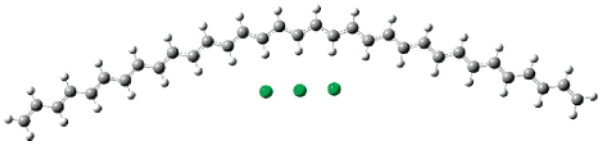
Figure 2. Changes in bond lengths of $C_{41}H_{43}-Cl_3$ compared to neutral $C_{40}H_{42}$ at MP2 (diamonds), DFT-hybrid (squares), and pure DFT (triangles) levels of theory.

atoms are placed next to chain. Only one kind of counterion was used here since experimental results[2,4] as well as test calculations agree that changes upon doping are independent of the nature of the dopant ion. Odd-numbered polyene cations and polyene-Cl_3 complexes are closed-shell singlets. Therefore MP2 geometries are expected to be reliable.

Geometries of $C_{13}H_{15}^+$ at CASSCF, pure DFT, DFT-hybrid, and MP2 are summarized in Table 2. Removing one electron from $C_{13}H_{15}$ converts the radical into a closed-shell cation. This change is associated with relatively small changes in bond lengths. Comparing data from Tables 1 and 2 shows that CASSCF, pure DFT, and DFT-hybrid agree that bond lengths tend to decrease upon ionization. The changes decrease in the order CASSCF < DFT-hybrid < pure DFT. Only MP2 predicts substantial increases in certain bond lengths. This problem is caused by problems of MP2 with the $C_{13}H_{15}$ radical. Compared to MP2, CASSCF predicts longer single and shorter double bonds for $C_{13}H_{15}^+$. This shows that dynamic correlation is necessary to predict accurate BLA. CASSCF geometries are therefore not very accurate but useful to evaluate closed- and open-shell systems on equal footing. Like for neutral polyenes, DFT-hybrid and MP2 predict similar BLA for $C_{13}H_{15}^+$. MP2 and DFT-hybrid are the methods of choice for closed-shell cations.

In Figures 1 and 2, changes in bond lengths compared to neutral $C_{40}H_{42}$ are plotted for $C_{41}H_{43}^+$ in the absence and in the presence of one Cl_3^- counterion at MP2/CEP-31G*, BP86-30%/CEP-31G*, and BP86/CEP-31G* levels of theory. The counterion is placed in the molecular plane in the vicinity of the hydrogen atoms (Scheme 4), since this conformation

Scheme 4



is lower in energy than the one with the counterion above the chain. The structures are fully optimized within C_{2v} symmetry. Charge transfer between polyene and counterion is virtually complete with natural populations (NPA)[79] of +0.97 e (MP2) and +0.96 e (DFT-hybrid) on the polyene chain. Figures 1 and 2 show that MP2 and DFT-hybrid lead to very similar changes in bond lengths. In the absence of a counterion, all bonds are involved in changes, but the major effect is observed at the center of the molecule. Pure-DFT gives a similar result at the chain ends and therefore a similar defect size. The response to charging in the middle of the chain is significantly less with pure DFT than with the other two methods. The reason for this is that neutral $C_{40}H_{42}$ is more delocalized with pure DFT, having BLA of only 0.023 Å (compared to 0.058 Å and 0.062 Å with MP2 and DFT-hybrid). Compared to DFT-hybrid the decreased BLA is due to longer double bonds.

In the presence of Cl_3^-, the defect is more localized. With MP2 the first five bonds are unaffected, with DFT, hybrid and pure, the first three bonds show no changes and the next four change very little. Defect sizes are therefore again similar with all three methods. At the center of the complex, DFT-hybrid leads to a slightly larger increase of the double

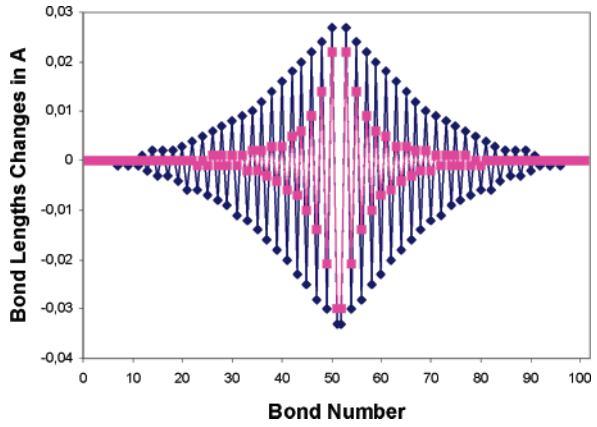


Figure 3. Changes in bond lengths of $C_{101}H_{103}^{+}$ (diamonds) and $C_{101}H_{103}-Cl_3$ (squares) compared to neutral $C_{102}H_{104}$ with DFT-hybrid.

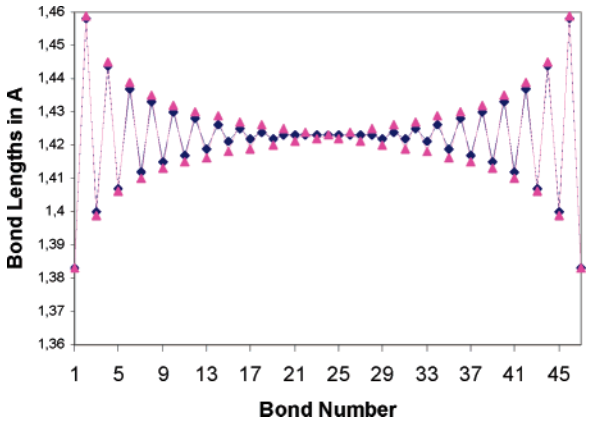


Figure 4. Bond lengths of $C_{48}H_{50}^{+}$ (diamonds) and $C_{49}H_{51}^{+}$ (triangles) with pure DFT.

bond lengths than MP2, and pure DFT leads to a much smaller response because of the small BLA of $C_{40}H_{42}$.

The defect size in the absence of a counterion was determined to be about 36 CH units at the MP2/6-31G level. B3LYP produced excessively delocalized defects.[35] In contrast, DFT employing a hybrid functional with 30% of HF exchange provides similar defect sizes compared to MP2. For systems longer than $C_{41}H_{43}$, MP2 becomes prohibitively expensive, and DFT-hybrid geometries can be used without loss of accuracy. For example in Figure 3, the defect size is evaluated for $C_{101}H_{103}^{+}$ and $C_{101}H_{103}-Cl_3$ at the DFT-hybrid level. For the naked cation, the first six bonds are completely unchanged compared to those in neutral $C_{102}H_{104}$. The next five bonds show negligible changes. Thus the defect spreads over about 80 CH units in the absence of a counterion. With counterion, the first 21 bond lengths are identical with those in neutral $C_{102}H_{104}$. The next 10 bonds change by only ± 0.001 Å. This indicates that the defect size is about 40 CH units.

Since more localized defects are obtained with HF, MP2/6-31G, and with semiempirical methods[7,16,20,23,29,35] and since forces during the geometry optimizations are very small, it seemed plausible that the differences are caused by very flat potential energy surfaces rather than the failure of certain methods. To investigate how big the energy lowering due to localization is, the counterions were removed, and single point energies were calculated on the geometries optimized in the presence of counterions. With DFT the energy difference is 1.66 kcal/mol for $C_{33}H_{35}^{+}$ and 1.89 kcal/mol for $C_{41}H_{43}^{+}$. At the MP2 level, the energies of the $C_{33}H_{35}^{+}$ with localized and delocalized defect differ by 2.14 kcal/mol or 0.06 kcal/mol per CH unit. Thus a segment size of about 15 CH units is required to obtain an energy difference of 1 kcal/mol. Extremely high levels of theory are required, if chemical accuracy of 1 kcal/mol is to be achieved. It is therefore not surprising that semiempirical methods and HF theory give results that differ from those at higher theoretical levels. With energy differences that small, theoretical methods well beyond MP2 are required to give a final answer about the size of the defects.[74] Such calculations are out of the question for systems long enough to have defect sizes converged with respect to chain lengths. The same sobering conclusion was drawn for neutral soliton defects in PA.[80] Thus in contrast to the generally believed strong electron phonon coupling,[27] high level ab initio calculations point toward very flat potential energy surfaces of doped conducting polymers.

Even-Numbered Polyene Cations. Calculating even-numbered long chain polyene cations is problematic at the MP2 level and with DFT-hybrid because spin-contamination increases dramatically with increasing chain length. CASSCF and pure DFT do not suffer from spin-contamination but the former overestimates, the latter underestimates BLA. Therefore, one should expect CASSCF and pure DFT to bracket the changes upon ionization. CASSCF calculations including 13 π-electrons and 13 π-orbitals in the active space were carried out for $C_{14}H_{16}^{+}$. Results at CASSCF, pure DFT, DFT-hybrid, and MP2 are compared with those for $C_{13}H_{15}^{+}$ in Table 2.

Removing one electron form $C_{14}H_{16}$ converts a closed-shell system to a radical cation. This leads to significant changes in bond lengths. CASSCF, pure DFT, and DFT-hybrid predict decrease in BLA, lengthening of the double bonds and shortening of the single bonds, especially in the center of the molecule. MP2 is the only method that predicts a shortening of the terminal double bond. Again MP2 shows erratic behavior for the radicals, which is most likely due to spin contamination. The spin expectation value for $C_{14}H_{16}^{+}$ is 1.96 with MP2.

Comparing the geometries of $C_{13}H_{15}^{+}$ and $C_{14}H_{16}^{+}$ at CASSCF and pure DFT levels indicates that the first two bonds from the chain ends differ very little $\sim$0.002 Å. In the center $C_{14}H_{16}^{+}$ shows slightly larger BLA. The maximum difference between any two bond lengths is 0.005 Å with both CASSCF and pure DFT. With DFT-hybrid, the trend is similar only the differences between $C_{13}H_{15}^{+}$ and $C_{14}H_{16}^{+}$ are larger. Thus DFT-hybrid seems to work fairly well for short polyene cation radicals. The spin expectation value for $C_{14}H_{16}^{+}$ is 0.92. Since spin contamination increases with chain length, DFT geometries were evaluated for longer oligomers. Bond lengths for open-shell $C_{48}H_{50}^{+}$ and closed-shell $C_{49}H_{51}^{+}$ are summarized in Figure 4 (pure DFT) and Figure 5 (DFT-hybrid). Like for the shorter systems, pure DFT predicts very similar geometries for cation and radical cation. In contrast, the hybrid functional yields a totally

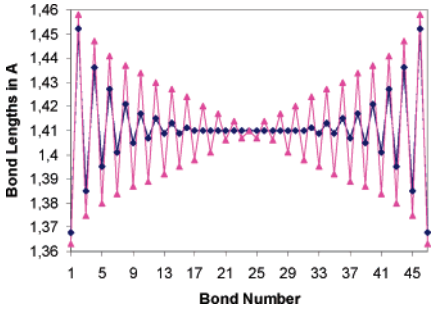


Figure 5. Bond lengths of $C_{48}H_{50}^{+}$ (diamonds) and $C_{49}H_{51}^{+}$ (triangles) with DFT-hybrid.

different geometry for $C_{48}H_{50}^{+}$ ($<S^2> = 2.11$) than for $C_{49}H_{51}^{+}$. It is noteworthy that DFT-hybrid predicts a slightly larger defect for the radical cation than pure DFT. Since pure DFT is consistent with CASSCF for the small systems and since it does not suffer from spin contamination, the pure DFT result seems to be more reliable, although BLA is certainly underestimated.

Excitation Energies. Short polyenes[81−83] and their cations[15] have been investigated with high level ab initio methods such as CASPT2 and MRMP. The first optically allowed excited state of neutral polyenes has ^1B_u symmetry and is dominated by a single-electron (HOMO−LUMO) excitation. There is a second forbidden transition of ^1A_g symmetry that becomes the first excited state between hexatriene and octatetraene. The ^1A_g state is one-electron forbidden and includes double excitations. This state cannot be treated reliably with single electron methods[39] and will not be considered here. Like neutral polyenes, odd-numbered cations have closed-shell ground states, and the first excitation energy to the $^1B_2$ state is mainly due to a single-electron HOMO−LUMO transition. The first optically allowed excited state of radical cations is produced by two single-electron transitions, one from the singly occupied molecular orbital (SOMO) to the LUMO and one from HOMO to SOMO. These two transitions correspond to separate HOMO−LUMO transitions in α- and β-electron spaces of the open-shell wavefunction. The two states mix with same and with opposite signs, generating a dipole-forbidden and a dipole-allowed excited state of ^2A_u or ^2B_g symmetry.[15] As a result there are two features in the optical spectrum, the one at lower energy is weak (dipole-forbidden), the one at higher energy is dipole-allowed and has high intensity.[84] The same state mixing happens between α- and β-electron transitions in closed-shell systems. The difference is that in a closed shell system, the cancellation is complete for the dipole forbidden state and its oscillator strength is exactly zero. For this reason neutral polyenes and odd-numbered cations show only one feature in the optical spectrum.

For neutral polyenes, the ^1B_u state is well described by TDHF. The results are of comparable accuracy to CASPT2 and MRMP predictions.[43,44,85] TDDFT, pure as well as hybrid, increasingly underestimates excitation energies of conjugated systems with increasing chain length.[42,45,46,86−88] Since high level ab initio methods become prohibitively expensive for long chain cations, TDHF and TDDFT will be tested here for excited-state calculations on polyene cations.

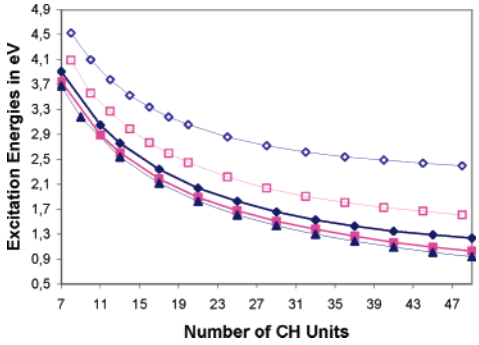


Figure 6. Excitation energies for neutral polyenes C_8H_{10}−$C_{48}H_{50}$ (open symbols) and polyene cations $C_7H_9^{+}$−$C_{49}H_{51}^{+}$ (filled symbols) at TDHF (diamonds), TDDFT-hybrid (squares), and pure TDDFT (triangles).

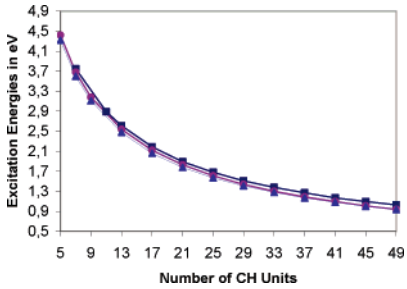


Figure 7. TDDFT-hybrid//DFT-hybrid (squares), pure TDDFT//DFT-hybrid (circles), and pure TDDFT//pure TDDFT (triangles) excitation energies for $C_5H_7^{+}$−$C_{49}H_{51}^{+}$.

Odd-Numbered Closed-Shell Polyene Cations. The first issue of interest is the behavior of TDHF and TDDFT excitation energies with increasing chain lengths. In Figure 6 excitation energies of neutral polyenes (C_6H_8−$C_{48}H_{50}$) and of monocations ($C_5H_7^{+}$−$C_{45}H_{47}^{+}$) at TDHF, TDDFT-hybrid, and pure TDDFT levels are plotted versus chain length. In contrast to neutral polyenes, where the TDDFT excitation energies are too low and fall off too fast with increasing chain lengths, TDHF and TDDFT excitation energies and their chain length dependence are similar for cations. The difference between TDDFT-hybrid and TDHF excitation energies for $C_{49}H_{51}^{+}$ is 0.21 eV as compared to 0.79 eV for neutral $C_{48}H_{50}$. Moreover, there is very little difference between results with the hybrid functional and with pure-TDDFT (0.09 eV). This is investigated in more detail in Figure 7, where TDDFT excitation energies are plotted vs chain lengths using TDDFT-hybrid, pure TDDFT on the hybrid geometry, and finally pure TDDFT on the pure DFT geometry. That the three curves lie almost on top of one another shows that HF exchange has little influence on TDDFT excitation energies and that the geometry does not influences the 1 ^1B_u excitation energies either. Thus TDHF and TDDFT, pure and hybrid, appear to be reliable for calculating 1 ^1B_u excitation energies of closed-shell cations even at long chain lengths.

Even-Numbered Open-Shell Polyene Cations. The preceding chapters have revealed that DFT geometries of open-shell systems have defect sizes which are too delocalized with DFT-hybrid. The comparison of excitation energies based on DFT-hybrid and pure DFT geometries for closed-shell cations has shown, however, that employing delocalized

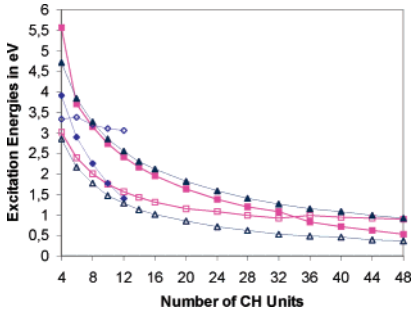


Figure 8. Excitation energies for $C_4H_8^+-C_{48}H_{50}^+$, dipole-forbidden states (open symbols), dipole-allowed states (filled symbols) at TDHF (diamonds), TDDFT-hybrid (squares), and pure TDDFT (triangles).

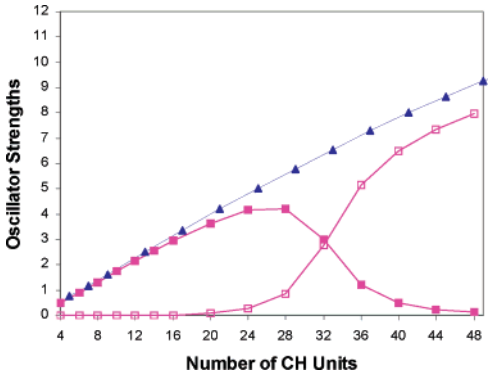


Figure 9. TDDFT-hybrid oscillator strengths for closed-shell cations $C_5H_7^+-C_{49}H_{51}^+$ (triangles) and for open-shell cations $C_4H_8^+-C_{48}H_{50}^+$ (squares).

pure DFT geometries does not influence the excitation energies very much (compare Figure 7). Therefore DFT-hybrid geometries were employed for the excited-state calculations.

In Figure 8 excitation energies at TDHF, TDDFT-hybrid, and pure TDDFT are presented for open-shell cations. The three methods find two excited states arising from the mixing of HOMO−SOMO and SOMO−LUMO transitions, as expected. However, TDHF predicts the correct order of oscillator strengths only for butadiene. For polyenes longer than butadiene, the lower energy excited state is intense and the higher one is weak. At the same time the energies of the two excited states are too close, underestimating the high-energy transition and overestimating the low-energy transition. The values deteriorate completely at longer chain lengths. In contrast, TDDFT, hybrid, and pure DFT, correctly predict lower oscillator strengths for the low-energy feature for short to medium sized polyenes up to $C_{28}H_{30}^+$. At $C_{32}H_{34}^+$ TDDFT-hybrid switches oscillator strengths of the two excited states. The dependence of oscillator strengths on chain lengths with hybrid and pure DFT functionals for even- and odd-numbered cations is shown in Figures 9 and 10.

It turns out that the crossover occurs with both TDHF and TDDFT-hybrid when the expectation value of the spin operator exceeds a value of 1. The value for a pure doublet state is 0.75. Thus, the relative oscillator strengths of the two transitions exchange when the wavefunction acquires increasing triplet character. As mentioned above, the two low-energy excitations of polyene cations arise from coupling

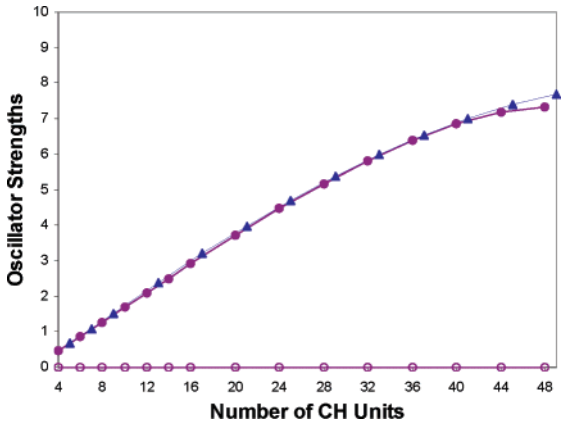


Figure 10. Pure TDDFT oscillator strengths for closed-shell cations $C_5H_7^+-C_{49}H_{51}^+$ (triangles) and for open-shell cations $C_4H_8^+-C_{48}H_{50}^+$ (circles).

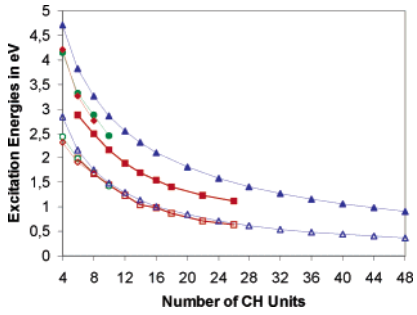


Figure 11. Excitation energies for $C_4H_8^+-C_{48}H_{50}^+$, dipole-forbidden states (open symbols), and dipole-allowed states (filled symbols). Experiment parent polyene cations (diamonds), experiment tert-butyl-capped polyene cations (squares), MRMP (circles), and pure TDDFT (triangles).

between the transitions of α- and β-electrons with the opposite sign (dipole-allowed) and with the same sign (dipole-forbidden). Since a triplet has two electrons with the same spin, the mixing of α- and β-excitations with same signs is dipole-allowed and mixing with opposite signs is dipole-forbidden. The spin contaminated wavefunction can therefore not be applied for calculating excitation energies and oscillator strengths. Pure DFT has no problem with spin contamination. The spin expectation value increases very slowly and reaches only 0.80 for $C_{56}H_{58}^+$. No crossover between oscillator strengths occurs with pure TDDFT (Figure 10), confirming that the crossover is an artifact caused by spin-contamination.

In Figure 11 pure TDDFT excitation energies are compared with experiment and with MRMP results. Experimental data are measured by Bally et al.[10,84] There are two sets of experimental data, parent polyene cations and tert-butyl capped systems. The weak absorption is very similar for both systems, but the strong peak occurs at lower energy for tert-butyl capped polyene cations. MRMP data are those of Kawashima et al.[15] MRMP excitation energies reproduce experimental data for parent polyene cations very well. However, it should be kept in mind that experiments are done in matrix and that theoretical values are obtained in the gas phase. For neutral polyenes the solvent effect is 0.3−0.4 eV.[89] Therefore one might expect accurate theoretical values to be a little higher than experimental ones. Interestingly,

Table 3. TDDFT Excitation Energies in eV and Oscillator Strengths for Polyene Cations[a]

	$C_4H_6^+$	$C_6H_8^+$	$C_8H_{10}^+$	$C_{10}H_{12}^+$
E(1)	2.84 (2.43)	2.17 (1.98)	1.76 (1.69)	1.48 (1.43)
osc. strength	0.010 (0.011)	0.009 (0.015)	0.007 (0.022)	0.005 (0.022)
E(2)	4.71 (4.16)	3.83 (3.32)	3.26 (2.88)	2.86 (2.46)
osc. strength	0.475 (0.626)	0.858 (0.998)	1.260 (1.380)	1.679 (1.706)

[a] MRMP values according to ref 15 are given in brackets.

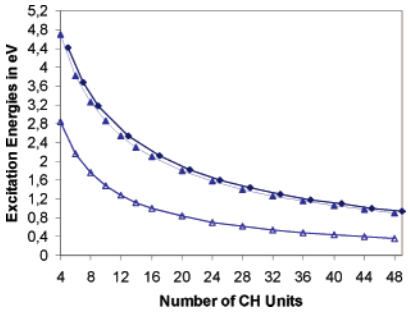


Figure 12. Comparison of excitation energies of open-shell (triangles) and closed shell (diamonds) cations with pure TDDFT.

the solvent effects for the butadiene and the hexatriene cations are reported to have opposite signs, increasing the excitation energy of butadiene but lowering the excitation energy of hexatriene by 0.2 eV.[84]

Pure TDDFT excitation energies for the low-energy transition agree very well with MRMP data (Table 3) and with experiment. All data sets fall on top of each other. For the high-energy feature there are larger differences. TDDFT predicts higher excitation energies than MRMP, but the difference is acceptable ~0.4−0.5 eV. The chain length dependence compared to the end-capped systems is also reasonable, although there is a tendency of TDDFT excitation energies of falling off too fast. Thus TDDFT, which is the only theoretical method that can be applied to large open-shell systems, produces reasonable results compared to experiment and MRMP. (A note on the side: Since relative signs of α- and β-orbitals are random in unrestricted calculations, relative signs of α- and β-coefficients contributing to a certain excited state are also random. Thus the dipole allowed state might appear to be generated from the positive instead of the negative sign combination of α- and β-excitations with TDDFT. Plotting of the orbitals reveals that the α- and β-orbitals are mirror images of each other in such cases.)

In Figure 12 pure TDDFT excitation energies for closed and open-shell systems are plotted vs increasing number of CH units. In Figures 13 and 14 spectra are simulated for selected odd- and even-numbered polyene cations of comparable size. It is visible that excitation energies of the intense feature in open-shell and closed shell species are very similar when calculated with pure TDDFT. For instance, the excitation energies of the strong absorptions are 0.98 and 1.01 eV for $C_{44}H_{46}^+$ and $C_{45}H_{47}^+$, respectively. The same agreement was found for oscillator strength as shown in Figure 10. At all chain lengths the low-energy absorption has very small oscillator strength and is invisible in the spectra. For the longer species $C_{40}H_{42}^+$ and $C_{41}H_{43}^+$, additional features appear at the long wavelengths side of the

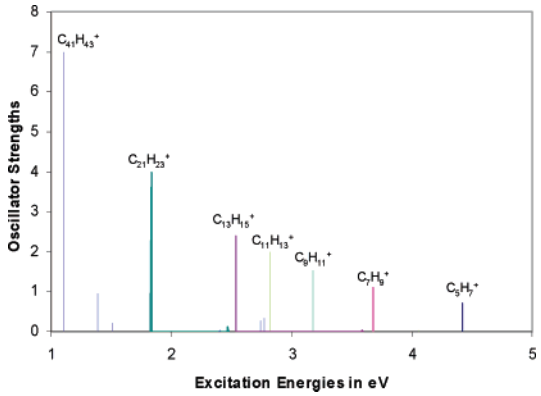


Figure 13. Simulated UV-spectra of odd-numbered closed-shell polyene cations at the pure TDDFT level.

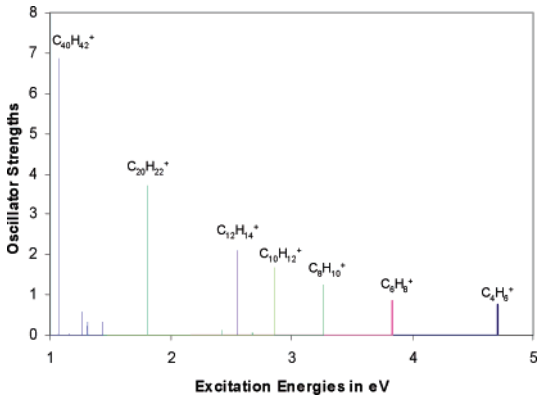


Figure 14. Simulated UV-spectra of even-numbered open-shell polyene cations at the pure TDDFT level.

main peak. That there are more such features with open-shell species might be due to slight differences in α- and β-orbital energy differences in unrestricted calculations. Overall spectra of even- and odd-numbered cations are predicted to differ little, which means that closed-shell odd-numbered cations are indeed very good models for doped PA and can be used to avoid the problems with spin-contamination.

The almost complete cancellation of the oscillator strength of the weak absorption of open-shell cations can be understood by investigating the orbital energy levels for α- and β-electrons. In Figures 15 and 16, DFT-hybrid π-orbital energies are plotted for closed-shell $C_{13}H_{15}^+$ through $C_{101}H_{103}^+$ and open-shell $C_{10}H_{12}^+$ through $C_{48}H_{50}^+$. In Figure 16, β-energy levels are slightly shifted right with respect to α-energy levels. Short open-shell cations have very different energies for α- and β-electrons, β-levels lying lower in energy. As the chain lengths increases, these energy differences decrease as the β-LUMO, the orbital from which the electron has been removed, merges with the α-LUMO. Likewise the α- and β-HOMO orbitals approach each other.

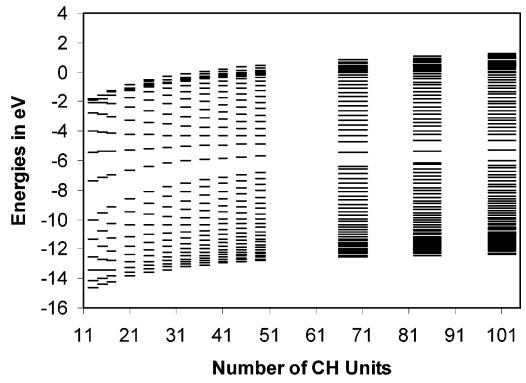


Figure 15. DFT-hybrid π-orbital energies for $C_{13}H_{15}^{+}$–$C_{101}H_{103}^{+}$.

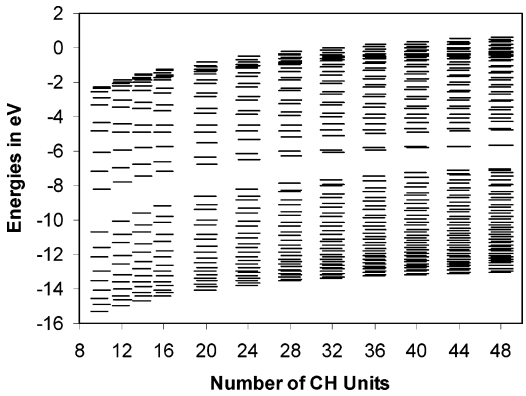


Figure 16. DFT-hybrid π-orbital energies for $C_{10}H_{12}^{+}$–$C_{48}H_{50}^{+}$. Values of β-electrons offset to the right.

Despite the energy differences between α- and β-levels for short cations, the α and β HOMO−LUMO gaps are almost identical at all chain lengths. Thus, like in closed-shell polyene cations, exciting α- and β-electrons require similar energies. For this reason, excitations are produced by mixing both states in almost equal amounts at all chain lengths. The equal mixing leads to the almost complete cancellation of the oscillator strength of the dipole-forbidden state and the resemblance of spectra of closed- and open-shell polyene cations. It will be shown in a forthcoming publication that such a cancellation does not occur with thiophene and pyrrole oligomer cations, which exhibit two absorptions of comparable strengths in the UV.

Comparing Figures 15 and 16 reveals that for long chains, closed- and open-shell polyene cations have very similar energy level patterns with one empty level lying slightly above the middle of the energy gap. This state moves to the middle of the gap in the long chain limit as is visible for $C_{101}H_{103}^{+}$ in Figure 15. For $C_{101}H_{103}^{+}$, the HOMO−LUMO gap is 1.37 eV, and the midgap state lies 0.70 eV above the HOMO. As discussed in the geometry section, the defect size for $C_{101}H_{103}^{+}$ is about 80 CH units in the absence of a counterion. Thus geometric localization of the defect is not responsible for the midgap state.

Higher Lying Excited States. With increasing chain lengths additional ^1B_u and $^1B_2$ states of decreasing energy and of increasing oscillator strength are predicted by TDDFT (compare Figures 13 and 14). There are also low lying excited states of other symmetries, but they have low oscillator strength at all chain lengths. For the longest odd-numbered systems, two additional $^1B_2$ states reach oscillator strengths comparable to that of the 1 $^1B_2$ state. The first five $^1B_2$ excitation energies are summarized for $C_{49}H_{51}^{+}$, $C_{69}H_{71}^{+}$, $C_{85}H_{87}^{+}$, and $C_{101}H_{103}^{+}$ at TDDFT-hybrid and pure TDDFT levels in Table 4.

Table 4. $^1B_2$ Excitation Energies in eV and Oscillator Strengths (in Brackets) for $C_{49}H_{51}^{+}$, $C_{69}H_{71}^{+}$, $C_{85}H_{87}^{+}$, and $C_{101}H_{103}^{+}$ at TDDFT-Hybrid and Pure TDDFT Levels of Theory

		$C_{49}H_{51}^{+}$	$C_{69}H_{71}^{+}$	$C_{85}H_{87}^{+}$	$C_{101}H_{103}^{+}$
1 $^1B_2$	hybrid	1.03 (9.29)	0.79 (11.65)	0.68 (11.94)	0.58 (10.96)
	pure	0.94 (7.70)	0.68 (8.34)	0.55 (7.92)	0.44 (7.79)
2 $^1B_2$	hybrid	1.70 (0.37)	1.26 (1.20)	1.05 (2.60)	0.88 (5.06)
	pure	1.19 (1.60)	0.89 (3.72)	0.75 (5.39)	0.65 (6.62)
3 $^1B_2$	hybrid	2.06 (0.08)	1.68 (0.08)	1.47 (0.01)	1.25 (0.32)
	pure	1.31 (0.39)	1.00 (1.24)	0.86 (2.05)	0.76 (2.80)
4 $^1B_2$	hybrid	2.43 (0.33)	1.84 (0.42)	1.56 (0.53)	1.38 (0.43)
	pure	1.76 (0.00)	1.27 (0.02)	1.03 (0.10)	0.87 (0.57)
5 $^1B_2$	hybrid	2.61 (0.60)	2.04 (0.99)	1.77 (1.35)	1.59 (2.08)
	pure	1.93 (0.02)	1.43 (0.08)	1.18 (0.24)	1.02 (0.48)

Although the 1 $^1B_2$ states of closed-shell cations have almost the same energies with TDDFT-hybrid and pure TDDFT, the energies of higher lying $^1B_2$ states differ with both DFT methods. Pure DFT predicts these excitation energies to be between 0.23−0.75 eV lower in energy and their oscillator strengths to be larger. The differences decrease with increasing chain lengths. The reason for this difference seems to be that energy levels lie closer with pure DFT than with DFT-hybrid. Thus with pure TDDFT, the additional excited states become significant at shorter chain lengths. The only number for comparison, that is available for a higher lying excited state, is the 3 ^2B_g state of decapentaene, obtained with MRMP by Kawashima et al.,[15] which lies at 3.53 eV. For this state both DFT levels seem to work fine. Pure TDDFT predicts 3.52 eV, TDDFT-hybrid 3.73 eV ($S^2 = 0.86$). The pure TDDFT value is closer to the MRMP value, but one number is not sufficient to assess the performance of the two methods with certainty.

The decrease in energy of the higher $^1B_2$ states with increasing chain lengths is consistent with the decrease in the spacing between energy levels (compare Figure 15). As the corresponding excitation energies are coming down, the corresponding states are becoming multiconfigurational. The first $^1B_2$ state is still dominated by the HOMO−LUMO transition, but a second transition between HOMO-2 and LUMO gains importance with increasing chain length. 2 $^1B_2$ is composed of two single electron transitions between HOMO-2 and LUMO and HOMO-1-LUMO+2. 3^1B_2 is composed of two contributions, HOMO-1-LUMO+1 and HOMO−LUMO+2. The energy spacing between the three states is 0.21 and 0.11 eV for $C_{101}H_{103}^{+}$ with pure TDDFT and 0.30 and 0.37 eV with TDDFT-hybrid.

Effect of Counterions. Depending on chain lengths of the polyene, energy levels of counterions may lie within the polyene band gap. For such systems, excitations from polyene to the counterions have very low energies but also low oscillator strengths. The main features in the UV-spectra

Table 5. Pure TDDFT Excitation Energies (eV) of Polyene Cations and Polyene-Cl_3 Complexes Based on Geometries with Increasingly Localized Defects

			B3P86-30% complex	
geometry	BP86	B3P86-30%	no counterion	with Cl_3^-
		$C_{41}H_{43}^+$		
$1\ ^1B_2$	1.08 (8.00)	1.10 (7.00)	1.05 (5.36)	1.14 (4.75)
$2\ ^1B_2$	1.32 (0.28)	1.39 (0.95)	1.39 (1.59)	1.18 (0.62)
$3\ ^1B_2$	1.39 (0.12)	1.51 (0.21)	1.54 (0.47)	1.50 (1.55)
		$C_{69}H_{71}^+$		
$1\ ^1B_2$		0.68 (8.34)	0.62 (5.92)	0.75 (4.05)
$2\ ^1B_2$		0.89 (3.72)	0.91 (3.67)	0.77 (1.92)
$3\ ^1B_2$		1.00 (1.24)	1.04 (1.57)	1.04 (4.73)
$4\ ^1B_2$			1.25 (0.39)	1.06 (0.54)
		$C_{101}H_{103}^+$		
$1\ ^1B_2$		0.44 (7.79)	0.38 (5.92)	0.54 (5.08)
$2\ ^1B_2$		0.65 (6.62)	0.64 (4.90)	0.81 (8.16)
$3\ ^1B_2$		0.76 (2.80)	0.75 (0.78)	0.83 (0.13)
$4\ ^1B_2$		0.87 (0.57)	0.86 (2.51)	0.99 (1.08)
$5\ ^1B_2$		1.02 (0.48)	1.04 (1.08)	1.12 (1.16)

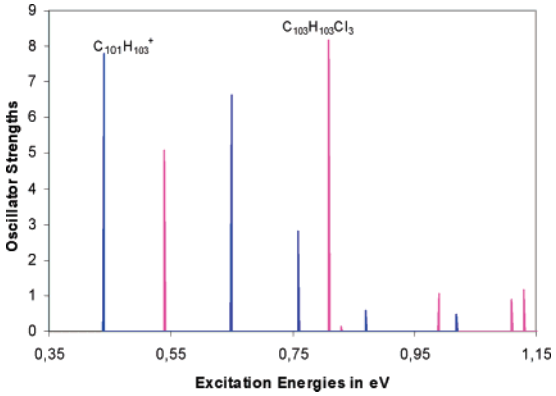


Figure 17. UV-spectra of $C_{101}H_{103}^+$ and of $C_{101}H_{103}-Cl_3$ with pure TDDFT.

are the same interchain excitations as in the absence of counterions in all systems investigated. Thus the effect of the counterions on the spectra is indirect, influencing geometric and electronic structures and therefore transition energies and oscillator strengths of the polyenes.

In Table 5 excitation energies of $C_{41}H_{43}-Cl_3$, $C_{69}H_{71}-Cl_3$, and $C_{101}H_{103}-Cl_3$ are compared with those of the corresponding naked cations. In Figure 17 spectra are simulated for $C_{101}H_{103}-Cl_3$ and $C_{101}H_{103}^+$. In all cases there is a blue shift due to the presence of a counterion. The energies of the first allowed transitions in the presence of the counterion differ by up to 0.10 eV from those of the naked cations. The differences are a little larger for the higher energy transitions. The main difference between complexes and cations is that the oscillator strengths of the first transitions decrease and those of the higher energy transitions increase in the presence of counterions. Thus the maximum absorption shifts to shorter wavelengths.

Counterions have a localizing effect on geometry and electronic structure. To separate the two effects, the counterions were removed from the complexes, and spectra were recalculated for the cations employing the geometries of the complexes. For $C_{41}H_{43}^+$ the delocalized pure-DFT geometry was used in addition. The data are included in Table 5. Using the delocalized pure-DFT geometry does not change the spectrum compared to the one based on the DFT-hybrid structure. Using the geometries of the complexes leads to a slight red-shift of the first peak and an increase in oscillator strength of the higher lying peaks. In Figure 18 spectra of $C_{101}H_{103}^+$ with different geometries are compared. The difference between the spectra with localized and delocalized defects is not dramatic. The first allowed peak is red-shifted by 0.06 eV, and the oscillator strengths of peaks four increase while those of peaks 1, 2, and 3 decrease. This result suggests that electron phonon coupling is not crucial to explain the spectral changes during doping. The bigger change occurs in the electronic structure and is caused by the presence of the counterion as can be seen in Figure 17.

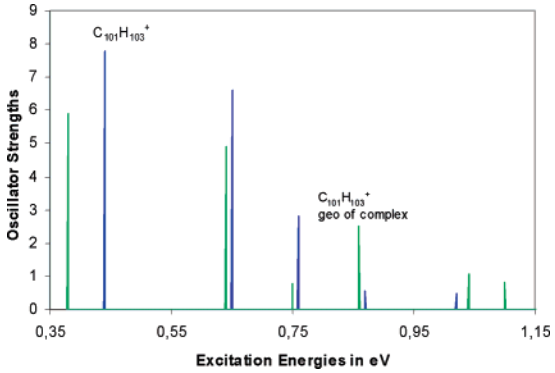


Figure 18. UV-spectra of $C_{101}H_{103}^+$ at its optimized geometry and at the geometry of $C_{101}H_{103}-Cl_3$ after removing the counterion with pure TDDFT.

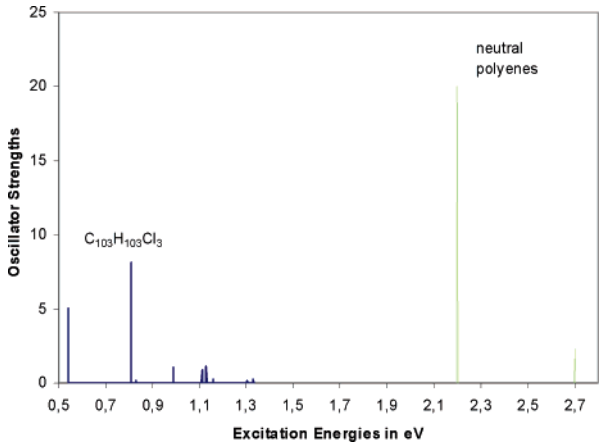


Figure 19. Comparison of the spectrum predicted with pure-TDDFT for $C_{101}H_{103}-Cl_3$ and that for neutral polyenes obtained with TDHF.

The last entry in Table 5 for $C_{101}H_{103}-Cl_3$ should resemble the excitation energies of a lightly doped isolated PA chain quite well. Therefore the spectrum of $C_{101}H_{103}-Cl_3$ is compared to the TDHF spectrum typical for neutral polyenes in Figure 19. TDHF λ_{max} values of neutral polyenes converge to ~2.2 eV. The λ_{max} values for example are 2.24 eV for

$C_{88}H_{90}$ and 2.17 for $C_{162}H_{164}$. Oscillator strengths for long chain polyenes are huge with values of up to 40 for $C_{162}H_{164}$. For cations oscillator strengths go through a maximum with increasing chain lengths as oscillator strengths shift from lower to higher energy transitions. Therefore the predicted spectral change upon doping (Figure 19) agrees well with experiment which shows a decrease of the intense 1.8 eV peak of neutral PA and the development of a much weaker and rather broad feature peaking at around 0.7 eV.[2]

Conclusions

In this investigation the theoretical level necessary to obtain excitation energies of closed- and open-shell polyene cations with up to 101 CH units was established, and spectral features of polyene monocations were examined in the absence and in the presence of a Cl_3^- counterion.

- Comparison of CASSCF, MP2, and DFT methods shows that geometries of closed-shell polyene cations can be obtained with DFT-hybrid functionals (30% of HF exchange) with the same accuracy as with MP2. This holds with respect to bond lengths alternation and defect size. For even-numbered open-shell cations, spin-contamination increases with increasing chain lengths with the DFT-hybrid functional and leads to deterioration of the results. Pure DFT is consistent for closed- and open-shell systems, but BLA is too small since double lengths are too large.
- Potential energy surfaces of polyene cations are very flat. The energy differences per repeat unit for structures with localized and delocalized defects are so small that the MP2 level of theory is probably not sufficient to determine the accurate defect size. Defect sizes are expected therefore to depend strongly on environmental effects, such as counterions, medium, and crystal packing.
- For excited-state calculations of closed-shell cations, TDDFT gives reliable values that do not deteriorate at long chain lengths like for neutral species. Pure DFT and DFT-hybrid lead to very similar excitation energies for the first allowed excited state. Higher lying excited states are obtained at lower energy with pure DFT than with DFT hybrid.
- Because of spin-contamination, hybrid functionals lead to qualitatively wrong oscillator strengths for open-shell polyene cations longer than 28 CH units. Pure DFT gives similar results for open- and closed-shell cations at all chain lengths and compares well with experiment and MRMP for butadiene through decapentaene cations.
- The lowest allowed excited state has the same energy and the same oscillator strength for odd- and even-numbered polyene cations. The weak feature that occurs only for open-shell systems has low oscillator strength at all chain lengths. Thus odd-numbered polyene cations are good models for PA.
- With increasing chain lengths additional allowed states with high oscillator strengths and only slightly higher energies are predicted. This might lead to a broad absorption or several very closely spaced features in the UV spectrum of lightly doped PA. Conjugation lengths of about 40−50 CH units are required for these high additional transitions to achieve significant oscillator strengths.
- Counterions have a localizing effect and decrease the size of the defect by about 40 CH units in the longest system considered here with 101 CH units. The defect is still predicted to be about 40 CH units wide. There is a midgap state for long polyene cations in the absence of counterions when the defect spreads over 80 CH units. Thus defect localization is not necessary for producing a midgap state.
- The energies of the first five excited states of polyene cations are practically independent of the geometry used. The sub-band transition in polyene cations is obtained in structures with localized and with delocalized defects at the same energy. Thus the sub-band feature in doped PA seems to be due to changes in the electronic rather than in the geometric structure.
- Localization influences oscillator strengths, shifting intensity from the HOMO−LUMO transition to higher lying excited states. This trend is further enhanced when a counterion is included explicitly. Excitations between polyene cation and counterion have low oscillator strengths.

Acknowledgment. This work is supported by TÜBITAK (TBAG-2461) and by Bilkent University. I would like to thank Prof. Stefan Grimme for helpful discussions and Dennis Salzner for writing a program that extracts orbital energies from G03 output files for plotting.

References

(1) Naarmann, C. H.; Theophilou, N. *Synth. Met.* **1987**, *22*, 1.

(2) Fincher, C. R., Jr.; Peebles, D. L.; Heeger, A. J.; Druy, M. A.; Matsamura, Y.; MacDiarmid, A. G.; Shirakawa, H.; Ikeda, S. *Solid State Commun.* **1978**, *27*, 489.

(3) Fincher, C. R., Jr.; Ozaki, M.; Tanaka, M.; Peebles, D.; Lauchlan, L.; Heeger, A. J. *Phys. Rev. B* **1979**, *20*, 1589.

(4) Fincher, C. R., Jr.; Ozaki, M.; Heeger, A. J.; MacDiarmid, A. G. *Phys. Rev. B* **1979**, *19*, 4140.

(5) Patil, A. O.; Heeger, A. J.; Wudl, F. *Chem. Rev.* **1988**, *88*, 183.

(6) Yakushi, K.; Lauchlan, L. J.; Clarke, T. C.; Street, G. B. *J. Chem. Phys.* **1983**, *79*, 4774.

(7) Su, W.-P.; Schrieffer, J. R.; Heeger, A. J. *Phys. Rev. Lett.* **1979**, *42*, 1698.

(8) Su, W. P.; Schrieffer, J. R.; Heeger, A. J. *Phys. Rev. B* **1980**, *22*, 2099.

(9) Su, W.-P. *Solid State Commun.* **1980**, *35*, 899.

(10) Bally, T.; Roth, K.; Tang, W.; Schrock, R. R.; Knoll, K.; Park, L. Y. *J. Am. Chem. Soc.* **1992**, *114*, 2440.

(11) Zinger, B.; Mann, K. R.; Hill, M. G.; Miller, L. L. *Chem. Mater.* **1992**, *4*, 1113.

(12) Hill, M. G.; Penneau, J.-F.; Zinger, B.; Mann, K. R.; Miller, L. L. *Chem. Mater.* **1992**, *4*, 1106.

(13) Bäuerle, P.; Segelbacher, U.; Maier, A.; Mehring, M. *J. Am. Chem. Soc.* **1993**, *115*, 10217.

(14) Furukawa, Y. *J. Phys. Chem.* **1996**, *100*, 15644.

(15) Kawashima, Y.; Nakayama, K.; Nakano, H.; Hirao, H. *Chem. Phys. Lett.* **1997**, *267*, 82.

(16) Ye, A.; Shuai, Z.; Kwon, O.; Brédas, J. L.; Beljonne, D. *J. Chem. Phys.* **2004**, *121*, 5567.

(17) Rubio, M.; Ortí, E.; Pou-Amerigo, R.; Merchán, M. *J. Phys. Chem. A* **2001**, *105*, 9788.

(18) Keszthelyi, T.; Grage, M. M.-L.; Offersgard, J. F.; Wilbrandt, R.; Svendsen, C.; Sonnich Mortensen, O.; Pedersen, J. K.; Jensen, H. J. A. *J. Phys. Chem. A* **2000**, *104*, 2808.

(19) Brédas, J. L.; Chance, R. R.; Silbey, R. *J. Phys. Chem.* **1981**, *85*, 756.

(20) Brédas, J. L.; Chance, R. R.; Silbey, R. *Phys. Rev. B* **1982**, *26*, 5843.

(21) Campbell, D. K.; Bishop, A. R. *Nucl. Phys. B* **1982**, *200*, 297.

(22) Campbell, D. K.; Bishop, A. R.; Fesser, K. *Phys. Rev. B* **1982**, *26*, 6862.

(23) Boudreaux, D. S.; Chance, R. R.; Brédas, J. L.; Silbey, R. *Phys. Rev. B* **1983**, *28*, 6927.

(24) Brédas, J. L.; Scott, J. C.; Yakushi, K.; Street, G. B. *Phys. Rev. B* **1984**, *30*, 1023.

(25) Brédas, J. L.; Thémans, B.; Fripiat, J. G.; André, J.-M.; Chance, R. R. *Phys. Rev. B* **1984**, *29*, 6761.

(26) Brédas, J. L. *J. Mol. Struct. (Theochem*) **1984**, *107*, 169.

(27) Brédas, J. L.; Street, G. B. *Acc. Chem. Res.* **1985**, *18*, 309.

(28) Kivelson, S.; Heeger, A. J. *Phys. Rev. Lett.* **1985**, *55*, 308.

(29) Heeger, A. J.; Kivelson, S.; Schrieffer, J. R.; Su, W.-P. *Rev. Mod. Phys.* **1988**, *60*, 781.

(30) Stafström, S.; Brédas, J. L. *Phys. Rev. B* **1988**, *38*, 4180.

(31) Villar, H. O.; Dupuis, M.; Clementi, E. *Phys. Rev. B* **1988**, *37*, 2520.

(32) Ehrendorfer, C.; Karpfen, A. *J. Phys. Chem.* **1994**, *98*, 7492.

(33) Ehrendorfer, C.; Karpfen, A. *J. Chem. Phys.* **1995**, *99*, 10196.

(34) Ehrendorfer, C.; Karpfen, A. *J. Phys. Chem.* **1995**, *99*, 5341.

(35) Champagne, B.; Spassova, M. *Phys. Chem. Chem. Phys.* **2004**, *6*, 3167.

(36) Tol, A. J. W. *Chem. Phys.* **1996**, *208*, 73.

(37) Irle, S.; Lischka, H. *J. Chem. Phys.* **1997**, *107*, 3021.

(38) Shimoi, Y.; Abe, S. *Phys. Rev. B* **1994**, *50*, 14781.

(39) Brocks, G. *J. Chem. Phys.* **2000**, *112*, 5353.

(40) Scherlis, D. A.; Marzari, N. *J. Phys. Chem. B* **2004**, *108*, 17791.

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

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

(43) Salzner, U. *Curr. Org. Chem.* **2004**, *8*, 569.

(44) Salzner, U. Conjugated Organic Polymers: From Bulk to Molecular Wire. In *Handbook of Theoretical and Computational Nanotechnology*; Rieth, M., Schommers, W., Eds.; ASP: 2006; Vol. 8, Chapter 5, p 203.

(45) Hsu, C.-P.; Hirata, S.; Head-Gordon, M. *J. Phys. Chem. A* **2001**, *105*, 451.

(46) Cai, Z.-L.; Sendt, K.; Reimers, J. R. *J. Chem. Phys.* **2002**, *117*, 5543.

(47) Grimm, S.; Nonnenberg, C.; Frank, I. *J. Chem. Phys.* **2003**, *119*, 11574.

(48) Cundari, T. R.; Stevens, W. J. *J. Chem. Phys.* **1993**, *98*, 5555.

(49) Stevens, W.; Basch, H.; Krauss, J. *J. Chem. Phys.* **1984**, *81*, 6026.

(50) Stevens, W. J.; Krauss, M.; Basch, H.; Jasien, P. G. *Can. J. Chem.* **1992**, *70*, 612.

(51) Salzner, U.; Karalti, O.; Durdagi, S. *J. Mol. Model.* **2006**, *12*, 687.

(52) Salzner, U.; Lagowski, J. B.; Pickup, P. G.; Poirier, R. A. *J. Comput. Chem.* **1997**, *18*, 1943.

(53) Salzner, U.; Lagowski, J. B.; Pickup, P. G.; Poirier, R. A. *J. Phys. Chem.* **1998**, *102*, 2572.

(54) Salzner, U.; Lagowski, J. B.; Poirier, R. A.; Pickup, P. G. *Synth. Met.* **1998**, *96*, 177.

(55) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

(56) Hegarty, D.; Robb, M. A. *Mol. Phys.* **1979**, *38*, 1975.

(57) Eade, R. H. E.; Robb, M. A. *Chem. Phys. Lett.* **1981**, *83*, 362.

(58) Schlegel, H. B.; Robb, M. A. *Chem. Phys. Lett.* **1982**, *93*, 43.

(59) Bernardi, F.; Bottini, A.; McDougall, J. J. W.; Robb, M. A.; Schlegel, H. B. *Far. Symp. Chem. Soc.* **1984**, *19*, 137.

(60) Frisch, M. J.; Ragazos, I. N.; Robb, M. A.; Schlegel, H. B. *Chem. Phys. Lett.* **1992**, *189*, 524.

(61) Yamamoto, N.; Vreven, T.; Robb, M. A.; Frisch, M. J.; Schlegel, H. B. *Chem. Phys. Lett.* **1996**, *250*, 373.

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

(63) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.

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

(65) Bauernschmitt, R.; Ahlrichs, R. *Chem. Phys. Lett.* **1996**, *256*, 454.

(66) Casida, M. E.; Jamorski, C.; Casida, K. C.; Salahub, D. R. *J. Chem. Phys.* **1998**, *108*, 4439.

(67) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*.

(68) Choi, C. H.; Kertesz, M.; Karpfen, A. *J. Chem. Phys.* **1997**, *107*, 6712.

(69) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; 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.; Gonzales, C.; Pople, J. A. *Gaussian 03; Revision C. 02 ed.*; Gaussian, Inc.: Pittsburgh, 2003.

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

(71) Cramer, C. J. *Essentials of Computational Chemistry, Theories and Models*; John Wiley: Chichester, 2004.

(72) Ciofini, I.; Adamo, C.; Chermette, H. *J. Chem. Phys.* **2005**, *123*, 121102.

(73) Jacquemin, D.; Femenias, A.; Chermette, H.; André, J.-M.; Perpète, E. A. *J. Phys. Chem. A* **2005**, *109*, 5734.

(74) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C. *Chem. Phys. Lett.* **2005**, *405*, 376.

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

(76) Fincher, C. R.; Chen, C.-E.; Heeger, A. J.; MacDiarmid, A. G.; Hastings, J. B. *Phys. Rev. Lett.* **1982**, *48*, 100.

(77) Yannoni, C. S.; Clarke, T. C. *Phys. Rev. Lett.* **1983**, *51*, 1191.

(78) Ferraro, J. R.; Hill, S.; Stout, P.; Furlani, A.; Polzonetti, G.; Russo, M. V. *Appl. Spectosc.* **1991**, *45*, 932.

(79) Reed, E. A.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899.

(80) Bally, T.; Hrovat, D. A.; Thatcher Borden, W. *Phys. Chem. Chem. Phys.* **2000**, *2*, 3363.

(81) Serrano-Andrés, L.; Merchán, M.; Nebot-Gil, I.; Lindh, R.; Roos, B. O. *J. Chem. Phys.* **1993**, *98*, 3151.

(82) Serrano-Andrés, L.; Lindh, R.; Roos, B. O.; Merchán, M. *J. Phys. Chem.* **1993**, *97*, 9360.

(83) Nakayama, K.; Nakano, H.; Hirao, K. *Int. J. Quantum. Chem.* **1998**, *66*, 157.

(84) Bally, T.; Nitsche, S.; Roth, K.; Haselbach, E. *J. Am. Chem. Soc.* **1984**, *106*, 3927.

(85) Cronstrand, P.; Christiansen, O.; Norman, P.; Ågren, H. *Phys. Chem. Chem. Phys.* **2001**, *3*, 2567.

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

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

(88) Parac, M.; Grimme, S. *Chem. Phys.* **2003**, *292*, 11.

(89) D'Amico, K. L.; Manos, C.; Christensen, R. L. *J. Am. Chem. Soc.* **1980**, *102*, 1777.

CT600223F

JCTC Journal of Chemical Theory and Computation

A Theoretical Databank of Transferable Aspherical Atoms and Its Application to Electrostatic Interaction Energy Calculations of Macromolecules

Paulina M. Dominiak,* Anatoliy Volkov,* Xue Li, Marc Messerschmidt, and Philip Coppens*

Department of Chemistry, State University of New York at Buffalo, New York 14260

Received June 14, 2006

Abstract: A comprehensive version of the theoretical databank of transferable aspherical pseudoatoms is described, and its first application to protein−ligand interaction energies is discussed. The databank contains all atom types present in natural amino acid residues and other biologically relevant molecules. Each atom type results from averaging over a family of chemically unique pseudoatoms, taking into account both first and second neighbors. The spawning procedure is used to ensure that close transferability is obeyed. The databank is applied to the syntenin PDZ2 domain complexed with four-residue peptides and to the PDZ2 dimer. Analysis of the electrostatic interactions energies calculated by the exact-potential/multipole-moment-databank method stresses the importance of the P_0 and P_{-2} residues of the peptide in establishing the interaction, whereas the P_{-1} residue is shown to play a much smaller role. Unexpectedly, the charged P_{-3} residue contributes significantly to the interaction. The class I and II peptides are bound with the same strength by the syntenin PDZ2 domain, though the electrostatic interaction energy of the P_{-2} residue is smaller for class I peptides. There is no difference between the interaction energies of the peptides with PDZ2 domains from single-domain protein fragments and those from PDZ1−PDZ2 tandems.

Introduction

In a preceding paper, we have described a databank of transferable aspherical atoms to be used in the evaluation of electrostatic interactions between large macromolecules and in the refinement of high-resolution structures of molecules such as proteins, polysaccharides, and nucleic acids.[1] The databank was shown to give an excellent reproduction of the electron density in a number of amino acids when compared with those calculated with conventional ab initio methods at the B3LYP/6-31G** level, while requiring only a small fraction of the computational time. It was shown that the databank allows calculation of the electrostatic interaction energies of amino acid dimers with an accuracy of ∼3 kJ per mole.[2] Its merits relative to the experimental databank pioneered by Pichon-Pesme et al.[3] have been the subject of a previous discussion.[4,5] In a more recent article, the databank has been used to evaluate the electrostatic component of the interaction between fragments of the antibiotic vancomycin and substrate molecules, with the nonelectrostatic contribution evaluated with perturbation-theory-based pairwise atom−atom potentials.[6]

The databank may be compared with the invariom approach, pursued by Dittrich and co-workers, in which aspherical atoms are defined by one atom in an as close as possible bonding environment.[7−9] In our procedure, a large number of molecules is analyzed and chemical similar atoms are grouped into families with closely related multipole population and expansion−contraction parameters. Standard deviations are calculated from the spread in the parameters within one family. A new family is spawned when significant

* Corresponding authors. Fax: (716) 645-6948 (P.C.). E-mail: pd24@buffalo.edu (P.M.D.), volkov@buffalo.edu (A.V.), coppens@buffalo.edu (P.C.).

10.1021/ct6001994 CCC: $37.00
Published on Web 11/10/2006

outliers occur. This procedure ensures that close transferability is obeyed and implicitly accounts for differences in bond order in the atom's coordination sphere.

We note that reconstructing the electron density from transferable aspherical molecular fragments stored in a databank has been pursued by other research groups using quantum-chemical partitioning quite different from that employed here. The fuzzy density fragmentation method, for example, has been introduced by Mezey et al.[10,11]

We describe here the extension of the databank by systematic application of the spawning procedure. The new version includes all atom types encountered in amino acid residues. In a first application to a large protein molecule, the databank is applied to the interactions between the PDZ2 domain of the scaffolding protein syntenin and peptides. The results shed light on the nature of the interaction between the protein and the various residues, which has been the subject of recent discussions.[12,13]

Methodology

Theoretical Calculations. Theoretical calculations were performed as described by Volkov et al.[1] To construct the databank, single-point calculations on selected small molecules (Table S1, Supporting Information) were performed with the Gaussian 98 and Gaussian 03 programs[14] using density functional theory (DFT) with a standard split-valence double-exponential 6-31G** basis set including polarization functions.[15] The DFT calculations were based on Becke's three-parameter hybrid method[16] combined with the nonlocal correlation functional of Lee, Yang, and Parr (B3LYP).[17] Experimental molecular geometries as recorded in the Cambridge Structural Database were used.[18] Hydrogen positions were obtained by extending X−H distances to their standard neutron diffraction values.[19] Complex static valence-only structure factors in the range $0 < \sin\theta/\lambda < 1.1$ Å^{-1} were obtained by the analytic Fourier transform of the molecular charge densities for reciprocal lattice points corresponding to a pseudo-cubic cell with 30 Å edges.[20]

Aspherical Atom Refinement of the Theoretical Structure Factors. The data were fitted with the Hansen−Coppens pseudoatom formalism,[21] using version 4.12 of the XD program suite.[22] Both radial screening factors (κ and κ') were refined independently for each atom, with the exception of the chemically equivalent hydrogen atoms which shared the same κ and κ' parameters. The multipolar expansion was truncated at the hexadecapolar level ($l_{max} = 4$) for the non-hydrogen atoms and at the quadrupolar level ($l_{max} = 2$) for hydrogen atoms, for which only bond-directed functions of $l,m = 1,0$ and $2,0$ were refined. To reduce the number of least-squares variables, local-symmetry constraints were imposed for some atoms. A molecular electroneutrality constraint was applied in all refinements. The phase ambiguity problem was solved by keeping the phases fixed at theoretically calculated values.

Atomic scattering factors were based on the atomic wave functions of Clementi and Roetti[23] and were, in the course of the refinement, modified by the κ parameter. For the deformation functions single-ζ exponents corresponding to weighted averages over the s- and p-shell values given by Clementi and Raimondi[24] were used, with powers of r (n_l) given in Table 1. They were multiplied by the refinable κ' parameters.

Table 1. Slater Radial Function Coefficients, n_l and ζ, Applied in the Refinements of Theoretical Structure Factors

element	n_l	ζ [bohr^{-1}]
C	2, 2, 3, 4	3.176
N	2, 2, 3, 4	3.839
O	2, 2, 3, 4	4.466
S	2, 4, 6, 8[a]	3.851
Cl	4, 4, 4, 4	4.259
H	1, 2	2.000

[a] As suggested by Dominiak and Coppens.[25]

The variables—P_{val}, the valence electrons population (in the model used equivalent to the spherical monopole population); P_{lm}, populations of the deformation functions; κ and κ', the expansion−contraction parameters—were averaged over each atom type and constitute the prime information stored in the databank.

Local Coordinate System Assignment. Averaging of the multipole populations of equivalent pseudoatoms requires that the spherical harmonic functions centered at these atoms be expressed in a common local frame. The program LSDB[1] includes a fully automatic definition of unique local coordinate systems based on the coordination environment of each atom. It is an essential tool in the charge density analysis of any large molecule, for which the manual assignment of coordinate systems becomes prohibitively cumbersome, and in the building of the pseudoatom databank based on a large number of small molecules.

Consistency in selecting the local coordinate system is crucial for defining the atom types and subsequent averaging. In the first step, the local coordinate system is oriented such as to allow local symmetry constraints. Whenever the local symmetry allows more than one possibility of selecting a particular local coordinate axis, the following procedure is used to choose the neighboring atom 1, which defines axis 1:[26]

(1) Atoms with the highest atomic number are selected, and if there is only one such atom, it is chosen as atom 1.

(2) If step 1 is inconclusive, atoms with the highest atomic number are ordered according to their hybridization state (sp first, then sp^2 and sp^3) and then by their valency (number of attached atoms; the smallest number first). The first atom from the list is chosen as atom 1.

(3) If the preceding procedure is not conclusive, that is, there is more than one atom at the top of the list with the same element type, hybridization, and valency, atoms are sorted according to their distance to the central atom, and the one with the shorter distance is chosen as atom 1.

After atom 1 is selected, atom 2 is chosen from the remaining neighboring atoms following the rules defined above. In some cases, it is necessary to define a dummy atom to make use of the symmetry, as for example for the central C atom of the −C−CH$_2$−C− group, for which the dummy atom will be placed at the midpoint between the heavier carbon atoms, and axis 1 will be oriented along the local 2-fold symmetry axis. Such use of local symmetry allows

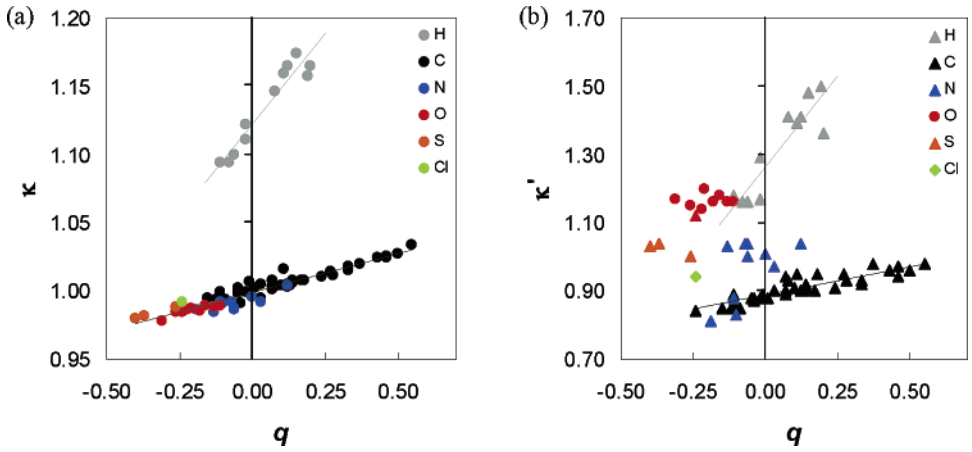


Figure 1. Correlation between the monopole-derived charge, q, and the expansion−contraction parameter (a) κ or (b) κ' for hydrogen (grey), carbon (black), nitrogen (blue), oxygen (red), sulfur (orange), or chlorine (green) atom types.

minimization of the number of multipole parameters. In case of atoms belonging to a planar ring, axis 1 is always oriented toward the center of the ring and mirror plane symmetry is always applied even if higher symmetry is possible. Usually, a right-handed coordinate system is defined, except in the case of chiral atoms, for which both right- and left-handed systems are allowed. Chirality is defined locally, that is, only by the character of the nearest neighbors.

In the initial stages, very low symmetry or no symmetry at all was applied in refinement. In case symmetry was evident after averaging over an atom type, that symmetry was adopted in the final refinement of the molecules. Residual density maps were examined carefully to ensure that the symmetry constraint was justified.

Atom Type Definitions. The key step in the construction of any databank is the choice of atom types, that is, the selection of atoms which are sufficiently similar to be treated identically. A compromise must be struck between achieving the most accurate representation of the charge density of many molecules and minimizing the number of atom types.

In the construction of the databank, a new atom type is introduced whenever the average value of any deformation density parameter over a group of atoms differs by an amount greater than one standard deviation of the sample consisting of the remaining atoms. This approach leads to the following set of general criteria for the definition of an atom type: (1) element type, (2) the number of attached atoms (atom valence, number of the nearest neighbors), (3) nearest neighbor type, which may be affected by the next-nearest neighbors, (4) aromaticity, and (5) local symmetry.

Atoms that are part of planar ring systems are given an aromaticity flag and treated as separate types. Usually, only element type and hybridization are taken into account to characterize any neighboring atom, except for neighboring sp^3 or sp^2-hybridized nitrogen atoms, which are additionally subdivided into $sp^3(4)$, $sp^3(3)$, $sp^2(3)$, and $sp^2(2)$, where the number in parentheses represents the number of neighbors of the nitrogen atom. Hybridization states of neighboring atoms are derived solely from the number of attached atoms and, in the case of nitrogens, planarity of the group. The atom types in the current version of the databank are listed in Table 2, whereas a list of neighbor types used in construction of the databank is given in Table 3.

In some cases, the criteria described above are modified by including the effect of next-nearest neighbors or, on the other hand, by ignoring differences between nearest neighbors. An example of the first situation is oxygen atom types O101−O104[27] (Table 2), while the second occurs for carbon atoms such as C405 for which the type of neighboring C atom does not affect the central atom's population parameters. A simpler definition of an atom type is sometimes used when the number of occurrences of a intended atom type in the molecular sample is too small to get statistically meaningful average values of the deformation-density parameters.

For each atom, the LSDB program analyzes the coordination environment and assigns the appropriate atom type. The corresponding charge-density parameters are then transferred from the databank for the subsequent analysis.

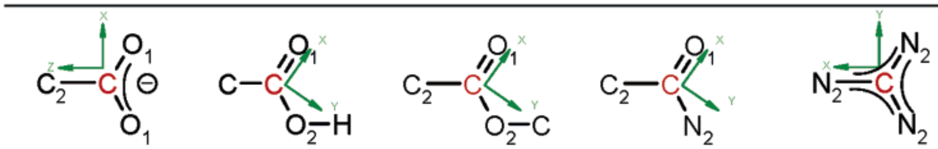
Atom Types in the Databank. At the time of writing, the databank contained 74 atom types (Table 2). (Note added in proof: 104 atom types are incorporated in the version made available on the Web in October 2006.) It includes all atom types encountered in peptides, proteins, and some other biologically interesting molecules.

As noted before,[28,29] the monopole-derived net atomic charges correlate with the expansion−contraction parameters. The κ parameters are strongly related to the charges (Figure 1a) and form two distinctive groups: those describing hydrogen atom types with a fitted linear equation $\kappa = 1.122(3) + 0.27(3)q$, $R = 0.96$, and those for non-hydrogen atoms which are fitted by $\kappa = 0.9987(5) + 0.056(2)q$, $R = 0.96$. A somewhat less clear linear dependence of κ' on q[30] is observed for hydrogen and carbon atom types (Figure 1b). For the other elements, the range of charges is insufficient to examine any correlation, and unlike for κ/q, the points in the graph for N and O do not fall on the line derived for carbon (Figure 1b).

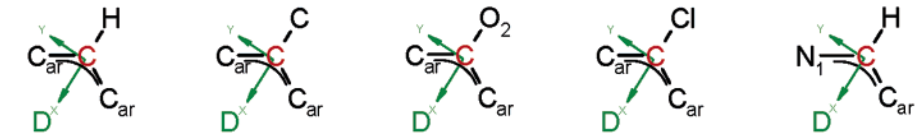
The net atomic charges of the pseudoatoms do not always reflect accepted electronegativity concepts, as they depend crucially on the division of the charge in the bonding regions. For the hydrogen atoms, for example, refinement of both bond-directed dipoles and quadrupoles effectively allows

Table 2. Atom Types in the Databank[a]

Carbon

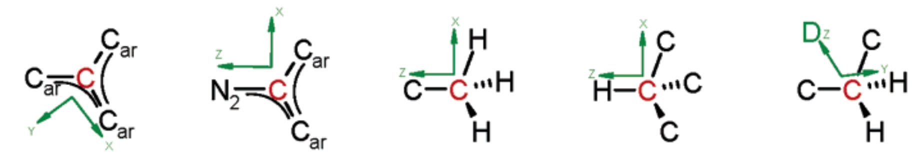


C301 C302 C303 C304 C306

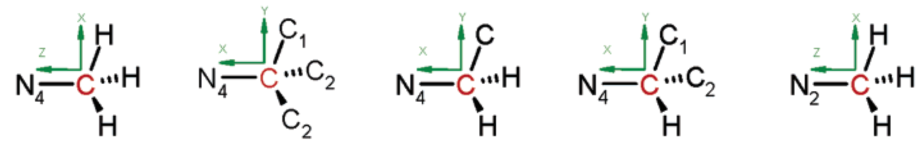


C351 C352 C355 C357 C358

C359 C360 C361 C362 C363

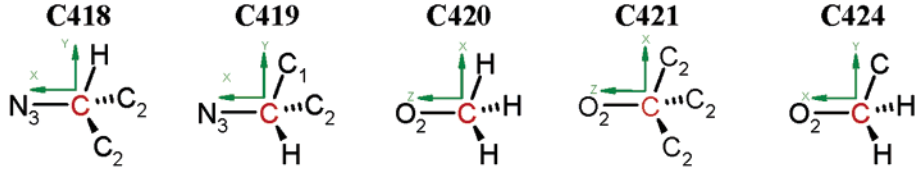


C370 C371 C402 C403 C405

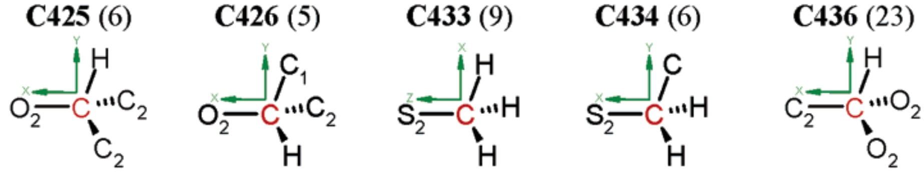


C409 C411 C412 C414 C415

C418 C419 C420 C421 C424

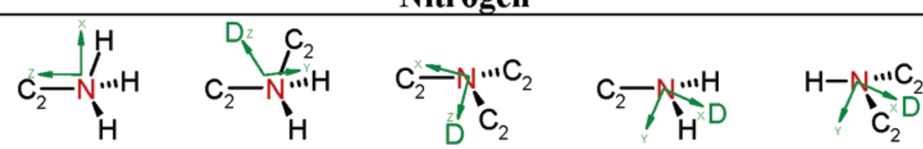


C425 (6) C426 (5) C433 (9) C434 (6) C436 (23)

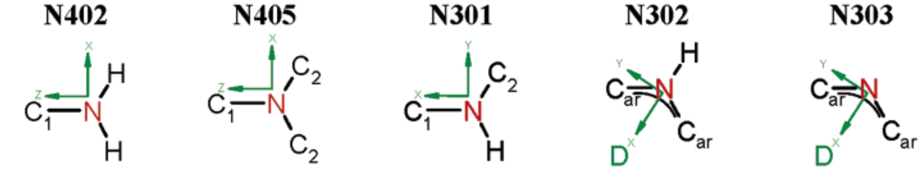


C437 C438 C439 C440 C443

Nitrogen



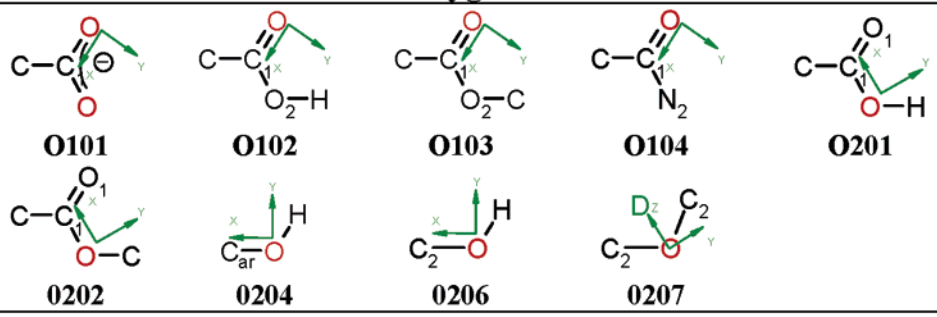
N402 N405 N301 N302 N303



N307 N309 N310 N311 N201

Table 2 (Continued)

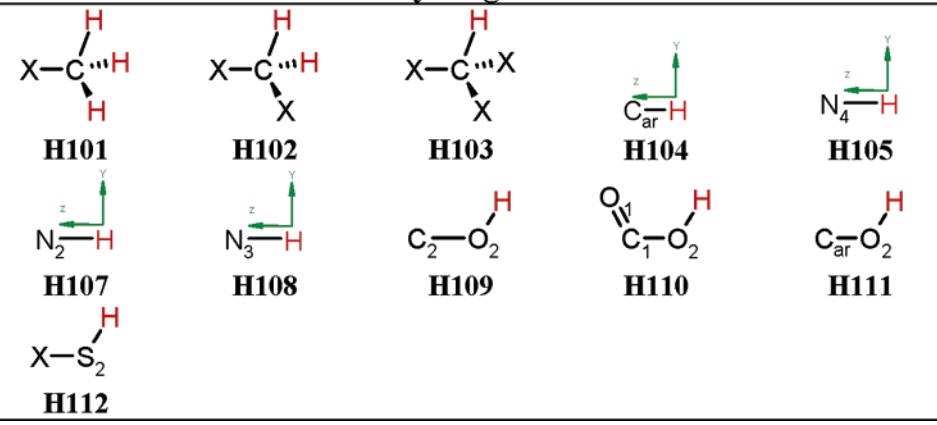
Oxygen



Sulfur

Chlorine

Hydrogen



[a] The symbol D represents dummy atoms, required for coordinate system definition.

Table 3. Labeling of the Nearest and Next-Nearest Neighbor Atom Types[a]

C	any Csp^2 or Csp^3 carbon
C_1	Csp^2 aromatic and nonaromatic
C_2	Csp^3
Car	Csp^2 in planar ring
N_1	$Nsp^2(2)$
N_2	$Nsp^2(3)$
N_3	$Nsp^3(3)$
N_4	$Nsp^3(4)$
O_1	$Osp^2(1)$
O_2	$Osp^3(2)$
S_2	$Ssp^3(2)$
X	any non-hydrogen atom

[a] Numbers in parentheses represent number of bonds to the atom.

assignment of the X−H bond density to the H atom, which becomes slightly negatively charged when bonded to carbon atoms. In general, the electron population on atoms linked to hydrogens decreases as the number of bonded hydrogen atoms increases. The effect is clearly visible for X−Csp^3−R_3 carbon atom types (with X = C, $Nsp^2(3)$, $Nsp^3(3)$, $Nsp^3(4)$, $Osp^3(2)$, or $Ssp^3(2)$ and R = C or H, where the number in parentheses represents the number of neighboring atoms) when atom types with the same X substituent are compared, as shown in Figure 2. The spherical hydrogen-atom population represented by the monopoles correlates with the density of the even and odd multipoles on the H atoms,[31] which give a total density more strongly concentrated along the bond axis and toward the bonded X atom (Figure 3). A similar situation occurs for the monopole populations of X−Csp^3−R_3 carbon-atom types and the octupole and hexadecapole populations (Figure 4).[32]

Carbon Atoms. The databank contains 40 carbon atom types, 23 of which are sp^3 carbons (numbers from C402 to C443). The deformation density parameters for Csp^3 are given in Table S2 (Supporting Information). The corresponding deformation density maps are presented in Figure 5. All sp^3-carbon types are positive, with the monopole-defined charge ranging from +0.03(3) *e* for the C434 type to +0.55(6) *e* for C409.

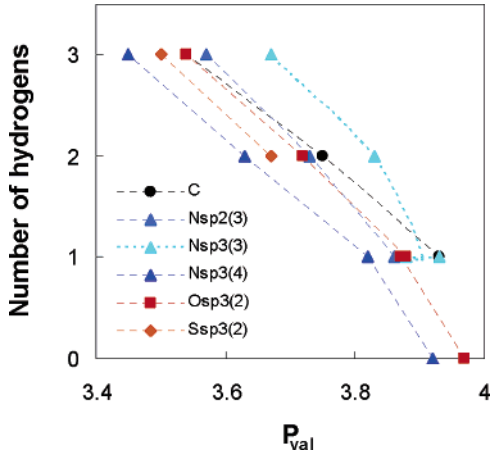


Figure 2. Dependence of the monopole population, P_{val}, of carbon on the number of hydrogen atoms attached in X−Csp^3−R$_3$ atom types, where X stands for C, Nsp^2(3), Nsp^3(3), Nsp^3(4), Osp^3(2), or Ssp^3(2) atom types (see Table 3) and R stands for C or H. The numbers in parentheses represent the number of bonded atoms.

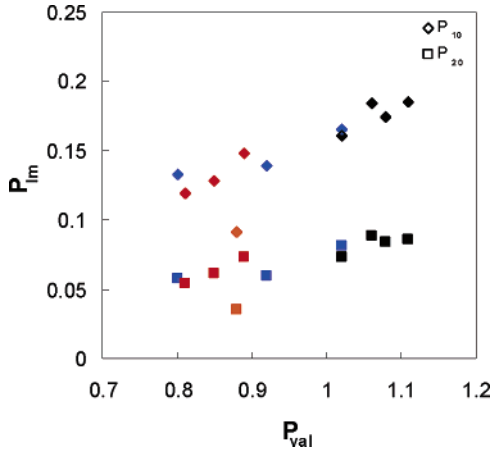


Figure 3. Correlations between the monopole population, P_{val}, and P_{lm} for hydrogen atoms connected to carbon (black), nitrogen (blue), oxygen (red), or sulfur (orange).

For C atoms in X−Csp^3−R$_3$ bonded to the same number of hydrogens, where X stands for C, Nsp^2(3), Nsp^3(3), Nsp^3(4), Osp^3(2), or Ssp^3(2) and R stands for C or H, the tetravalent nitrogens {Nsp^3(4)} induce the largest positive charge, followed by divalent sulfur {Ssp^3(2)}, oxygen {Osp^3(2)}, trivalent sp^2-hybridized nitrogen {Nsp^2(3)}, carbon, and trivalent sp^3-hybridized nitrogen {Nsp^3(3)}, as illustrated in Figure 2. Moreover, the more populated the d_{11+} dipole (oriented along the C−X bond), the more populated the d_{20} and d_{22+} quadrupoles become (Figure 6).[33] The populations of d_{11+}, d_{20}, and d_{22+} decrease in the following order of the substituents X: Nsp^3(4), Osp^3(2), Nsp^2(3), Ssp^3(2), Nsp^3(3), and C.

With the exception of the S atom, the order of the substituent effect agrees well with the concept of electronegativity. According to Huheey's group electronegativity concept,[34] the tetravalent sp^3-hybridized nitrogen is the most electronegative group among X (3.65 for $-NH_3^+$ on the Huheey scale), followed by Osp^3(2) (3.51 for −OH), Nsp^3(3) (2.61 for $-NH_2$), Ssp^3(2) (2.32 for −SH), and

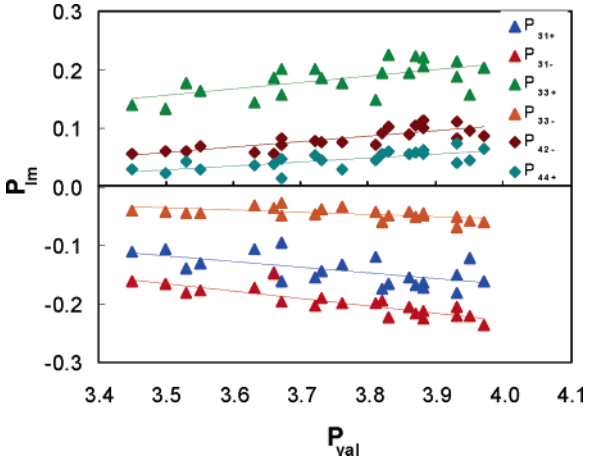


Figure 4. Correlation between the monopole population, P_{val}, and selected P_{lm} in X−Csp^3−R$_3$ carbon atom types, where X stands for C, Nsp^2(3), Nsp^3(3), Nsp^3(4), Osp^3(2), or Ssp^3(2) (as defined in Table 3) and R stands for C or H. The local coordinate system for C402, C403, C405, C409, C415, C421, C433, C434, and C439 is rotated to coincide with that of other X−Csp^3−R$_3$ carbon atom types such as C411.

Csp^3(4) (2.27 for $-CH_3$). It is not surprising, then, that Nsp^3(4) generates the biggest positive charge on carbon (i.e., the smallest values of P_{val} among carbon types with the same number of hydrogens). The number of data points for the sulfur atom is less than those for the other atoms included in the databank, so a discussion of this exception appears premature.

The influences of hydrogen and Csp^3 carbon on the aspherical deformation density of the central carbon atom are very comparable, leading to mirror plane symmetry relating H and Csp^3 with each other in the chiral Csp^3 atoms C414, C420, C426, and C438, as evident from the group-averaged values of the P_{lm} coefficients (Table S2, Supporting Information).

The difference between sp^3 and sp^2 of the nearest neighboring carbon appears to have only a minor influence on the central carbon atom. For example, the C405 atom type was originally divided into two subgroups, with two Csp^3 atoms as first neighbors (*mm*2 local symmetry) and with Csp^2 and Csp^3 atoms as neighbors (*m* symmetry). However, after appropriate coordinate system rotation, the two subtypes became statistically indistinguishable.

There are currently 12 aromatic carbon atom types in the databank. Their deformation density parameters are given in Table S3 (Supporting Information), and deformation density maps are presented in Figure 7. The atoms are close to neutral and vary from +0.11(6) (C363) to −0.15(4) (C360) electron units. In the first 10 atom types, representing atoms participating only in a single ring, the multipoles d_{33+}, d_{20}, and d_{11+} are dominant. They describe respectively the deformation density along the three bonds, the π density above and below the ring, and the difference between the density in the exocyclic and cyclic bonds.

Nonaromatic sp^2 carbon atoms are represented by five different types in the current version of the databank. They are described in Table S4 (Supporting Information); the graphical representation of the deformation densities is shown

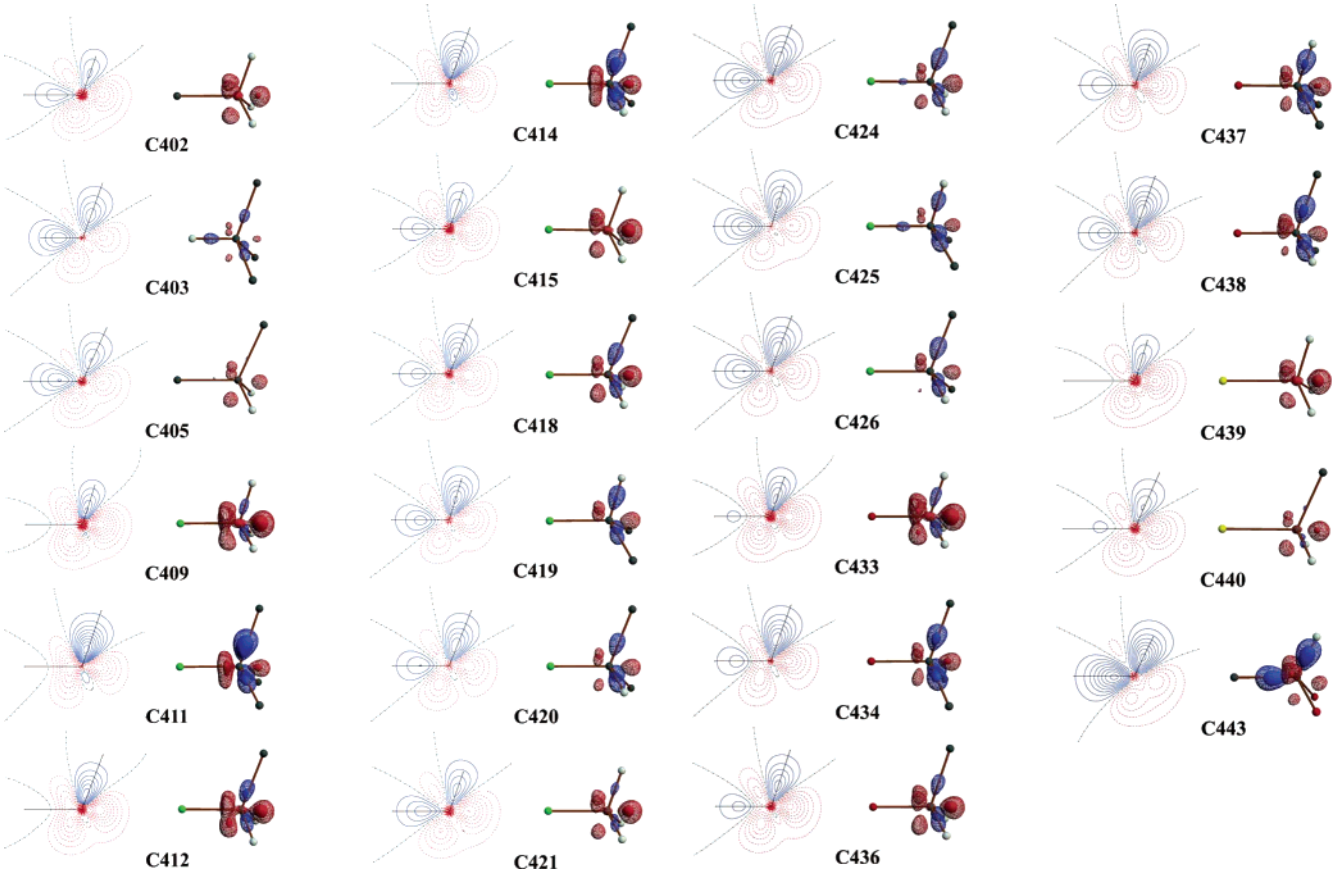


Figure 5. Deformation densities of the sp^3 carbon atom types. Left: In the plane of two neighboring atoms, orientation as in Table 2, contour levels at 0.05 *e*/Å^3. Right: 3D representation, contour levels at ±0.2 and ±0.3 *e*/Å^3. Positive contours, blue; negative contours, red; zero contour, black. Color codes for atoms: C, dark gray; H, white; N, green; O, red; S, yellow.

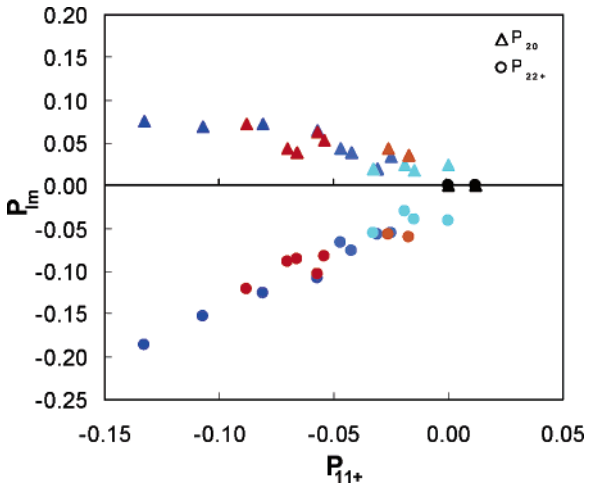


Figure 6. Correlation between P_{11+} and P_{20} (triangles) or P_{22+} (circles) of X−Csp^3−R$_3$ carbon atom types, where X stands for C (black), Nsp^2(3) (blue), Nsp^3(3) (light blue), Nsp^3(4) (dark blue), Osp^3(2) (red), or Ssp^3(2) (orange) atom types (as defined in Table 3) and R stands for C or H. The local coordinate system for C402, C4035, C405, C409, C415, C421, C433, C434, and C439 is rotated to coincide with that of other X−Csp^3−R$_3$ carbon atom types such as C411.

in Figure 8. Four of the types describe carbonyl carbon in, respectively, carboxylate anions, carboxylic acids, esters, and amides. The last type represents carbon atoms in the guanidine group. To distinguish between carboxylic acid and ester type C atoms, the atom type definition includes second neighbors to preserve uniqueness leading to atom types C302 and C303 (Table 2).

Nitrogen Atoms. On the basis of the deformation density, the 10 nitrogen atom types stored in the databank can be divided into two groups: those dominated by lone pair electrons (N301, N302, N303, and N201) and those which do not carry a lone electron pair (N402, N405, N307, N309, N310, and N311; Table S5 of the Supporting Information, Figure 9). The former have large positive deformation densities in the lone pair region and very small positive deformation densities in the bond directions. Additionally, trivalent sp^3-hybridized nitrogens have relatively large electron deficiencies between bond regions and, in the case of aromatic divalent nitrogen (N201), below and above the aromatic ring plane.

Characteristic features for planar trivalent nitrogen atom types (N307, N309, and N310) are charge excess lobes above and below the plane, extending toward the neighboring Csp^2 atom.

Deformation densities for tetravalent nitrogen types are similar to those for tetravalent carbons connected only to C or H atoms. The only difference is that the nitrogen charge is more negative [−0.10(4) and − 0.19(8) *e* for N402 and N405 respectively], causing the electron-charge-accumulating lobes to be more pronounced.

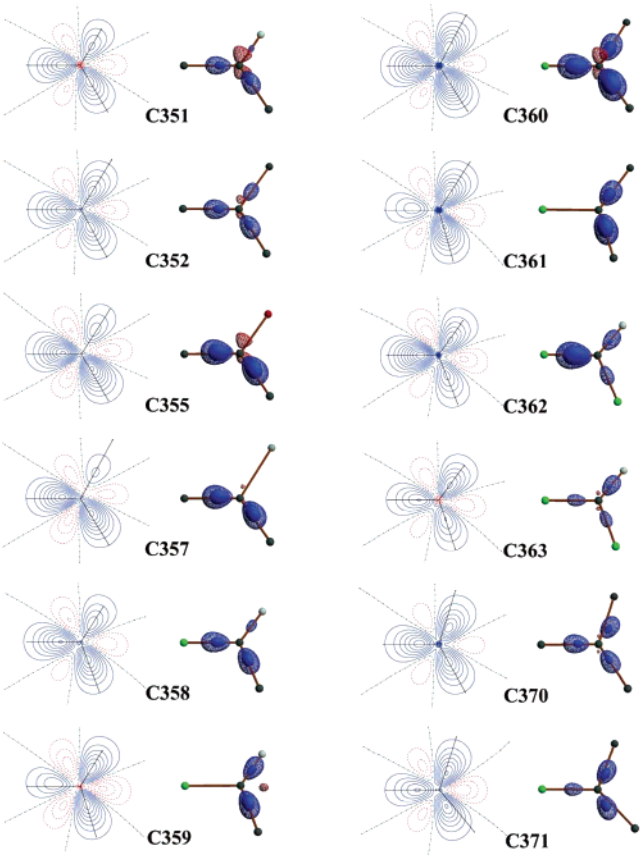


Figure 7. Deformation densities of the aromatic sp^2 carbon atom types. Left: In the plane of two neighboring atoms, orientation as in Table 2, contour levels at 0.05 *e*/Å^3. Right: 3D representation, contour levels at ±0.2 and ±0.3 *e*/Å^3. Positive contours, blue; negative contours, red; zero contour, black. Color codes for atoms: C, dark gray; H, white; N, green; Cl, white.

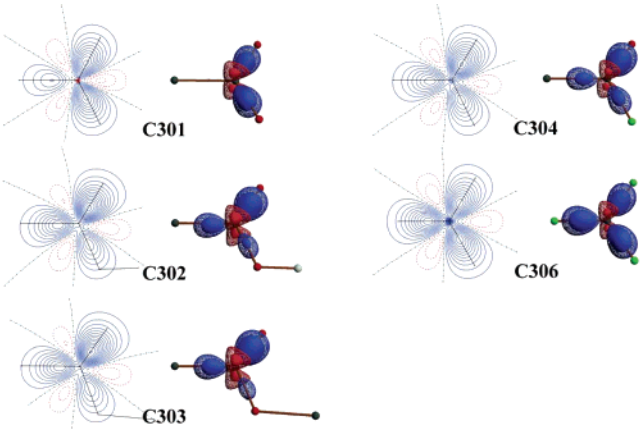


Figure 8. Deformation densities of the nonaromatic sp^2 carbon atom types. Left: In the plane of two neighboring atoms, orientation as in Table 2, contour levels at 0.05 *e*/Å^3. Right: 3D representation, contour levels at ±0.2 and ±0.3 *e*/Å^3. Positive contours, blue; negative contours, red; zero contour, black. Color codes for atoms: C, dark gray; H, white; N, green; O, red.

The aromatic trivalent nitrogen type, N311, differs from aromatic carbons by not having electron deficiency lobes above and below the plane (d_{20} is not populated).

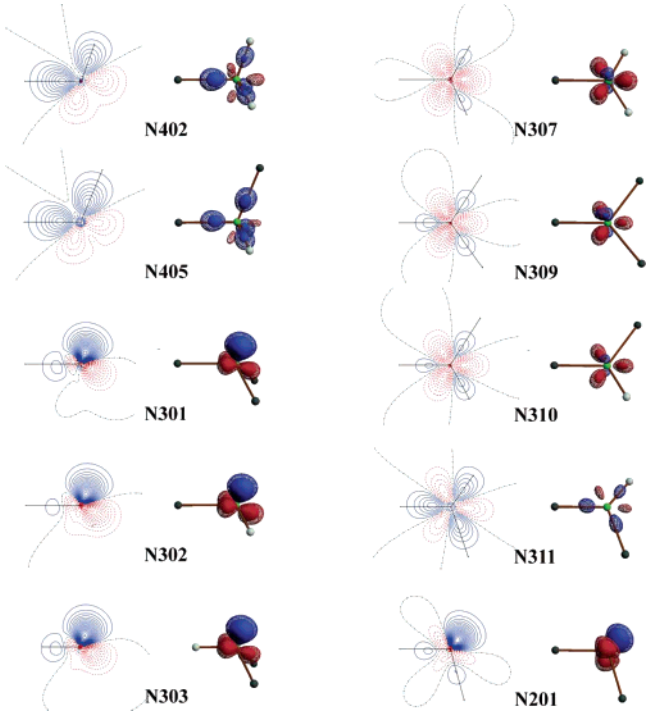


Figure 9. Deformation densities of the nitrogen atom types. Left: In the plane of two neighboring atoms, orientation as in Table 2, contour levels at 0.05 *e*/Å^3. Right: 3D representation, contour levels at ±0.2 and ±0.3 *e*/Å^3. Positive contours, blue; negative contours, red; zero contour, black. Color codes for atoms: C, dark gray; H, white; N, green.

Oxygen Atoms. The deformation densities for all nine oxygen atom types stored in the databank are dominated by the lone pair electron region (Table S6 of the Supporting Information, Figure 10). For monovalent oxygens, two lone pairs are clearly visible; in divalent ones, the lone pairs do not split in two disjunctive regions. The electron deficiency lobe is always oriented perpendicular to the lone pair plane and embraces the bond region. There is no excess density in the bonds originating from the oxygen atoms. All oxygen types bear negative charges ranging from −0.11(3) *e* for carbonyl oxygen in esters to −0.31(3) *e* for carboxylate oxygens.

Sulfur Atoms. There are only three sulfur types in the databank at the moment (Table S7 of the Supporting Information, Figure 11). The deformation of density around them is considerably less pronounced than that for oxygens. Still, the lone pair region dominates, but the concentration of electrons is much smaller, and additionally, there is an electron concentration in the bonding regions. Like oxygens, all sulfur types are negative, in the range from −0.26(2) to −0.40(3).

Chlorine Atoms. Currently, only one type of chlorine atom is stored in the databank, corresponding to Cl connected to aromatic rings. The deformation density has *mm*2 symmetry, see C*l*01 in Table S7 of the Supporting Information and Figure 11. Moreover, with a coordinate system with the *z* axis pointing to the carbon, only one multipole component, the quadrupole 22+ with population −0.013(2), violates cylindrical symmetry and flattens the density toward the

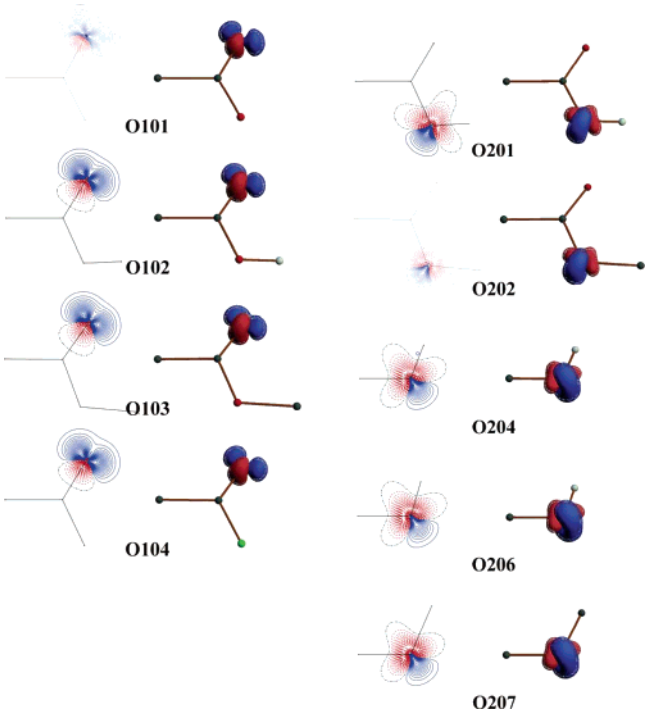


Figure 10. Deformation densities of the oxygen atom types. Left: In the plane of two neighboring atoms, orientation as in Table 2, contour levels at 0.05 *e*/Å^3. Right: 3D representation, contour levels at ±0.2 and ±0.3 *e*/Å^3. Positive contours, blue; negative contours, red; zero contour, black. Color codes for atoms: C, dark gray; H, white; N, green; O, red.

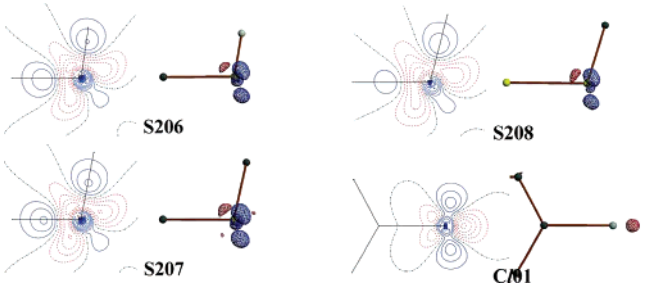


Figure 11. Deformation densities of the sulfur and chlorine atom types. Left: In the plane of two neighboring atoms, orientation as in Table 2, contour levels at 0.05 *e*/Å^3. Right: 3D representation, contour levels at ±0.2 and ±0.3 *e*/Å^3. Positive contours, blue; negative contours, red; zero contour, black. Color codes for atoms: C, dark gray; H, white; S, yellow; Cl, white.

aromatic plane. The population of the monopole is 7.24(3), indicating the Cl to be negative.

Hydrogen Atoms. The databank contains 11 hydrogen types. In the case of hydrogens connected to sp^3-hybridized carbons and oxygens, the definition of atom type has to be extended beyond the first neighbor type defined in Table 3. As a result, hydrogens connected to tetravalent carbons are divided into three groups on the basis of the total number of hydrogens attached to the carbon, whereas hydrogens connected to oxygens are split into those in alcohols (H109), carboxylic acids (H110), and phenols (H111). As discussed above, all deformation density parameters for hydrogens correlate with the monopole populations (Figure 3, Table S8 of the Supporting Information).

Application

In a first macromolecular application, the databank is applied to the 0.7-Å-resolution crystal structure of the second PDZ domain (PDZ2) from the scaffolding protein syntenin[35] and subsequently to all PDZ2−peptide complex structures available in the Protein Data Bank (PDB).[36,10,11]

PDZ domains are the most ubiquitous protein−protein interaction domains.[37,38] They are structurally conserved modules of about 90 amino acids in length and with a distinct fold of six β strands and two α helices.[39,40] Their function is achieved by binding of the C terminus of the target protein as an antiparallel β strand in a highly conserved hydrophobic groove between the β2 sheet and α2 helix. The side chains of the P_0 and P_{-2} ligand residues point into the groove and account for the specificity (P_0 and P_{-n} denote the C-terminal residue of the bound ligand and the *n*th upstream amino acid residue, respectively). The PDZ binding motifs have been divided into three major classes: type I with the consensus sequence −(S/T)XΦ, type II with −ΦXΦ, and type III with −(D/E)XΦ, where X is any amino acid, Φ is a residue with hydrophobic side chains, and S/T is either serine or threonine and D/E is either aspartic or glutamic acid, respectively.[38]

However, it has been pointed out that this simple classification is unable to explain the increasing number of PDZ-mediated interactions that do not conform to this canonical type of recognition.[11] Some PDZ domains are able to bind with internal, rather than C-terminal, peptide sequences. Others recognize noncanonical sequences or more than one class of ligands. The syntenin PDZ2 domain, for example, demonstrates degenerate specificity. It was shown to bind template peptides modeling IL5Rα (SVF, class I), syndecan-4, neurexin, and ephrin B (FYA, YYV, and YKV, respectively, all class II) protein ligands (see Table 4 for nomenclature). Additionally, it has been established that interactions of syntenin with proteins are often mediated cooperatively by both of its PDZ domains.

Preparation of the Structures. The high-resolution structure of the syntenin's PDZ2 domain (PDB code: 1R6J, called 1R6J in the following), including hydrogen atom coordinates, was kindly provided by Prof. Z. S. Derewenda. For disordered parts of the protein, only major conformers were taken into account. The missing hydrogen of the thiol group of Cys239 was generated by the program Reduce[41] from the MolProbity Web service.[42] Two tautomers of His208 were tested with the proton at ND1 (original position) and at NE2. The latter, proposed by Reduce, was chosen for further calculations.

Structures of the PDZ2 domains in the complexes with peptides were taken from the protein data bank (PDB codes: 1OBX, 1OBY, 1OBZ, 1YBO, 1W9O, 1W9E, 1W9Q, and 1V1T). The Reduce program was used to add hydrogen atoms to the protein−peptide complexes by optimization of the hydrogen-bond network. All histidines were treated as singly protonated. The side chains of the Arg, Lys, Asp, and Glu residues were assigned as ionized by the program. Only major conformers of side chains were taken into account in

Table 4. Electrostatic Energy [kJ/mol] of Interaction between the Syntenin PDZ2 Domain and Truncated Peptides[a]

		P_{-3}	P_{-2}	P_{-1}	P_0	total ES
			Class I			
			X(S/T)XΦ[b]			
			DSVF			
1OBX: A	single domain	−172(−24)	−37(−32)	4(13)	−249(−235)	−454(−278)
1OBZ: A	tandem	−168(−64)	−45(−32)	−1(8)	−266(−233)	−480(−321)
			Class II			
			XΦXΦ			
			EFYA			
1OBY: A	single domain	−199(−48)	−106(−94)	27(11)	−193(−194)	−471(−325)
B	single domain	−206(−50)	−113(−101)	5(8)	−197(−201)	−512(−344)
1YBO: A	tandem	−169(−60)	−121(−107)	0(−10)	−170(−171)	−460(−348)
B	tandem	−178(−20)	−87(−69)	−12(10)	−188(−185)	−464(−264)
			EYYV			
1W9O: A	tandem	−163(−47)	−85(−56)	−25(10)	−138(−138)	−411(−231)
B	tandem	−213(−65)	−125(−100)	5(11)	−143(−86)	−476(−240)
			EYKV			
1V1T: A	tandem	−163(−49)	−100(−81)	20(12)	−176(−180)	−419(−298)
B	tandem	−173(−50)	−134(−103)	−15(3)	−181(−191)	−502(−341)
			EFYF			
1W9E: A	tandem	−180(−49)	−104(−94)	−5(1)	−247(−204)	−536(−346)
B	tandem	−199(−60)	−169(−159)	−20(−5)	−237(−218)	−626(−442)
			EFAF			
1W9Q: A	tandem	−193(−56)	−120(−107)	16(17)	−283(−246)	−580(−392)

[a] Columns labeled "P_{-3}" to "P_0": individual peptide residues interacting with the protein. Column labeled "total ES": total electrostatic energy of four-residue peptide interacting with the protein. Numbers in parenthesis correspond to interactions for main-chain atoms of peptides only.[b] X is any amino acid, Φ is a residue with hydrophobic side chains, S/T is either serine or threonine, D is aspartic acid, V is valine, F is phenylalanine, E is glutamic acid, Y is tyrosine, A is alanine, and K is lysine.

the calculations. All protein chains were truncated to 73-residue-long fragments starting from Thr196 and finishing with the Thr268 residue, because the N and C termini of the PDZ2 domain are not in the vicinity of the peptide-binding groove. The N and C termini of the protein chains were capped with the neutral acetyl and methylamino blocking groups, respectively. All peptides were truncated to C-terminus four-residue fragments, as the residues upstream from P_{-3} are solvent-exposed and not involved in the protein−peptide interaction. The C termini of peptides were deprotonated. Ideally, the NH_2-terminal group of P_{-3} residues should be capped with the neutral acetyl group. However, the lack of coordinates for the P_{-4} residue in some structures did not allow construction of the capping groups. Therefore, the neutral form of the peptide N termini was used to avoid introducing artificial positive charges, which would have biased the calculation.

In all cases, ions and water molecules were omitted in the calculations. The type and position of the ions vary among the structures considered, which prevents a direct analysis of their relevance for the electrostatic interactions. The variability among the ions in the structures results most likely from crystallization conditions and does not reflect the native structure of the syntenin PDZ2 domain. Only a few water molecules have similar positions from one structure to another, but none of them is present in all structures. The position of the water molecules suggests that their contribution to the binding of the ligand is minor.

Results

Charge Density Reconstruction. Less than 1 min on an AthlonMP 1800+ CPU is required by LSDB to assign atom types to more than 1000 protein atoms. The CPU-time scales approximately linearly with the number of atoms. The sum of the valence populations obtained with the databank differs by only 0.08% from the target value, with 2.6 and 2.4 electrons missing for the assembly of 1100−1200-atom 1R6J and truncated PDZ domains, respectively. This demonstrates the excellent self-consistency of the databank. To ensure the electroneutrality, the monopole populations of the pseudo-atoms were scaled a posteriori, using the Faerman and Price[43] scaling algorithm implemented in LSDB.[1] The deformation densities shown in Figure 12 illustrate that all bonding and electron pair features are very well reconstructed.

Electrostatic Potentials. Electrostatic potentials for uncomplexed PDZ2 domains are computed from the charge density distributions according to the method of Su and Coppens[44] available in the XDPROP module of the XD package. An example of the electrostatic potential +0.05 and −0.05 *e*/Å isosurfaces and the electrostatic potential mapped on the 0.01 *e* isodensity surface (for 1W9E:B) is shown in Figure 13. All PDZ2 domains exhibit the strong polarization evident in Figure 13a. The binding groove starts on the border between negative and positive potentials and extends into the positive potential region. The electrostatic potentials of the PDZ2 domain and the peptide (both calculated for the nonbonded molecules) mapped on the molecular surface reveals as expected that the PDZ2−peptide complex is

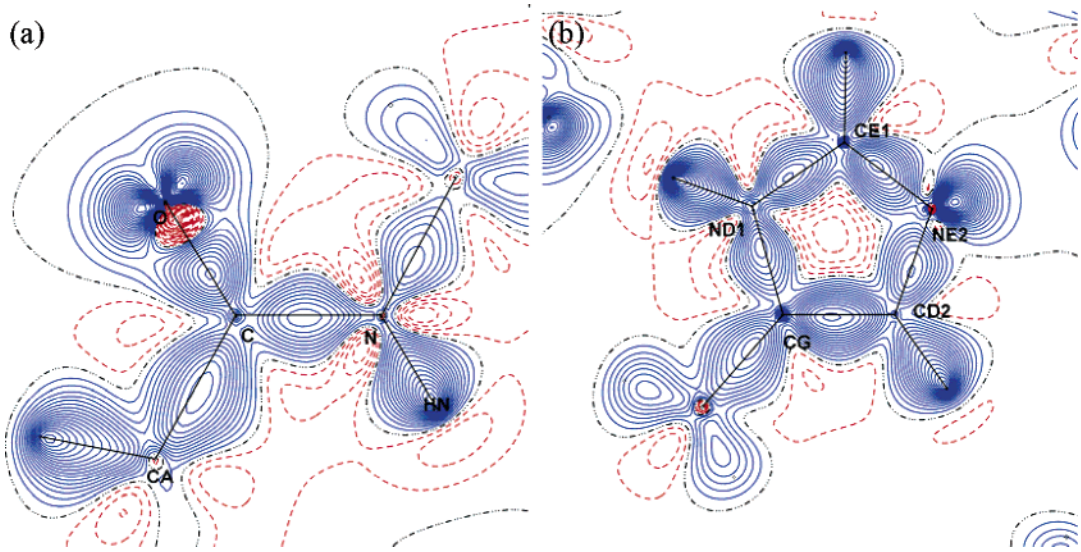


Figure 12. Deformation densities in the syntenin PDZ2 domain in the plane of (a) the peptide bond between Gly210 and Phe211 and (b) the imidazole ring of His208 with ND1 nitrogen protonated. Contour levels at 0.05 *e*/Å^3; positive contours, blue; negative contours, red; and zero contour, black.

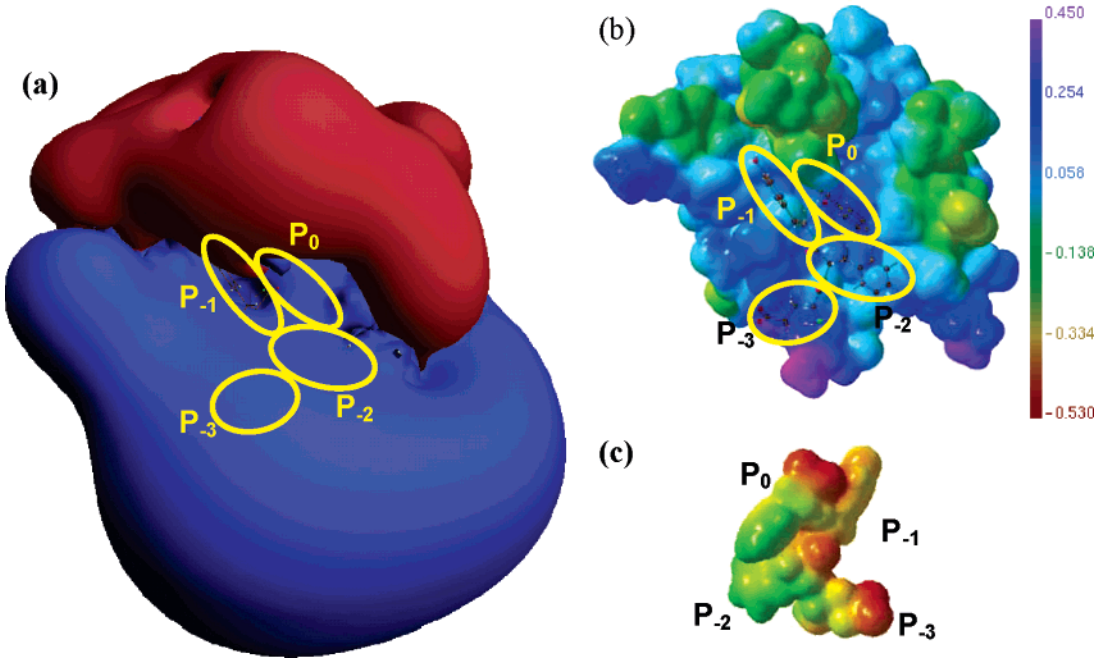


Figure 13. Electrostatic properties of the uncomplexed syntenin PDZ2 domain exemplified by the 1W9E:B structure. (a) Electrostatic potential of the PDZ2 domain calculated from the databank-generated charge density; isosurfaces: blue +0.05 *e*/Å, red −0.05 *e*/Å. Electrostatic potential [*e*/Å] of the PDZ2 domain (b) and peptide (c) mapped on the 0.01 *e* isodensity surface. Orientation of the protein as in Figure 16a; the peptide is rotated by 180°. The peptide molecule in the ball-and-stick model is presented only to illustrate the binding site and was not included in the electrostatic potential calculations.

formed between surfaces whose electrostatic potentials are complementary (Figure 13b and c).

Electrostatic Interactions. Electrostatic interactions are evaluated with the exact potential/multipole moment (EPMM) method of Volkov et al.[2,45] implemented in the XDPROP module of the XD package. The EPMM method combines numerical evaluation of the exact Coulomb integral for the short-range interactions between atoms within a certain limiting distance (here, taken as 4.5 Å) with the Buckingham-type multipole approximation for the long-range interatomic interactions between atoms for which no charge density overlap occurs. The computational effort of the calculation of the interaction between distant atoms within the Buckingham approximation is negligible when compared to the numerical evaluation, the latter scales linearly with the number of the short-range interactions. It takes 78 s to process 100 interactions on an AthlonMP 1800+ CPU.

PDZ2 Dimer. In the 1R6J structure, the peptide binding groove is occupied by the C-terminal tails of the neighboring molecule, mimicking the recognition of the peptide ligand. The PDZ2 dimer serves as an example of protein−protein interactions. Application of the EPMM/databank method to

the 1R6J structure gives a value of −526 kJ/mol (−388 kJ/mol for the HisD tautomer) for the electrostatic energy of interaction between the PDZ2 domain and its symmetry-related ligand-mimicking counterpart. For comparison, the MMFF94 force field[46] implemented in Sybyl 7.1[47] gives a value of −742 kJ/mol for the electrostatic interaction (−345 kJ/mol for the HisD tautomer).

The 138 kJ/mol difference between the electrostatic interaction energies of the PDZ2 dimer with different tautomeric forms of His208 originates almost entirely from the difference in the His208−Glu235 interaction, which is +97 kJ/mol for the D tautomer and −33 kJ/mol for the E tautomer. The NE2 nitrogen from the imidazole ring of His208 is only 2.64 Å from the OE2 oxygen from the side-chain carboxylic group of Glu235 of the ligand-mimicking molecule. Such a close distance can only be explained by formation of a NE2−HE2···OE2 hydrogen bond, which is clearly supported by the above results. Therefore, the PDZ2 domain with His208 protonated at NE2 was chosen for further calculations.

The individual electrostatic contributions of each residue of the receptor molecule interacting with the whole ligand-mimicking molecule in the PDZ2 complex are presented in Figure 14a−c. Only a small number of receptor protein residues have large contributions to the binding energy, both attractive and repulsive. Among them are Lys203, Asp204, Thr206, His208, Val209, Gly210, Phe211, and Lys214 from the β2 strand and the preceding conserved glycine-rich loop. Additionally, isolated strongly attractive and repulsive electrostatic interactions with the ligand-mimicking molecule are observed for the charged residues (Lys223 and Asp224, Lys250, Asp251, and Arg229). Three PDZ domain fingerprint residues, Val209, Gly210, and Phe211, interact electrostatically with the ligand protein, mainly through main-chain atoms (Figure 14b). The remaining interactions are with side chains of the interacting residues (Figure 14c).

Comparison of the individual residue−residue electrostatic interactions reveals that the larger absolute values are observed for interactions between charged residues (in the range from 162 to −126 kJ/mol, on average ±26 kJ/mol). However, interactions between these almost cancel each other, resulting in only a −86 kJ/mol contribution to the total electrostatic interaction energy.

PDZ2−Peptide Complexes. PDZ2 domains complexed with short peptides are an excellent example of protein−ligand interactions. We analyzed all eight PDB-available structures of syntenin domains interacting with short peptides. Six of these structures (1OBY, 1YBO, 1W9O, 1W9E, 1W9Q, and 1V1T) represent tandems of PDZ1 and PDZ2 domains. The remaining two (1OBX and 1OBY) consist solely of the PDZ2 domain. Only the PDZ2 domains were analyzed in this study; the PDZ1 domains were removed from the tandem structures. The structures of complexes superimposed on each other are shown in Figure 15a. The class II peptides (EFYA, EYYV, EYKV, EFYF, and EFAF) interact with the PDZ2 domain as the canonical model predicts, see Figure 15b. The DSVF (class I) peptides are not fully inserted in the binding groove.

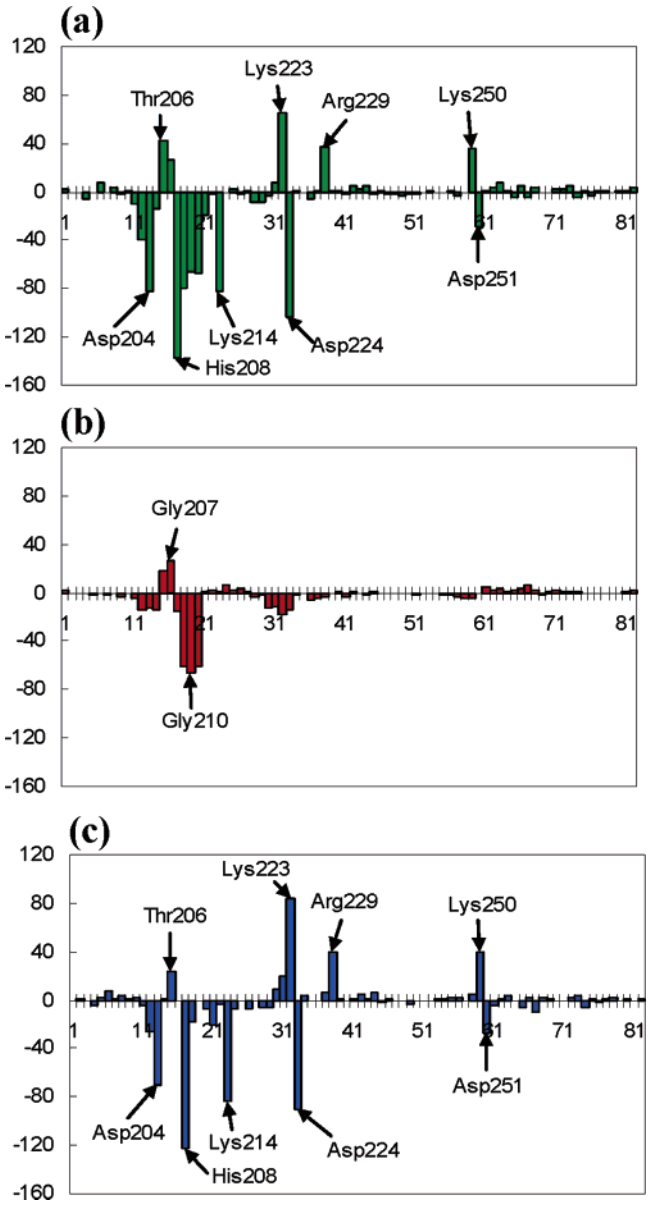


Figure 14. Electrostatic contributions (kJ/mol) of individual amino acid residues of the PDZ2 domain in the 1R6J structure interacting with the symmetry-related domain, mimicking the native ligand/PDZ2 interaction (a) all interactions, (b) main-chain/PDZ2 interactions, and (c) side-chain/PDZ2 interactions.

For the PDZ2 domains, the electrostatic energy of interaction between the protein and the truncated peptide is found to range from −411 to −626 kJ/mol (Table 4).

Several of the peptides contain Phe as P_0. It is the strongest interacting residue out of all of the residues examined, with the electrostatic interaction energy averaged over all occurrences equal to 256(18) kJ/mol. The main-chain atoms of P_0 (Phe) are the major contributors to the interaction energy, whereas 11% (29 kJ/mol) is provided by side-chain atoms. When P_0 is Ala or Val, the interaction energies are lower by about 80 kJ/mol. With the exception of the 1W9O:B structure, no significant contributions of the side-chain atoms of Ala or Val are observed. Interestingly, the extra interaction energy of Phe is almost equally distributed between the side-chain and main-chain atoms, suggesting that the side-chain phenyl ring enforces a better fit of the main chain. The results

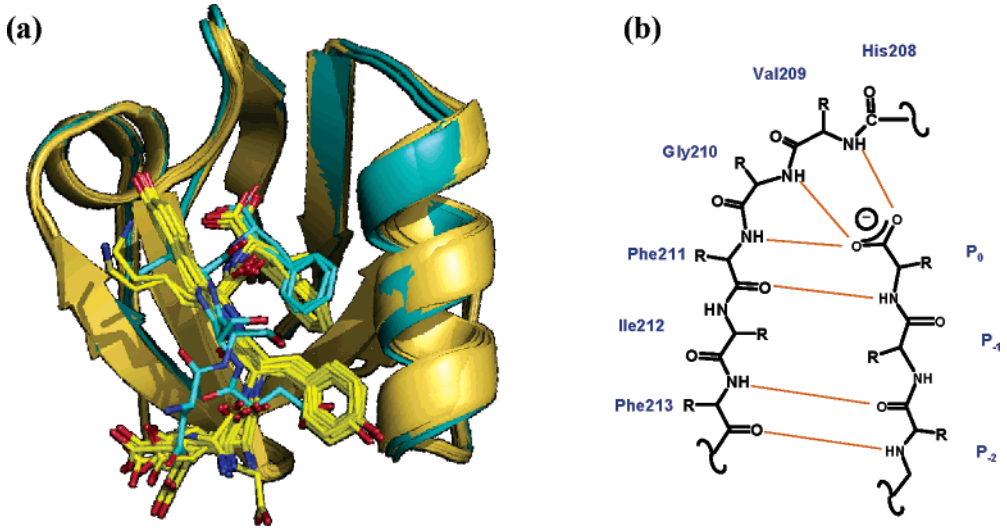


Figure 15. Structural details of the syntenin PDZ2 domain complexes. (a) Superposition of PDZ2 domains interacting with class I (blue; 1OBY and 1OBZ) or class II peptides (yellow; 1OBY, 1YBO, 1W9O, 1VLT, 1W9E, 1W9Q). (d) The canonical hydrogen bonds formed between the protein (left) and the peptide (right) in the syntenin PDZ2/peptide complex.

indicate stronger binding for phenylalanine than for alanine, as noted by Grembecka et al.,[12] but are not in agreement with the experimental dissociation constants for the interaction of the TNEFYA (with alanine at P_0) and TNEFYF (with phenylalanine at P_0) hexapeptides with PDZ2 in tandem, which are 0.16(1) and 0.79 (3) mM, respectively.[10] Grembecka et al. attributed this discrepancy to an energetically unfavorable rearrangement of the PDZ2 domain to accommodate the large phenyl group, which is reflected in the dissociation constants but not in the interaction energies.

It is interesting that residues P_{-1} do not contribute to electrostatic binding energy [av 0(15) kJ/mol]. Studies have shown that this residue is not conserved among the C termini of ligand proteins that interact with PDZ domains and is considered unspecific. However, this does not explain the fact that the replacement of Tyr (Y) at the P_{-1} position by Lys (K) is observed to decrease the binding. Dissociation constants for interaction of the TNEYYV and TNEYKV hexapeptides with PDZ2 in tandem have been found to be equal to 0.10(1) and 1.15(9) mM, respectively.[10]

The P_{-2} residue of the ligand is the major cause of PDZ specificity. The Ser P_{-2} residue common to class I peptides interacts much weaker with the protein [av −41(6) kJ/mol] than the aromatic P_{-2} residues [Phe and Tyr, av −115(24) kJ/mol] of class II peptides. There is no clear difference in interaction strength between Phe and Tyr at the P_{-2} position.

It is noteworthy that the P_{-3} residues [negatively charged Asp (D) or Glu (E)] significantly strengthen the electrostatic interaction [av −183(17) kJ/mol], even though they do not have short contacts with the protein and are solvent-exposed. The interactions are mainly maintained by side-chain atoms, while the main chains contribute only in 27% (−49 kJ/mol in average).

Although the interaction energies of the individual residues vary, there is no substantial difference between the electrostatic energies of class I and class II peptides interacting with the syntenin PDZ2 domain. This result is in agreement with fluorometric dissociation constants for the interactions of IL5Rα- and syndecan-derived dansylated peptides with PDZ2 in tandem, which are very similar [1.9(3) and 2.3(5) μM for EDSVF and NEFYA, respectively].[48] Although the interaction strength of the P_{-2} residue is indeed weaker for class I peptides, this loss of interaction energy is compensated by stronger binding of the P_0 and P_{-3} residues.

There is no difference between the interaction energies of the short peptides with PDZ2 domains from single-domain protein fragments and those from tandem protein fragments from which PDZ1 has been removed (see single domains vs tandems, Table 4).

A more detailed analysis of the individual atom−atom interactions gives information on the types of electrostatic interactions which contribute most to peptide binding by the syntenin PDZ2 domain. Major contributions come from hydrogen bonds between main-chain atoms of the Val209−Phe211 fragment and main-chain atoms of P_0 peptide residues (including the COO^--terminal group; Table 5). The two strongest electrostatic interactions are Val209 H···OXT P_0 and Gly210 H···O P_0 hydrogen bonds, which average −53(10) and −49(6) kJ/mol, respectively. Hydrogen bonds between Phe213 and P_{-2} also contribute to the binding. Equally important are charge−charge interactions between side chains of Lys214 and the P_{-3} residue [−80(18) kJ/mol on average], though they are less numerous than the hydrogen bonds.

Conclusions

A comprehensive version of the theoretical databank of transferable aspherical pseudoatoms has been developed and applied for the first time to protein−ligand interaction energies. The databank consists of all atom types encountered in natural amino acid residues and a number of other biologically relevant molecules. Each atom type results from averaging over a family of chemically unique pseudoatoms, taking into account both first and second neighbors. A new atom type is spawned when one of the parameters for a subgroup of the family deviates more than one standard

Table 5. Most Important Atom−Atom Contributions [kJ/mol] to the Total Electrostatic Interaction Energies for the Syntenin PDZ2 Domain Complexed with Truncated Peptides

interacting atoms	1OBX: AB	1OBZ: AP[a]	1OBY: AB	1OBY: BQ	1YBO: AC	1YBO: BD[a]	1W9O: AT	1W9O: BS	1VLT: AT	1VTL: BS	1W9E: AT	1W9E: BS	1W9Q: BS
						In H Bonds[b]							
Val209 H···OXT P_0	−60	−57	−58	−52	−41	−35	−40	−45	−63	−52	−55	−56	−72
Val209 H···O P_0	−16	−18	−17	−17	−16	−17	−16	−19	−15	−15	−15	−17	−18
Gly210 H···OXT P_0	−16	−16	−16	−17	−16	−15	−16	−16	−19	−18	−17	−17	−19
Gly210 H···O P_0	−63	−51	−51	−52	−40	−52	−42	−46	−45	−46	−49	−48	−53
Phe211 H···O P_0	−37	−38	−32	−32	−33	−29	−28	−27	−24	−28	−33	−35	−30
Phe211 O···HP_0	−31	−37	−38	−38	−42	−70	−31	−18	−36	−37	−46	−38	−69
Phe213 H···O P_{-2}	−8	−8	−30	−33	−30	−33	−22	−29	−27	−32	−36	−79	−38
Phe213 O···H P_{-2}	−4	−2	−32	−30	−34	−16	−19	−36	−25	−34	−25	−34	−37
						In Ionic Interactions							
Lys214 NH_3^+···$^-$OOC P_{-3}	−109	−86	−105	−102	−54	−87	−62	−86	−62	−72	−65	−70	−77

[a] OXT and O oxygens of P_0 residue are interchanged compared to labeling scheme in original PDB file. [b] Only interactions between hydrogen and acceptor atoms are taken into account because interactions with other atoms are very weak, 0(1) kJ/mol on average.

deviation of the sample from the average of the remaining atoms. The algorithm for atom-type definition ensures that close transferability is obeyed.

The pseudoatoms as derived with the Hansen−Coppens formalism from theoretical wave functions show several regularities. The monopole-derived net atomic charge correlates strongly with the expansion−contraction parameter κ and somewhat less with κ', as found before.[26,27] In general, the atomic charges reflect the group-electronegativity concept as defined by Huheey. However, hydrogen pseudoatoms, for which the division of the charge in the bonding regions is particularly difficult, behave differently. Their populations significantly affect charges on the neighboring atoms. The charge of the atoms linked to hydrogens becomes more and more positive as the number of bonded hydrogen atoms increases, while the H-atom charge shifts toward negative values. The deformation density of atoms possessing lone pairs is dominated by the lone pair electrons, while the bonding regions contribute very little.

Application of the databank to the syntenin PDZ2 domain complexes allows a detailed analysis of the electrostatic interaction energies. The results stress the importance of the P_0 and P_{-2} residues of the peptide in establishing the interaction, whereas the P_{-1} residue is shown to play a much smaller role. Unexpectedly, the charged P_{-3} residue contributes significantly also.

Class I and class II peptides are bound with the same strength by the syntenin PDZ2 domain. Although the electrostatic interaction energy of the P_{-2} residue is smaller for class I peptides, this loss is compensated by stronger binding of the P_0 and P_{-3} residues.

In agreement with previous conclusions,[10] our results confirm that there is no difference between the interaction energies of short peptides with PDZ2 domains from single-domain protein fragments and those from PDZ1−PDZ2 tandem protein fragments. The experimentally observed cooperation of PDZ1 and PDZ2 domains upon binding the target protein has to be attributed to interactions in other regions of the complexes.

Although the electrostatic energy is the dominant component in the hydrogen-bonded systems studied, other contributions such as dispersion, exchange-repulsion, and induction terms must be considered. We are developing a set of symmetry-adapted perturbation-theory-based pairwise atom−atom potentials which will be applied in a subsequent study. Entropic effects, such as solvation and configurational changes in solution, must also be taken into account in a comprehensive treatment.

Acknowledgment. We thank Prof. Zygmunt S. Derewenda for providing the coordinates of 1R6J including hydrogen atoms. Support of this work by the National Institutes of Health through Grant GM56829 and the National Science Foundation through Grant CHE0236317 is gratefully acknowledged. Molecular graphics have been made with programs MolIso[49] and PyMOL.[56]

Supporting Information Available: Table S1 with the CSD codes, chemical diagrams, names, formulas, and references for all compounds used for the construction of the databank and Tables S2−S8 with deformation density parameters for atom types present in the current version of the databank. This material is available free of charge via the Internet at http://pubs.acs.org. The LSDB program and the full databank are available at http://harker.chem.buffalo.edu.

References

(1) Volkov, A.; Li, X.; Koritsánszky, T.; Coppens, P. *Ab Initio* Quality Electrostatic Atomic and Molecular Properties Including Intermolecular Energies from a Transferable Theoretical Pseudoatom Databank. *J. Phys. Chem.* **2004**, *108*, 4283−4300.

(2) Volkov, A.; Koritsánszky, T.; Coppens, P. Combination of the Exact Potential and Multipole Methods (EP/MM) for Evaluation of Intermolecular Electrostatic Interaction Energies with Pseudoatom Representation of Molecular Electron Densities. *Chem. Phys. Lett.* **2004**, *391*, 170−175.

(3) Pichon-Pesme, V.; Lecomte, C.; Lachekar, H. On Building a Data Bank of Transferable Experimental Electron Density Parameters Applicable to Polypeptides. *J. Phys. Chem.* **1995**, *99*, 6242−6250.

(4) Pichon-Pesme, V.; Jelsch, C.; Guillot, B.; Lecomte, C. A Comparison between Experimental and Theoretical Aspherical-Atom Scattering Factors for Charge-Density Refinement of Large Molecules. *Acta Crystallogr., Sect. A* **2004**, *60*, 204−208.

(5) Volkov, A.; Li, X.; Koritsánszky, T.; Coppens, P. Response to the Paper A Comparison between Experimental and Theoretical Aspherical-Atom Scattering Factors for Charge-Density Refinement of Large Molecules, by Pichon-Pesme, Jelsch, Guillot & Lecomte 2004. *Acta Crystallogr., Sect. A* **2004**, *60*, 638−639.

(6) Li, X.; Volkov, A. V.; Szalewicz, K.; Coppens, P. Interaction Energies between Glycopeptide Antibiotics and Substrates in Complexes Determined by X-ray Crystallography: Application of a Theoretical Databank of Aspherical Atoms and a Symmetry-Adapted Perturbation Theory-Based Set of Interatomic Potentials. *Acta Crystallogr., Sect. D* **2006**, *62*, 639−647.

(7) Dittrich, B.; Koritsánszky, T.; Luger, P. A Simple Approach to Nonspherical Electron Densities by Using Invarioms. *Angew. Chem., Int. Ed.* **2004**, *43*, 2718−2721.

(8) Dittrich, B.; Hübschle, C. B.; Messerschmidt, M.; Kalinowski, R.; Girnt, D.; Luger, P. The Invariom Model and Its Application: Refinement of D,L-Serine at Different Temperatures and Resolution. *Acta Crystallogr., Sect. A* **2005**, *61*, 314−320.

(9) Dittrich, B.; Strumpel, M.; Schäfer, M.; Spackman, M. A.; Koritsánszky, T. Invarioms for Improved Absolute Structure Determination of Light-Atom Crystal Structures. *Acta Crystallogr., Sect. A* **2006**, *62*, 217−223.

(10) Walker, P. D.; Mezey, P. G. Molecular Electron Density Lego Approach to Molecule Building. *J. Am. Chem. Soc.* **1993**, *115*, 12423−12430.

(11) Szekeres, Zs.; Exner, T. E.; Mezey, P. G. Fuzzy Fragment Selection Strategies, Basis Set Dependence, and HF − DFT Comparisons in the Applications of the ADMA Method of Macromolecular Quantum Chemistry. *Int. J. Quantum Chem.* **2005**, *104*, 847−860.

(12) Grembecka, J.; Cierpicki, T.; Devedjiev, Y.; Derewenda, U.; Kang, B. S.; Bushweller, J. H.; Derewenda, Z. S. The Binding of the PDZ Tandem of Syntenin to Target Proteins. *Biochemistry* **2006**, *45*, 3674−3683.

(13) Kang, B. S.; Cooper, D. R.; Devedjiev, Y.; Derewenda, U.; Derewenda, Z. S. Molecular Roots of Degenerate Specificity in Syntenin's PDZ2 Domain: Reassessment of the PDZ Recognition Paradigm. *Structure* **2003**, *11*, 845−853.

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

(15) Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta*. **1973**, *28*, 213−222.

(16) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(17) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785−789.

(18) Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr., Sect. B* **2002**, *58*, 380−388.

(19) C−H: 1.099 Å for primary and 1.092 Å for secondary carbons, 1.059 Å for methyl groups connected to carbon and 1.066 Å for methyl groups connected to a non-carbon atom, 1.083 Å for aromatic carbons and 1.077 Å for nonaromatic sp^2 carbons. N−H: 1.033 Å for tetravalent and 1.009 Å for trivalent nitrogens. O−H: 1.015 Å for carboxylic acids and 0.967 for other hydroxyl groups. S−H 1.338 Å for thiols. Wilson, A. J. C.; Prince, E. *International Tables for Crystallography*; Kluwer Academic Publishers: Dordrecht, The Netherlands1992; Vol. C.

(20) Koritsánszky, T. S.; Volkov, A. V.; Coppens, P. Aspherical-Atom Scattering Factors from Molecular Wave Functions. 1. Transferability and Conformation Dependence of Atomic Electron Densities of Peptides within the Multipole Formalism. *Acta Crystallogr., Sect. A* **2002**, *58*, 464−472.

(21) Hansen, N. K.; Coppens, P. Testing Aspherical Atom Refinements on Small-Molecule Data Sets. *Acta Crystallogr., Sect. A* **1978**, *34*, 909−921.

(22) Koritsánszky, T. S.; Howard, S.; Macchi, P.; Gatti, C.; Farrugia, L. J.; Mallinson, P. R.; Volkov, A.; Su, Z.; Richter, T.; Hansen, N. K. *XD, a Computer Program Package for Multipole Refinement & Analysis of Electron Densities from Diffraction Data*, version 4.12, October 2004; Free University of Berlin: Berlin, Germany; University of Wales: Cardiff, U. K.; Universitá di Milano: Milano, Italy; CNR−ISTM: Milano, Italy; University of Glasgow: Glasgow, U. K.; State University of New York: Buffalo, NY; University of Nancy: Nancy, France, 2004. Home page: http://xd.chem.buffalo.edu/ (accessed May 2006).

(23) Clementi, E.; Roetti, C. Roothan-Hartree−Fock Atomic Wave Functions. *At. Data Nucl. Data Tables* **1974**, *14*, 177−478.

(24) Clementi, E.; Raimondi, D. L. Atomic Screening Constants from SCF Functions. *J. Chem. Phys.* **1963**, *38*, 2686−2692.

(25) Dominiak, P. M.; Coppens, P. Finding Optimal Radial-Function Parameters for S Atoms in the Hansen−Coppens Multipole Model through Refinement of Theoretical Densities. *Acta Crystallogr., Sect. A* **2006**, *62*, 224−227.

(26) In the XD procedure, two vectors are defined by neighboring atoms. The first from the original atom to atom 1 defines vector 1, which is also axis 1; vector 2 is defined by atoms 2 and 3, atom 2 being usually identical to the original atom. Axis 2 is defined as v1 × v2, and axis 3 is the cross product of axes 1 and 2. For full details, see the XD manual.

(27) The first number of the atom-type designation represents the number of bonded atoms; the following numbers are a sequential number.

(28) Coppens, P.; Guru Row, T. N.; Leung, P.; Stevens, E. D.; Becker, P. J.; Yang, Y. W. Net Atomic Charges and Molecular Dipole Moments from Spherical-Atom X-ray Refinements, and the Relation between Atomic Charge and Shape. *Acta Crystallogr., Sect. A* **1979**, *35*, 63−72.

(29) Volkov, A.; Abramov, Y. A.; Coppens, P. Density-Optimized Radial Exponents for X-ray Charge-Density Refinement from ***ab initio*** Crystal Calculation. *Acta Crystallogr., Sect. A* **2001**, *57*, 272−282.

(30) Fitted linear equation for H: $\kappa' = 1.26(2) + 1.1(2)q$, $R = 0.90$; for C: $\kappa' = 0.888(3) + 0.17(1)q$, $R = 0.89$.

(31) Fitted linear equation for P_{10}/P_{val}: $P_{10} = -0.06(4) + 0.22(5)P_{val}$, $R = 0.85$; for P_{20}/P_{val}: $P_{20} = -0.04(3) + 0.12(3)P_{val}$, $R = 0.81$.

(32) Fitted linear equation for P_{31+}/P_{val}: $P_{31+} = 0.2(1) - 0.10(3)P_{val}$, $R = -0.60$; for P_{31-}/P_{val}: $P_{31-} = 0.27(6) - 0.12(2)P_{val}$, $R = -0.86$; for P_{33+}/P_{val}: $P_{33+} = -0.2(1) + 0.11(3)P_{val}$, $R = 0.62$; for P_{33-}/P_{val}: $P_{33-} = 0.10(4) - 0.04(1)P_{val}$, $R = -0.60$; for P_{42-}/P_{val}: $P_{42-} = -0.27(6) + 0.09(2)P_{val}$, $R = 0.81$; for P_{44+}/P_{val}: $P_{44+} = -0.21(6) + 0.07(1)P_{val}$, $R = 0.72$.

(33) Fitted linear equation for P_{20}/P_{11+}: $P_{20} = 0.014(4) - 0.57(7)P_{11+}$, $R = -0.88$; for P_{22+}/P_{11+}: $P_{22+} = -0.018(4) + 1.125(7)P_{11+}$, $R = 0.97$.

(34) Huheey, J. E. The Electronegativity of Groups. *J. Phys. Chem.* **1965**, *69*, 3284−3291.

(35) Kang, B. S.; Devedjiev, Y.; Derewenda, U.; Derewenda, Z. S. The PDZ2 Domain of Syntenin at Ultrahigh Resolution: Bridging the Gap between Macromolecular and Small Molecule Crystallography. *J. Mol. Biol.* **2004**, *338*, 483−493.

(36) 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. http://www.pdb.org/ (accessed Apr 10, 2006).

(37) Hung, A. Y.; Sheng, M. PDZ Domains: Structural Modules for Protein Complex Assembly. *J. Biol. Chem.* **2002**, *277*, 5699−5702.

(38) Sheng, M.; Sala, C. PDZ Domains and the Organization of Supramolecular Complexes. *Annu. Rev. Neurosci.* **2001**, *24*, 1−29.

(39) Doyle, D. A.; Lee, A.; Lewis, J.; Kim, E.; Sheng, M.; MacKinnon, R. Crystal Structures of a Complexed and Peptide-Free Membrane Protein-Binding Domain: Molecular Basis of Peptide Recognition by PDZ. *Cell* **1996**, *85*, 1067−1076.

(40) Morais Cabral, J. H.; Petosa, C.; Sutcliffe, M. J.; Raza, S.; Byron, O.; Poy, F.; Marfatia, S. M.; Chishti, A. H.; Liddington, R. C. Crystal Structure of a PDZ Domain. *Nature* **1996**, *382*, 649−652.

(41) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-Chain Amide Orientation. *J. Mol. Biol.* **1999**, *285*, 1735−1747.

(42) Lovell, S. C.; Davis, I. W.; Arendall, W. B., III; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. Structure Validation by Cα Geometry: Φ, Ψ, and Cβ Deviation. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 437−450.

(43) Faerman, C. H.; Price, S. L. A Transferable Distributed Multipole Model for the Electrostatic Interactions of Peptides and Amides. *J. Am. Chem. Soc.* **1990**, *112*, 4915−4926.

(44) Su, Z. W.; Coppens, P. On the Mapping of Electrostatic Properties from the Multipole Description of the Charge Density. *Acta Crystallogr., Sect. A* **1992**, *48*, 188−197.

(45) Volkov, A.; King, H. F.; Coppens, P. Dependence of the Intermolecular Electrostatic Interaction Energy on the Level of Theory and the Basis Set. *J. Chem. Theory Comput.* **2006**, *2*, 81−89.

(46) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(47) *SYBYL 7.1*.; Tripos Inc.: St. Louis, Missouri.

(48) Kang, B. S.; Cooper, D. R.; Jelen, F.; Devedjiev, Y.; Derewenda, U.; Dauter, Z.; Otlewski, J.; Derewenda, Z. S. PDZ Tandem of Human Syntenin: Crystal Structure and Functional Properties. *Structure* **2003**, *11*, 459−468.

(49) Hübschle, C. B. *MolIso−A Program for Color Mapped Iso-surfaces*, 2005; http://userpage.chemie.fu-berlin.de/~chuebsch/moliso/ (accessed Apr 7, 2006).

(50) DeLano, W. L. *The PyMOL Molecular Graphics System*; DeLano Scientific: San Carlos, CA, 2002. http:/www.pymol.org. (accessed Jun 6, 2005).

CT6001994

JCTC Journal of Chemical Theory and Computation

Computational Analysis of Current and Noise Properties of a Single Open Ion Channel

Enrico Piccinini,[†] Fabio Affinito,[‡] Rossella Brunetti,*[,‡] Carlo Jacoboni,[‡] and Massimo Rudan[†]

Advanced Research Center on Electronic Systems and Dipartimento di Ingegneria Elettronica, Informatica e Sistemistica, Alma Mater Studiorum, Università di Bologna, Bologna, Italy, and Consiglio Nazionale delle Ricerche−Istituto Nazionale di Fisica della Materia S3 Research Center and Dipartimento di Fisica, Università di Modena e Reggio Emilia, Modena, Italy

Received June 21, 2006

Abstract: This paper presents a computational analysis of the noise associated with the ion current in single open ion channels. The study is performed by means of a coupled molecular dynamics/Monte Carlo approach able to simulate the conduction process on the basis of all microscopic information today available from protein structural data and atomistic simulations. The case of potassium ions permeating the KcsA channel is considered in the numerical calculations. The results show a noise spectrum different from what is theoretically predicted for Poisson noise, confirmed by the existence of a correlation in ion-exit events.

Introduction

Rather than merely a bothersome phenomenon, noise in biological systems is a useful property.[1] Before the patch-clamp technique was developed, current noise across biological membranes provided the first experimental evidence for the existence of ion-conducting pores with discrete conductance levels. Single-channel recording techniques are nowadays widely accessible, and open-channel noise can be measured with a standard experimental setup. However, noise analysis to study the kinetics of ions permeating a membrane channel is at present still a difficult task, because it is actually not possible to detect the individual shot events using the patch-clamp technique. The associated noise is so much faster than the time resolution of typical amplifiers used that it would not contribute much variance to the signal at the experimental filter settings;[2] however, it is possible to measure the increase in the white noise level when the channel opens.[3,4]

From the computational side, the availability of X-ray crystallographic structures with atomic resolution[5] provided the necessary input to molecular dynamics (MD) analyses.[6] MD provides atomistic information about the system and confirms that the permeation process takes place as a single-file concerted motion of ions. MD, however, is unable to provide directly the electrical properties of ionic flux due to the long time scale involved in the physiological permeation process. In the literature, a coupled MD−Brownian dynamics method has already been proposed[7] to simulate conduction properties of potassium ions inside KcsA which reproduces the experimental data only at low applied voltages (also referred to in the following as "low-bias range"). From the computational point of view, in a Brownian dynamics trajectory, only few events resulting in the hopping of ions can be observed, and such translocations normally occur a few times per nanosecond: most of the simulation time is spent for intrasite thermal motion. Alternatively, the analysis of conduction properties through an analytical solution of a kinetic model has recently been reported.[8]

The aim of this work is to study the conduction and noise properties of potassium ions in the KcsA channels by means of a coupled molecular dynamics/Monte Carlo (MD/MC) simulation that yields both the current and its noise for a single channel under open-gate conditions and overcomes the numerical problems discussed above. The MC approach presented in this paper provides a numerical solution of the

* Corresponding author tel.: +39 059 205 5277; fax: +39 059 205 5616; e-mail: brunetti.rossella@unimore.it.

[†] Università di Bologna.

[‡] Università di Modena e Reggio Emilia.

10.1021/ct6002077 CCC: $37.00
Published on Web 11/03/2006

kinetic equation and allows the exploration of a more detailed and accurate microscopic model. Furthermore, the simulation can be "biased" in order to enhance the sampling of rare events.

As in the methods available in the literature, a multi-ion model based on the existence of ion binding sites inside the protein and at the protein boundaries is the starting point of the framework. The MD/MC procedure can be summarized in the following steps:

1. Select relevant states for the channel's binding-site occupation during the permeation process, and classify them into categories on the basis of translocation homology.

2. Run atomistic classical MD simulations without any applied external electric field, and obtain free-energy profiles for the previously identified transitions.

3. Determine rates for transitions between minima on the basis of a kinetic model. The same rate is associated with all of the transitions belonging to the same category, but differences may exist within the same category, depending on the direction of the flux.

4. Alter transition rates through the introduction of the effect of the external electric field assuming stepwise potential drops.

5. Perform MC simulations with the obtained rates to get the ion current as a numerical solution to this kinetic model.

6. Analyze ion-current fluctuations using tools from the noise analysis.

Partial results obtained from the MD/MC simulation of the ion permeation process include evaluations of the relevant energy barriers along given transition paths on the free-energy profile (end of step 2), current−voltage characteristics comparable with experiments[9] (end of step 5), and noise power spectra for different transmembrane potentials (end of step 6).

The analysis of open-channel current fluctuations confirms that the motions of the ions in the selectivity filter are strongly correlated, and a quantitative estimate of the degree of correlation between consecutive ion exits from the channel can be evaluated.

Theoretical Basis

(a) Conduction Model. It has been established from structural[5,10] and computational[11] analyses that the KcsA channel exhibits six stable binding sites, usually referred to as S_0 through S_4 and S_{cav}, in which, alternatively, potassium ions and water molecules reside. Furthermore, because of the strong electrostatic repulsion between ions inside the filter, ion hopping inside the channel takes place as a synchronous concerted motion of many ions, each ion moving from one site to an adjacent one. A free-energy barrier exists between two given ion occupancy configurations. These barriers range from few k_BT's to many tens of k_BT's (k_B is the Boltzmann constant and T the temperature in Kelvin), which suggests that only a restricted set of transitions actually determines the conduction properties of the channel.

The identification of the relevant configurations must account for a number of constraints imposed by what we know today from the atomic knowledge of the channel protein. The configurations used in the simulation of the permeation process presented in this paper have been selected taking into account the following constraints:[7] (a) single-file ion motion; (b) transitions involving configurations where ions moving inside the selectivity filter are allowed only through concerted motion; (c) fewer than two ions in the selectivity filter (sites S_1-S_4) are not accepted to prevent an unstable conformation of the selectivity filter itself.

As previously reported,[7] for an ion flux toward the extracellular environment, conduction is driven by the motion of ions from the cavity to site S_4 and, subsequently, by the concerted motion of ions in S_3 and S_1 into sites S_2 and S_0. In the following, we describe with greater detail the translocation path of a single ion from the intracellular reservoir to the extracellular environment (and vice versa).

The four-step cycles

$$(S_1,S_3,S_{cav}) \leftrightarrow (S_0,S_2,S_4) \leftrightarrow (S_2,S_4) \leftrightarrow (S_1,S_3) \leftrightarrow (S_1,S_3,S_{cav}) \text{ or}$$

$$(S_1,S_3,S_{cav}) \leftrightarrow (S_0,S_2,S_4) \leftrightarrow (S_2,S_4) \leftrightarrow (S_2,S_4,S_{cav}) \leftrightarrow (S_1,S_3,S_{cav})$$

are composed of a three-ion concerted motion [later indicated with the label (t)], a two-ion concerted motion [label (d)], an entry/exit in the cavity site S_{cav} [label (c)], and an exit/entry in the outer mouth site S_0 [label (m)], as already adopted.[8] Generally speaking, all of the possible transitions resulting from the reported rules can be classified into the aforementioned four groups, on the basis of how many ions simultaneously move, the occupation of the cavity site, or the outer mouth site. We also point out that the three-ion motion described above does not conflict with the translocation picture reported in the literature.[7] As a matter of fact, once the potassium ion in the cavity enters site S_4, previously filled with a water molecule, it forces this latter to the intermediate position between sites S_3 and S_4.[12] Because of both steric constraints and electrical repulsion between adjacent potassium ions, this configuration immediately either turns back into its original one or evolves to (S_0,S_2,S_4). The lifetime of the intermediate state is so short that this state may be ignored in the simulation for conduction purposes.

All of the transition rates can, in principle, be obtained, in the absence of an external field, by MD simulations as averages over many occurrences of the same pathway in the configuration space. In practice, each transition event requires MD simulation times too long to be spontaneously observed, and techniques designed in order to map specific regions of the free-energy profile (typically, umbrella sampling,[13] steered MD,[14,15] and metadynamics[16]) must be applied to circumvent this difficulty. The accuracy of the above-mentioned numerical estimates is dependent on the choice of the selected force field and reaction coordinates, and the numerical techniques used to speed up computational convergence. A comparison among results obtained from the three techniques on a test case based on an internal ion translocation within the KcsA protein is discussed elsewhere.[17] The probability per unit time $k_{A\to B}$ to undergo a transition from channel configuration A to configuration B is then calculated using the transition-state theory:

$$k_{A\rightarrow B} = A \exp(-\beta\epsilon_b) \quad (1)$$

The prefactor A may be expressed by[18] $A = \beta(M\omega_1\omega_2 D/2\pi)$, where D is the ion diffusion coefficient inside the channel, ϵ_b is the free-energy barrier separating the two states, M is the ion mass, ω_1 and ω_2 represent the parabolic-approximation parameters around the maximum and minimum of the energy profile as functions of the appropriate reaction coordinate, and $\beta = 1/k_B T$.

Equation 1, above, has been derived for a generic one-dimensional reaction coordinate. The representation of a multidimensional free-energy landscape, as the present case, strictly depends on the particular choice of the reaction coordinates, so that around minima and maxima paraboloids can be obtained with a different curvature for each chosen coordinate. Thus, the approximating frequencies depend on the curvilinear coordinate following the considered transition pathway and may vary from one representation to another. The order of magnitude of the parameters used in our simulations has been determined by a parabolic approximation of the maxima and minima in the one-dimensional free-energy profile as functions of the curvilinear coordinate; their final values have been fine-tuned applying detailed balance, so that in equilibrium conditions a net zero current results at zero bias.

The resulting transition rates used to generate the permeation paths in the MC procedure range from 1.08×10^8 s^{-1} to 9.93×10^8 s^{-1} and have been determined for a symmetric concentration of 100 mM of potassium ions in the reservoirs. A different concentration may affect both prefactors and barriers.

(b) Noise Analysis. Our noise analysis is restricted to the contribution coming from the electric current fluctuations generated by single-file diffusion through the very narrow channel in its open configuration. These current fluctuations are expected to arise from the discrete nature of current flow (ions move in discrete steps across the membrane) yielding "shot noise" analogous to what is observed in electronic devices. Additional sources of noise are associated with fluctuations of the energy barriers due to both thermal structural fluctuations of the protein and fluctuations on the protein structure coming from the interaction between the protein and the lipid bilayer. A third source of noise is associated with the gating process of the channel. The characteristic time scales of the three fluctuations listed above are different: nanoseconds for shot noise, picoseconds to nanoseconds for protein structural fluctuations (as MD simulations suggest), and milliseconds for fluctuations due to gating.[9]

Noise spectra in the frequency domain of the last two processes above have been experimentally obtained and studied in the past.[3,19] The contribution to the total noise spectrum coming from shot noise associated with ion flow is still difficult to measure. It can be observed in the low-frequency range as an increase in the "white" noise background when the channel opens. In the high-frequency range, the main limit comes from the noise associated with electrodes and electronics. A theoretical analysis of ion current fluctuations in these nanometric biological conductors on the nanosecond time scale is feasible only through the implementation of advanced atomistic approaches. This is actually the focus of our calculations. To this purpose, we assume that the channel is permanently in the open configuration. Furthermore, we will not consider the fluctuations in excess coming from internal motions of the channel protein. Theoretical calculations of current-noise spectra during single-file transport have been performed in the past by means of the analytical solution of macroscopic time-dependent single-file transport equations, based on a number of limiting hypotheses concerning the number of binding sites and ions inside the channel.[20] These calculations suggest that the presence of interactions between the ions in the channel can produce damped oscillations in the autocorrelation function of the microscopic current fluctuations. The availability of protein structural information and atomistic simulators allows us to calculate these noise properties on the basis of a more sound and general physical scheme and, in the future, to compare computational results with experimental results, now achievable with modern noise equipment. The autocorrelation function $C(\tau)$ of the current fluctuations $\delta I(t)$ can be numerically computed from the MC simulation assuming a stationary process. The spectral density of the current $S_I(\omega)$ is then obtained either by Fourier transforming $C(\tau)$ or by directly averaging the squared Fourier transform of the current fluctuations.

The classical theory of shot noise[21] assumes that charge transport occurs by means of instantaneous processes and that the charge movements are totally uncorrelated in time. A parameter defined as the noise-to-signal ratio for a random variable n, which could be estimated from a time window that, on average, contains several random events, is the Fano factor F:[22]

$$F = \frac{\sigma^2(n)}{\langle n \rangle} \quad (2)$$

where $\sigma^2(n)$ is the variance associated with the random variable and $\langle n \rangle$ is its mean value, both evaluated within the time step Δt. For the case here investigated, n is the number of ions crossing the channel mouth in the time interval Δt. Considering that ions flow across the channel in both directions, it holds that

$$n = n^+ - n^- \quad (3a)$$

where n^+ and n^- refer to the number of ions moving in the two opposite directions, and

$$\sigma^2(n) = \sigma^2(n^+ - n^-) = \sigma^2(n^+) + \sigma^2(n^-) - 2\,\mathrm{cov}(n^+,n^-) \quad (3b)$$

$\mathrm{cov}(n^+,n^-)$ being the covariance of the two fluxes. By means of eqs 3a and 3b, eq 2 can be rewritten as

$$F = \frac{\sigma^2(n^+) + \sigma^2(n^-) - 2\,\mathrm{cov}(n^+,n^-)}{\langle n^+ \rangle - \langle n^- \rangle} \quad (4)$$

When the total current is determined by the sum of two independent, totally uncorrelated, opposite currents, the covariance term appearing in eq 3b vanishes and eq 4 reads

$$F^* = \frac{\sigma^2(n^+) + \sigma^2(n^-)}{\langle n^+ \rangle - \langle n^- \rangle} = \frac{\langle n^+ \rangle + \langle n^- \rangle}{\langle n^+ \rangle - \langle n^- \rangle} \quad (5)$$

Thus, the comparison between F and F^* gives information about the degree of correlation in the charge motion. Furthermore, it can be easily proved that

$$\lim_{\Delta t \to \infty} \frac{1}{\Delta t} \langle \delta Q_t^2 \rangle = \frac{1}{2} S_I(0) \quad (6)$$

where Q_t is the estimator of the total charge flowing across the channel mouth in the time interval Δt and $S_I(0)$ is the current noise power spectrum at zero frequency. We observe that $I = Q_t/\Delta t$ and $Q_t = n(t)q$, q being the charge of the moving particle and I being the current. Combining eq 2 with eq 6, one can obtain the well-known expression for the Fano factor:[21]

$$F = \frac{S_I(0)}{2q\langle I \rangle} \quad (7)$$

Methods

The simulated system is built by embedding the KcsA structure from *Streptomices lividans* solved at 2.0 Å (PDB code 1K4C) in a water−octane−water bilayer.[11] The crystallographic structure has been obtained when the channel is in its closed state; we suppose here that the open-state structure does not modify significantly the selectivity filter region, which is under investigation in this study. For an analogous reason—a small influence of the membrane on the energy profile in the selectivity filter—we mimic the membrane environment with a simple octane slab and not with more sophisticated lipids in order to minimize the computational effort.

The system is thus composed of 420 amino acids, 500 octane molecules, 8802 water molecules (including crystallographic waters), 8 potassium ions, and 24 chlorine ions to ensure an electroneutral environment, for a total of 34 434 atoms.

The free-energy profiles have been determined via MD simulations using the GROMACS 3.3 package with the default GROMACS (or modified GROMOS87) force field[23,24] in the NVT ensemble. Even though this force field is somewhat obsolete, it always proved to be satisfactory in correctly keeping the position of the binding sites after the equilibration time and was adopted because of this evidence. Several previous simulations with AMBER6 did not satisfy this preliminary test. Periodic boundary conditions have been applied to the simulation box (69 × 69 × 97 Å), and electrostatic interactions have been calculated with the particle mesh Ewald method.[25] The temperature has been kept constant with the Nosé−Hoover thermostat with a time constant of 5 fs; the Parrinello−Rahman barostat is used to control the pressure. Time steps of 1 fs have been applied.

The X-ray crystallographic structure has initially been fully equilibrated for 1.2 ns, in which the first 200 ps have been spent to increase the temperature from 100 to 300 K. Umbrella sampling free-energy calculations have been performed by applying a biasing harmonic potential (force constant used: 8368 kJ/mol nm^2) to one of the ions involved in the conduction process. Because of the physical constraint of concerted motion inside the selectivity filter, this choice proved to be effective to induce the complete concerted translocation of all of the ions involved. The biasing potential was moved in steps of a half angstrom along the channel axis for a total of 18 space windows per trajectory. Each simulation lasted 500 ps, the first 300 ps being discarded as equilibration time in the presence of the added potential. The WHAM algorithm[26] was finally used to determine the free-energy profile.

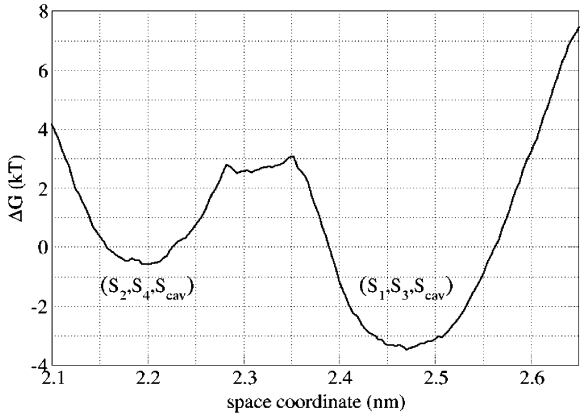


Figure 1. Free-energy profile as a function of the center of mass coordinate of the two top ions for transition (S_2,S_4,S_{cav}) → (S_1,S_3,S_{cav}). The energy barrier is approximately 3.5 k_BT for the forward transition and 6.5 k_BT for the reverse transition.

An example of an equilibrium multi-ion free-energy profile is reported in Figure 1. The 1D curve represents the free energy as a function of the center of mass coordinate of the ions initially in the 2 and 4 binding sites for the transition (S_2,S_4,S_{cav}) ↔ (S_1,S_3,S_{cav}). The energy barrier obtained for this transition is approximately 3.5 k_BT for forward motion and 6.5 k_BT for backward motion, which is consistent with what is reported in the literature[7] for two-ion concerted motion. Similar MD calculations have been performed to estimate the energy barriers affecting each class of translocations. Rates involving entries to or exits from the cavity site S_{cav} have been inferred from corresponding rates affecting analogous events in site S_0, because under open-gate conditions the cavity is directly linked to the "inner bath", as the outer mouth can freely communicate with its corresponding "outer bath".

Finally, a few words must be spent on the 1D umbrella sampling, which may appear to be an oversimplified method to evaluate the free-energy profiles. In the past, calculations were performed using the positions of two or three ions in the selectivity filter as reaction coordinates:[6,7] such calculations however are very demanding in terms of CPU time. From the analysis of the results available in the literature, one may note that in every single step of the conduction cycle only one coordinate varies significantly, the others being close to their initial values. This evidence suggests the possibility of implementing a 1D sampling of the free-energy profile. On the other hand, a good estimate of the ionic current and of the error affecting the energy barriers is of great importance for the appropriate evaluation of the free-

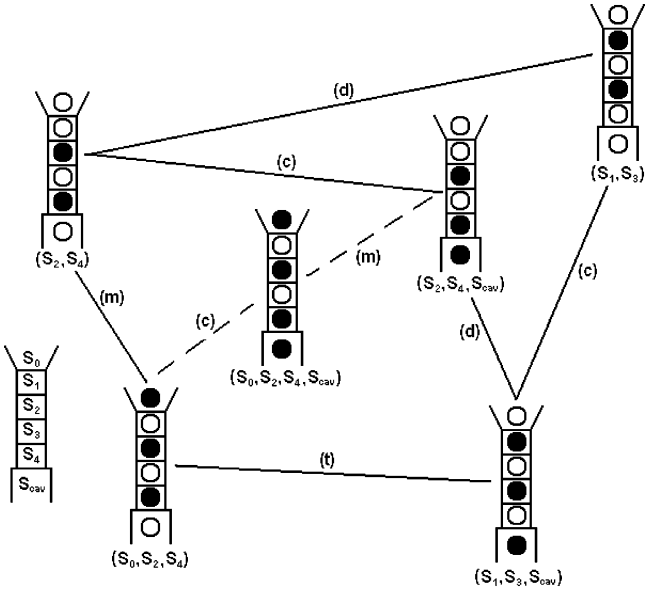


Figure 2. Configurations considered in the model and transitions among them. The sketch on the left is intended to show the six binding sites. Open circles stand for water molecules, solid circles for potassium ions. The dashed lines represent a transition with four ions involved that seldom happens. (m) stands for an ion entry/exit into/from the outer mouth S_0, (c) for an exit/entry from/into the cavity site S_{cav}, (d) for the two-ion concerted motion, and (t) for the three-ion concerted motion. The set of transition rates at zero bias used in the MC simulations is the following: $(S_1,S_3) \rightarrow (S_2,S_4)$ and $(S_1,S_3,S_{cav}) \rightarrow (S_2,S_4,S_{cav})$: 1.08×10^8 s^{-1}; $(S_2,S_4) \rightarrow (S_1,S_3)$ and $(S_2,S_4,S_{cav}) \rightarrow (S_1,S_3,S_{cav})$: 2.14×10^8 s^{-1}; $(S_1,S_3) \rightarrow (S_1,S_3,S_{cav})$, $(S_2,S_4) \rightarrow (S_2,S_4,S_{cav})$, and $(S_0,S_2,S_4) \rightarrow (S_0,S_2,S_4,S_{cav})$: 9.93×10^8 s^{-1}; $(S_1,S_3,S_{cav}) \rightarrow (S_1,S_3)$, $(S_2,S_4,S_{cav}) \rightarrow (S_2,S_4)$, and $(S_0,S_2,S_4,S_{cav}) \rightarrow (S_0,S_2,S_4)$: 2.07×10^8 s^{-1}; $(S_2,S_4) \rightarrow (S_0,S_2,S_4)$ and $(S_2,S_4,S_{cav}) \rightarrow (S_0,S_2,S_4,S_{cav})$: 2.58×10^8 s^{-1}; $(S_0,S_2,S_4) \rightarrow (S_2,S_4)$ and $(S_0,S_2,S_4,S_{cav}) \rightarrow (S_2,S_4,S_{cav})$: 1.73×10^8 s^{-1}; $(S_0,S_2,S_4) \rightarrow (S_1,S_3,S_{cav})$: 8.61×10^8 s^{-1}; $(S_1,S_3,S_{cav}) \rightarrow (S_0,S_2,S_4)$: 1.35×10^8 s^{-1}.

energy profile. More results can be found in the literature[17] and in a further work in preparation.

The conduction process is simulated by generating a sequence of transitions between different channel configurations through a Monte Carlo code. The available configurations and their transition rates at zero bias are reported in Figure 2.

The ion current has been evaluated as the net flux of incoming and outgoing charges at a channel boundary in a defined time interval Δt.

Power spectra have finally been calculated by Fourier transforming the current fluctuations $\delta I(t)$ and averaging results over 1000 independent sequences of 10 000 time steps.

To extract $I(V)$ characteristics starting from an atomistic approach, one must first deal with the problem of the effect of the external electric field on the transition rates. This has been done with the following constraint: after a complete cycle, the system (protein + ions inside) changes its total energy by an amount equal to qV, where V is the external applied potential and q is the electronic charge that crossed the channel during the cycle.

Equation 1 now reads

$$k'_{A\rightarrow B} = A \exp[-\beta(\epsilon_b - q\alpha_b V)] = k_{A\rightarrow B} \exp(\beta q \alpha_b V) \qquad (8)$$

to take into account the effect of the applied potential. Parameters α_b represent the fractions of the total potential applied to the transitions considered and will be further discussed in this section; the subscript b is a label classifying the kind of the considered transition.

The energetics of ion conduction in the presence of an electric field, to our knowledge, have not been deeply developed with a full self-consistent procedure so far. Even if technically possible,[27,28] the presence of charged ions in the aqueous solution and the periodic boundary conditions make it difficult to implement an explicit electric field in the MD simulations. To overcome this problem, we choose to perform all of the MD simulations and the evaluation of the corresponding free-energy profiles without any external electric field. Only after calculating the transition rates at zero applied bias, we did consider the effect of the electric field by introducing a correcting factor, as reported in eq 8. To estimate how much a transition is affected by the applied external electric field, we referred to the milestone work of Bernèche and Roux,[7] where the potential drop through the channel is calculated by means of a Poisson−Boltzmann scheme. It should be noted that this procedure applies rigorously only for the motion of a single ion at a time, that is, when all of the process can be described using one physical coordinate that also coincides with the reaction coordinate. On the contrary, when a multiplicity of coordinates must be taken into account, the procedure fails, and the potential drop should be further tuned to fit the experimental data, after its first evaluation via MD at zero bias. For the sake of truth, this fitting procedure could be avoided if the chosen coordinates *really are* the reaction coordinates, but in practice, this situation never comes at hand if dealing with biological systems.

Results

In our numerical results, we have found that $I(V)$ characteristics are very sensitive to the values assumed by α_b's in eq 8, which express the fraction of the total applied bias associated with each of the four steps reported in the previous section. Barriers used in our MC calculations range from 2.5 to 5.3 k_BT; higher values, as sometimes MD simulations provide, lead to current saturation at voltages much higher than those suggested by experiments. Furthermore, the existence of voltage-independent transition rates dominating the ion flow is suggested above 100 mV to account for current leveling-off, which in turn means that a dominating transition exists and somehow influences the conduction process at lower bias values.

We identify this crucial transition in the three-ion concerted motion $(S_1,S_3,S_{cav}) \leftrightarrow (S_0,S_2,S_4)$: it exhibits the highest barriers at zero bias, and it determines the order of magnitude of the current. To tune simulated data to experiments,[9] we have found that most of the energy variation due to the external potential (about 90%) is affecting this transition (α_b

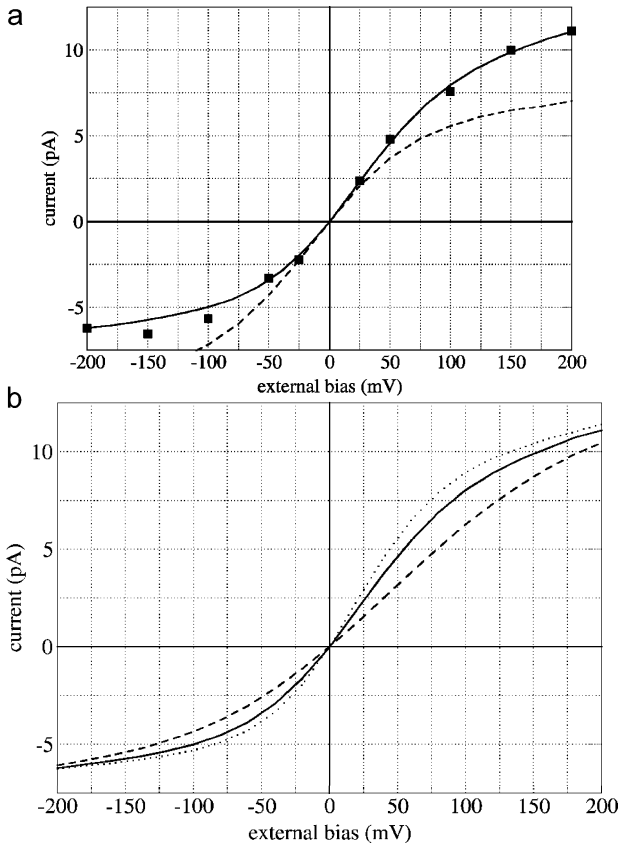


Figure 3. (a) Experimental and simulated $I(V)$ characteristics for K^+ ions flowing across the KcsA potassium channel (full squares refer to experiments[9]); the two curves refer to MC simulations obtained with a different distribution of the internal potential drop (dashed line mainly on the two-ion concerted motion, solid line mainly on the three-ion concerted motion). (b) Effect of the barrier height of the (t) transition on the current. The dashed and dotted lines represent calculations where 1 k_BT has been added to or subtracted from the original barrier height (solid line), respectively. A symmetric potassium concentration of 100 mM is used in all of the cases investigated. Positive bias is meant for a flux from the intracellular to the extracellular reservoir.

$= 0.9$), the residual energy being distributed $^1/_3$ to the transition to and from cavity S_{cav} ($\alpha_b = 0.033$) and $^2/_3$ to the transition from and to the outer mouth S_0 ($\alpha_b = 0.067$). The two-ion motion's rate is substantially unaltered. A similar qualitative behavior can be obtained considering the two-ion motion as the crucial transition and leaving unaltered the three-ion's one, but results are less satisfactory, giving origin to a worse representation of experimental data (see also Figure 3). This framework appears to be consistent with the one reported in the literature,[7] where the transmembrane potential drops mainly in the selectivity filter and approximately $^1/_{10}$ elsewhere. The resulting $I(V)$ characteristics for potassium atoms are reported in Figure 3a together with available experimental results.[9]

To investigate how the results reported in Figure 3a are sensible to the values assumed by the barrier heights, we performed two simulations adding or subtracting 1 k_BT to the previously determined height of the (t) transition, which is supposed to mostly affect the permeation process, keeping

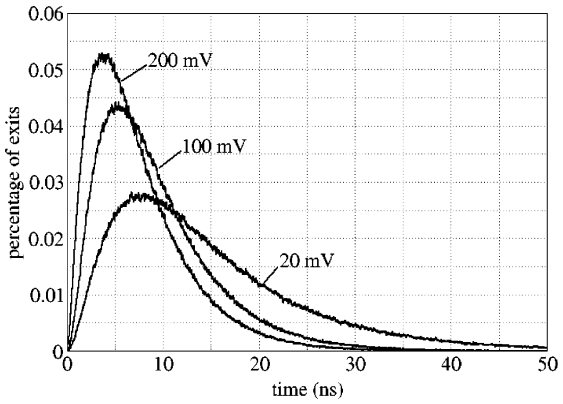


Figure 4. Number of ion exits as a function of the time interval between two successive ion exits from the simulation. Mean time and most frequent time are 14 ns and 7.8 ns at 20 mV, 9.3 ns and 5.3 ns at 100 mV, and 7.4 ns and 4.1 ns at 200 mV, respectively.

the prefactor A in eq 8 fixed. This value is approximately the minimum standard deviation associated with the estimates from MD simulations. Results are show in Figure 3b. It is seen that the three curves in the figure are qualitatively similar and converge to values close to each other at the highest (positive or negative) fields. This is so because the electrostatic energy contribution coming from the external bias gains more importance in the determination of the exponent in eq 8 under the above-mentioned conditions. Moreover, the considered barrier is significantly lowered by the effect of the field, and the barriers associated with other transitions still play an important role, as discussed above. On the contrary, at low and intermediate biases, the difference in the conductance for the three cases is relevant, and the agreement with experiments is lost, especially when the barrier height is increased. It should be observed, however, that the prefactor and the exponent in eq 8 are equally important in the determination of the numerical value of the rate itself.

From the computational point of view, shot noise can be analyzed even at frequencies corresponding to the characteristic microscopic time of single shots, provided that the current sampling time is properly chosen. To estimate this characteristic time for the process under investigation, we have evaluated the number of ion exits as a function of the time interval elapsed since the previous exits. The obtained distribution exhibits a maximum, and the length of its tail depends on the external bias, as shown in Figure 4. For a bias of 100 mV, the peak of the curve is obtained at a time of 5.3 ns, while the average time $\langle T_{exit} \rangle$ between two successive ion exits is estimated approximately as 9.3 ns.

Figure 5 shows the power spectrum of current fluctuations normalized to $2q\langle I \rangle$ as a function of frequency from simulations which use different values of the current sampling time Δt. It can be noted that the low- and high-frequency ranges are not affected by the choice of Δt. On the contrary, at frequencies on the order of $1/\langle T_{exit} \rangle$, the shape of the spectrum strongly depends on the choice of Δt, as it should happen also in experimental measurements, because of the limitations imposed in the frequency range of the Fourier transform.

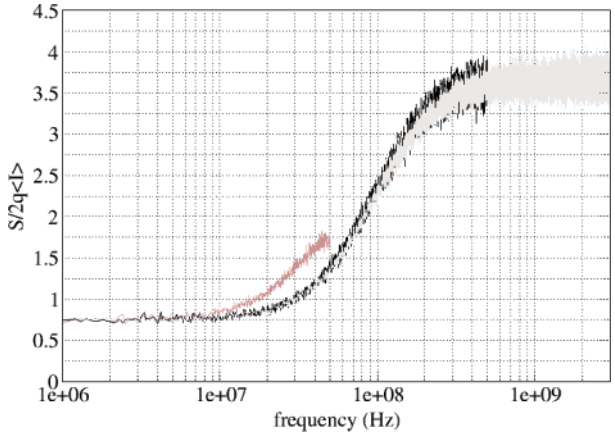


Figure 5. Power spectrum of the current fluctuations as a function of frequency as obtained from the simulations using different sampling times for the current signal. The dark line is obtained with a sampling time of $\Delta t = 10^{-9}$ s, the half-tone gray line with a sampling time of $\Delta t = 10^{-8}$ s, and the pale gray line with a sampling time of $\Delta t = 10^{-10}$ s. The three results have been obtained at the same bias of 100 mV.

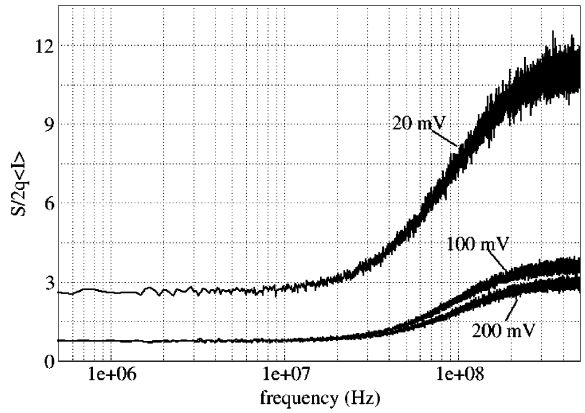


Figure 6. Power spectrum of the current fluctuations as a function of frequency obtained with different values of the external bias, indicated in the figure. All of the results refer to the same potassium concentration of 100 mM.

Examples of calculated noise power spectra as functions of frequency at different external biases are shown in Figure 6.

White noise is found until about 2×10^7 Hz; at frequencies on the order of $1/\langle T_{exit} \rangle$, an increase of the spectrum is observed, up to the highest considered frequencies, where single charge output is seen.

The asymptotic value of $S_I(\omega)$ in the low-frequency range can be exploited to calculate the system conductance G without application of any transmembrane potential by means of the Nyquist formula:[29]

$$S_I(\omega) = 4k_B T\, Re[1/Z(\omega)] \tag{9}$$

where $Z(\omega)$ is the system impedance. In the considered low-frequency range, the dispersion can be neglected, and $1/Z(\omega) \cong G$. Equation 9 can be reformulated as follows:

$$G = S_I(0)/4k_B T \tag{10}$$

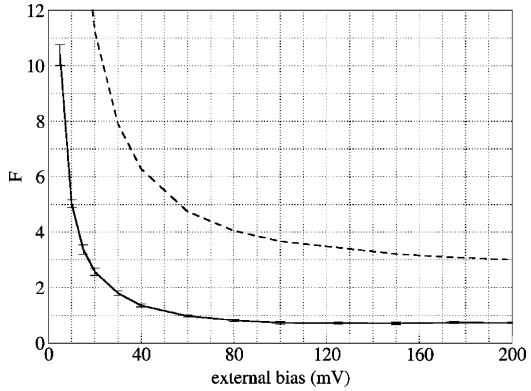


Figure 7. Fano factor as a function of the external bias for the power spectrum shown in Figure 5. The dashed line refers to a process with two totally uncorrelated opposite currents, as from eq 5 (see text).

$S_I(0)$ being the limit of $S_I(\omega)$ in the low-frequency region. For this purpose, the average value in the range 50−500 kHz was considered (lower frequencies do not introduce significant differences in the evaluation of the average). For the zero-bias condition, we obtain a conductance $G = 91 \pm 2$ pS; the corresponding value that can be inferred from experimental data reported in Figure 3 is 96 pS. The agreement between these two data further confirms the consistency of the calculations.

Finally, in Figure 7 the Fano factor obtained from eq 4 is plotted as a function of the external bias. The same result is obtained from both positive and negative biases, and for this reason, negative biases are not reported in the figure. F rapidly decreases at increasing bias until about 100 mV, where an asymptotic value of 0.73 is reached. To estimate the effect of correlations in the ions' motion on the noise spectrum, we reported in the same figure the Fano factor F^* calculated from the MC simulation with eq 5, assuming that the two opposite currents (with net flux equal to I) are totally uncorrelated.

The two curves are qualitatively similar, and their ratio ranges from a factor 4 to a factor 5 over the range of the considered biases. The comparison between F and F^* allows us to state that the increase of the Fano factor at low biases is mainly because of the numbers of ions' crossings in the two directions tend to become equal, and exits immediately followed by a re-entrance are very frequent. The presence of a significant external bias imposes the dominance of one of the two fluxes over the other. Moreover, the numerical differences between F and F^* confirm the existence of a significant correlation in the ions' motion due to the peculiar permeation paths affecting the conduction process.

Conclusions

A numerical procedure for the evaluation of the conduction and noise associated with ion flow across membrane channels has been presented. The occupancy configurations of the channel, the transition rates between different configurations, and the parameters contained in the transition-rate formula have been evaluated through MD simulations and from a comparison with experimental data. Numerical results obtained from a MC procedure for the case of potassium ions

permeating the KcsA channel provide $I(V)$ characteristics comparable with experimental data, once the drop of the external field is accurately modeled. The current noise power spectrum exhibits a structure suggesting significant correlation between successive ion exits. This computational analysis shows that noise measurements in single open channels, today achievable with modern equipment, can contribute to the comprehension of the permeation mechanisms inside the channel.

Acknowledgment. We thank A. Bigiani, J. Kona, C. Miller, M. Rosini, and F. Sigworth for useful discussions.

References

(1) Traynelis, S. F.; Jaramillo, F. Getting the Most out of Noise in the Central Nervous System. *TINS* **1998**, *21*, 137−145.

(2) Miller, C. Private communication.

(3) Heinemann, S. H.; Sigworth, F. J. Open Channel Noise. V. Fluctuating Barriers to Ion Entry in Gramicidin A Channels. *Biophys. J.* **1990**, *57*, 499−514.

(4) Krasilnikov, O. V.; Bezrukov, S. M. Polymer Partitioning from Nonideal Solutions into Protein Voids. *Macromolecules* **2004**, *37*, 2650−2657.

(5) Zhou, Y.; Morais-Cabral, J.; Kaufman, K.; MacKinnon, R. Chemistry of Ion Coordination and Hydration Revealed by a K+ Channel−Fab Complex at 2.0 A. *Nature* **2001b**, *414*, 43−48.

(6) Roux, B.; Allen T.; Bernèche S.; Im W. Theoretical and Computational Models of Biological Ion Channels. *Q. Rev. Biophys.* **2004**, *37*, 15−103.

(7) Bernèche, S.; Roux, B. A Microscopic View of Ion Conduction through the K^+ Channel. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 8644−8648.

(8) Kutluay, E.; Roux, B.; Heginbotham, L. Rapid Intracellular TEA Block of the KcsA Potassium Channel. *Biophys. J.* **2005**, *68*, 1018−1029.

(9) LeMasurier, M.; Heginbotham, L.; Miller, C. KcsA: It's a Potassium Channel. *J. Gen. Physiol.* **2001**, *118*, 303−313.

(10) Morais-Cabral, J.; Zhou, J. H.; MacKinnon, R. Energetic Optimization of Ion Conduction Rate by the K^+ Selectivity Filter. *Nature* **2001**, *414*, 37−42.

(11) Compoint, M.; Carloni, P.; Ramseyer, C.; Girardet, C. Molecular Dynamics Study of the KcsA Channel at 2.0-Å Resolution: Stability and Concerted Motions within the Pore. *Biochim. Biophys. Acta* **2004**, *1661*, 26−39.

(12) Roux, B. Ion Conduction and Selectivity in K^+ Channels. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 153−171.

(13) Souaille, M.; Roux, B. Extension to the Weighted Histogram Analysis Method: Combining Umbrella Sampling with Free Energy Calculations. *Comput. Phys. Commun.* **2001**, *35*, 40−57.

(14) Hummer, G.; Szabo, A. Free Energy Reconstruction from Nonequilibrium Single-Molecule Pulling Experiments. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 3658−3661.

(15) Hummer, G.; Szabo, A. Free Energy Surfaces from Single-Molecule Force Spectroscopy. *Acc. Chem. Res.* **2005**, *38*, 504−513.

(16) Laio, A.; Parrinello, M. Escaping Free Energy Minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562−12566.

(17) Piccinini, E.; Ceccarelli, M.; Affinito, F.; Brunetti, R.; Jacoboni, C. Exploring Free-Energy Profiles through Ion Channels: Comparison on a Test Case. *J. Comput. Electron*. **2006**, to be published

(18) Hanggi, P.; Talkner, P.; Borkovec, M. Reaction-Rate Theory: Fifty Years after Kramers. *Rev. Mod. Phys*. **1990**, *62*, 251−341.

(19) Siwy, Z.; Ausloos, M.; Ivanova, K. Correlation Studies of Open and Closed State Flluctuatons in an Ion Channel: Analysis of Ion Current through a Large-Conductance Locust Potassium Channel. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2002**, *65*, 0319071−0319075.

(20) Frehland, E.; Stephan, W. Theory of Single-File Noise. *Biochim. Biophys. Acta* **1979**, *553*, 326−341.

(21) Schottky, W. Über Spontane Stromschwankungen in Versichiedenen Eletrikzitätsleitern. *Ann. Phys. (Leipzig)* **1917**, *57*, 541−567.

(22) Fano, U. Ionization Yield of Radiation. II. The Fluctuations of the Number of Ions. *Phys. Rev.* **1947**, *72*, 26−29.

(23) Lindhal, 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.

(24) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation. *Comput. Phys. Commun.* **1995**, *91*, 43−56.

(25) Essman, U.; Perela, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577−8592.

(26) Kumar, S.; Bouzida, D.; Swedesen, R. H.; Kollman, P. A.; Rosenberg, J. M. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13*, 1011−1021.

(27) Yeh, I.; Berkowitz, M. L. Dielectric Constant of Water at High Electric Fields: Molecular Dynamics Study. *J. Chem. Phys.* **1999**, *110*, 7935−7942.

(28) Yeh, I.; Hummer, G. Nucleic Acid Transport through Carbon Nanotube Membranes. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12177−12182.

(29) Nyquist, H. Thermal Agitation of Electric Charge in Conductors. *Phys. Rev.* **1928**, *32*, 110−113.

CT6002077

JCTC Journal of Chemical Theory and Computation

Linear Interaction Energy (LIE) Models for Ligand Binding in Implicit Solvent: Theory and Application to the Binding of NNRTIs to HIV-1 Reverse Transcriptase

Yang Su,[†] Emilio Gallicchio,*[,†] Kalyan Das,[‡] Eddy Arnold,[‡] and Ronald M. Levy*[,†]

BioMaPS Institute of Quantitative Biology, Department of Chemistry and Chemical Biology, and Center for Advanced Biotechnology and Medicine, Rutgers University, Piscataway, New Jersey 08854

Received August 8, 2006

Abstract: Expressions for Linear Interaction Energy (LIE) estimators for the binding of ligands to a protein receptor in implicit solvent are derived based on linear response theory and the cumulant expansion expression for the free energy. Using physical arguments, values of the LIE linear response proportionality coefficients are predicted for the explicit and implicit solvent electrostatic and van der Waals terms. Motivated by the fact that the receptor and solution media may respond differently to the introduction of the ligand, a novel form of the LIE regression equation is proposed to model independently the processes of insertion of the ligand in the receptor and in solution. We apply these models to the problem of estimating the binding free energy of two non-nucleoside classes of inhibitors of HIV-1 RT (HEPT and TIBO analogues). We develop novel regression models with greater predictive ability than more standard LIE formulations. The values of the regression coefficients generally conform to linear response predictions, and we use this fact to develop a LIE regression equation with only one adjustable parameter (excluding the intercept parameter) which is superior to the other models we tested and to previous results in terms of predictive accuracy for the HEPT and TIBO compounds individually. The new models indicate that, due to the different effects of induced steric strain of the receptor, an increase of ligand size alone opposes binding for ligands of the HEPT class, whereas it favors binding for ligands of the TIBO class.

1. Introduction

The binding free energy of a ligand to a receptor is given by the difference of the free energies of inserting the ligand in the receptor and in solution. In principle the free energy for each process can be calculated exactly for a given force field using the free energy perturbation (FEP) or thermodynamic integration (TI) methods. In practice, however, the complexities involved in setting up suitable mutation paths and the long simulation times needed to reach convergence have limited the applicability of FEP and TI methods to the investigation of the variations of the binding free energy for small ligand modifications in the final stages of lead optimization.[1] Linear Interaction Energy (LIE) models[2,3] offer attractive approximate alternatives to the full FEP methodology because they require only the computation of average interaction energies at the end points of the mutation.

LIE models can be described as empirical Quantitative Structure−Activity Relationships (QSAR) which employ physically motivated energetic estimators. As opposed to methods that predict the binding free energy on the basis of the structure of the ligand alone, LIE estimators also reflect properties of the ligand−receptor complex. LIE methods are expected to perform better than methods based on the ligand

* Corresponding author e-mail: emilio@biomaps.rutgers.edu (E.G.) and ronlevy@biomaps.rutgers.edu (R.M.L.).

[†] BioMaPS Institute of Quantitative Biology and Department of Chemistry and Chemical Biology.

[‡] Center for Advanced Biotechnology and Medicine and Department of Chemistry and Chemical Biology.

10.1021/ct600258e CCC: $37.00
Published on Web 11/10/2006

alone because they are more intimately related to the structure of the complex. One of the most popular LIE formulations employs the following regression expression for the binding free energy ΔF_b[3,4]

$$\Delta F_b = \alpha \Delta \bar{V}_{vdw} + \beta \Delta \bar{V}_{el} + \gamma \Delta \bar{A} + \delta \quad (1)$$

where $\Delta \bar{V}_{vdw}$, $\Delta \bar{V}_{el}$, and $\Delta \bar{A}$ are differences between quantities measured for the ligand complexed with the receptor and the ligand free in solution. $\bar{V}_{vdw}$ is the average van der Waals interaction energy between the ligand and its environment (the solvent or the receptor and the surrounding solvent). Similarly, $\bar{V}_{el}$ is the average electrostatic interaction energy between the ligand and its environment. $\bar{A}$ is the average solvent accessible surface area of the ligand. These quantities are typically obtained from Molecular Dynamics (MD) or Monte Carlo (MC) simulations started from known or modeled conformations of the ligand and the ligand−receptor complex. α, β, γ, and δ are empirical adjustable parameters whose values are obtained by fitting the model over a training set of ligands of known binding affinity. A trained LIE model can then be used to predict the binding free energy of ligands of unknown affinity provided that the LIE estimators for these ligands can be reliably calculated. Each LIE estimator in eq 1 reflects physical forces that affect binding. $\Delta \bar{V}_{vdw}$ and $\Delta \bar{V}_{el}$ measure the balance between the desolvation penalty caused by the loss of ligand−solvent interactions and the gain of ligand−receptor interactions which favor binding, whereas the surface area estimator measures the hydrophobic driving force toward complexation. These LIE estimators do not take into account explicitly thermodynamic forces related to the receptor that affect the binding affinity, such as the desolvation of receptor atoms and reorganization free energy of the receptor for accommodating the ligand. Nevertheless it can be shown (see below) that, under the assumption of linear response, these effects are, in fact, included in the model and are encoded in the values of the LIE regression coefficients.

LIE models have their origin in physical theories of solvation based on the linear response approximation to the free energy,[5−7] applied to the problem of binding free energy estimation.[2,8−11] The introduction of the ligand in either the solution or receptor environments can be regarded as a perturbation applied to the system. If the system responds perfectly linearly (as formally defined below) to the perturbation, the free energy of introducing the ligand can be shown to be exactly proportional to the interaction energy between the ligand and its environment. Given that the ligand constitutes a large perturbation to the system, it is unlikely that linear response applies to the entire processes of introducing the ligand in solution and in the receptor. The LIE regression equation (eq 1) assumes instead that linear response applies to the individual processes of introducing hydrophobic, van der Waals, and electrostatic interactions, albeit with different linear response proportionality coefficients. Even so, nonlinearities in practice limit the applicability of LIE models to within a related class of ligands. This is reflected in the fact that in practice LIE relationships are used for estimating *relative* binding free energies of similar ligands rather than absolute binding free energies. The accuracy of a LIE model therefore hinges on whether the perturbation corresponding to the mutation of each ligand into another is small enough so that linear response applies to the relative binding free energy. The effect of absolute binding free energies is collectively absorbed by the intercept parameter δ; deviations from linear response are expected to be reflected in the limited range of applicability of a LIE model.

In this paper we investigate a series of outstanding issues with regard to LIE models and their applications to ligand binding in structural biology. The first question is to what extent linear response applies to a given ligand set and the consequences of deviations from linear response in terms of the accuracy of the LIE model. Although in principle addressing this question requires comparing LIE predictions with relative free energies evaluated using the rigorous FEP and TI methods, we take a first step in this direction by comparing the values of the LIE adjustable parameters obtained by fitting training sets of ligand binding data with those expected based on linear response theory. The second question concerns the form of the LIE regression equation. Equation 1 assumes that the response of the solution and receptor environments, as measured by the LIE coefficients α, β, and γ, is the same. To our knowledge this assumption is ubiquitous in LIE applications. We develop and validate an alternative formulation in which the processes of insertion of the ligand in the solution and in the receptor are decoupled so that each is allowed to have different linear response proportionality coefficients. Finally, based on linear response techniques we analyze the appropriate form of the LIE estimators when the solvent is treated implicitly. We show that the form for the electrostatic LIE estimator we derive based on linear response theory agrees with the corresponding estimator recently proposed by Carlsson et al.[12] and differs from the more empirical expression proposed earlier by Zhou et al.[4] We develop, following the linear response formalism, electrostatic, van der Waals, and cavity implicit solvent estimators that best represent the corresponding LIE estimators in explicit solvent and show that these lead to improved accuracy.

We apply these ideas to ligand−protein complexes that have been studied previously using the LIE method: the binding of the HEPT and TIBO classes of Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTI) to the HIV-1 Reverse Transcriptase (RT) enzyme.[13] HIV-1 RT is essential for the life cycle of the virus. It converts the single stranded genomic RNA into double stranded DNA which is subsequently integrated into the host chromosome and passed on to all progeny cells.[14] Computer-aided structure-based drug discovery technologies have made a significant contribution to the development of medicinally active NNRTI anti-AIDS compounds.[15−17] Three of these compounds, dapivirine, etravirine, and rilpivirine, are currently in clinical trials. The goal is the design of NNRTIs of greater potency and resilience with respect to common drug-resistance mutations.[18] However the mode of binding of NNRTIs, which includes extensive receptor conformational reorganization and mainly nonspecific hydrophobic ligand−receptor contacts, does not offer obvious chemical modification leads for

binding free energy optimization. In this context, LIE models have been applied to the estimation of the binding affinities of NNRTI inhibitors with encouraging results.[19,20]

The renewed interest of our laboratories in the LIE modeling of HIV-1 RT inhibitors is motivated in part by recent progress in obtaining a crystal structure of HIV-1 RT in complex with an inhibitor of the N-acyl hydrazone class.[21] The crystal structure identifies a novel non-nucleoside binding site adjacent to but distinct from the NNRTI binding site. The NAH inhibitors targeting this site are expected to suffer from little or no cross-resistance from existing drug resistance mutations, thus providing new options for novel therapeutic strategies in the treatment of AIDS. This crystal structure provides the initial framework for the application of LIE methodologies for lead optimization and the development of a new class of inhibitors of HIV-1 RT. The work presented in this paper on NNRTIs provides the theoretical and computational basis for the application of LIE modeling in implicit solvent for binding free energy estimation of the new NAH class of HIV-1 RT inhibitors which is currently being investigated in our lab.

In the following section we review the theoretical foundations of the LIE method and discuss some of the approximations. We then derive a LIE formalism appropriate for situations where the solvent is modeled implicitly. We propose novel LIE regression equations which emerge naturally from the statistical mechanics derivation. We then apply these regression equations to analyze the binding of a series of 20 HEPT inhibitors and 37 TIBO inhibitors of HIV-1 RT. The models are validated using jack-knife prediction tests. The predictive ability of different forms and parametrizations of the LIE equations are compared. We compare features of the HEPT and TIBO binding modes and the corresponding LIE estimators. We conclude the paper with a discussion of the accuracy and physical interpretation of LIE models.

2. Theory and Methods

2.1. Linear Response Approximation. The idea of adopting a linear response approximation expression for binding free energy estimation was first stated by Lee et al.,[8] who suggested the use of a "two-point" Linear Response Approximation (LRA) estimation formula previously derived for electrostatic solvation;[6] see, for example, discussion in ref 10. This formula was later simplified by Åqvist et al.[2] who introduced the Linear Interaction Energy (LIE) model which, unlike the more accurate "two-point" LRA formula,[22] does not require the evaluation of estimators at more than one state of the ligand. The LIE method was used by Åqvist and collaborators to estimate relative binding affinities of endothiapepsin and HIV-protease inhibitors.[2,9] We review here some basic concepts related to the linear response approximation and derive a linear response expression—eq 12 below—for the LIE proportionality coefficients, which will be used in the following to interpret the values of the LIE fitting coefficients obtained from the analysis of experimental binding affinities.

The insertion of the ligand into either the receptor or the solvent can be thought of as turning on, by means of a charging parameter λ, the interactions of the ligand with the surrounding environment (the receptor and/or the solvent). It is convenient to think about these processes in stages. First the ligand cavity is formed, then the van der Waals interactions between the ligand and the environment are turned on, and finally ligand-environment electrostatic interactions are established. At each stage the λ-dependent potential energy of the system is

$$U(x;\lambda) = U_0(x) + \lambda V(x) \tag{2}$$

where V represents the ligand-environment interactions which are being added, and U_0, the reference potential energy, contains receptor−receptor, solvent−solvent, and ligand−intramolecular interactions as well as the ligand−environment interactions established in the previous stages. Starting from the expression of the configurational partition function

$$Z(\lambda) = e^{-\mathrm{F}(\lambda)/\mathrm{kT}} = \int dx\, e^{-\lambda \mathrm{V(x)/kT}} e^{-\mathrm{U_0(x)/kT}} \tag{3}$$

it is straightforward to show that the first and second derivatives of the free energy $F(\lambda) = -kT \ln Z(\lambda)$ with respect to λ correspond, respectively, to the first and second moments of the probability distribution of V:

$$\frac{\partial F}{\partial \lambda} = \langle V \rangle_\lambda \tag{4}$$

$$\frac{\partial^2 F}{\partial \lambda^2} = -\langle (\delta V)^2 \rangle_\lambda / kT \tag{5}$$

If the second moment is approximately constant along the thermodynamic path from $\lambda = 0$ to $\lambda = 1$, the third and higher order derivatives of F can be neglected, and it is possible to express $F(\lambda)$ as a Taylor series (known as the cumulant expansion of the free energy) centered at $\lambda = 0$ and truncated at the second order

$$F(\lambda) - F(0) = \langle V \rangle_0 \lambda - \frac{c}{2}\lambda^2 \tag{6}$$

where

$$c = \langle (\delta V)^2 \rangle / kT \tag{7}$$

is assumed constant for $0 \le \lambda \le 1$. It is of interest to note that when the fluctuations of the interaction potential V are Gaussian distributed this assumption is verified and eq 6 is exact.[6] According to eq 6 and under these assumptions, the free energy is quadratic with respect to the charging parameter. This is a manifestation of linear response behavior defined as when the solute-environment average interaction energy $\langle V \rangle_\lambda$ is linearly related to the charging parameter. Indeed, using eqs 4 and 5, and the same assumptions that have lead to eq 6, we obtain

$$\langle V \rangle_\lambda = \langle V \rangle_0 - c\lambda \tag{8}$$

which confirms that $\langle V \rangle_\lambda$ is linearly related to λ.

By evaluating eqs 6 and 8 at $\lambda = 1$ we obtain the free energy change and average interaction energy for adding solute-environment interactions under the assumption of linear response

$$\Delta F = F(1) - F(0) = \langle V \rangle_0 - \frac{c}{2} \tag{9}$$

and

$$\langle V \rangle_1 = \langle V \rangle_0 - c \tag{10}$$

Linear interaction energy models are based on the assumption that ΔF is proportional to $\langle V \rangle_1$:[2]

$$\Delta F = \alpha \langle V \rangle_1 \tag{11}$$

It is therefore of interest to compute, under the assumption of linear response, the proportionality coefficient α given by the ratio of ΔF to $\langle V \rangle_1$. From eqs 9 and 10

$$\frac{\Delta F}{\langle V \rangle_1} = \frac{\langle V \rangle_0 - c/2}{\langle V \rangle_0 - c} \tag{12}$$

The limiting values for this ratio are

$$\frac{\Delta F}{\langle V \rangle_1} = \begin{bmatrix} 1/2 \text{ if } |\langle V \rangle_0| \ll c \\ 1 \text{ if } |\langle V \rangle_0| \gg c \end{bmatrix} \tag{13}$$

Thus, under the assumption of linear response, the limiting values of the ratios between the free energy of adding solute-environment interactions and the corresponding average interaction energy are 1/2 and 1. The free energy change is half the interaction energy when the fluctuation of V, measured by $c = \langle (\delta V)^2 \rangle / kT$, is much larger than $\langle V \rangle_0$, the mean ligand-environment interaction energy calculated within the ensemble of conformations obtained in the absence of ligand-environment interactions. In the opposite limit at which the fluctuations of V are much smaller than $\langle V \rangle_0$, eq 12 gives $\Delta F / \langle V \rangle_1 = 1$, that is the free energy change is equal to the average solute-environment interaction energy.

In the following we apply linear response to the problem of binding free energy estimation and use eq 12 and physical arguments to derive values of the linear response coefficients. We will first review the derivation in explicit solvent and then examine the case in which the solvent is treated implicitly.

2.2. LIE Models in Explicit Solvent. *2.2.1. Hydration Free Energy − Explicit Solvent.* To successfully apply linear response ideas to the hydration free energy estimation, the process of hydration is described as occurring in stages. First the solute cavity is formed in the solvent, then solute−solvent van der Waals interactions are turned on, and finally solute−solvent electrostatic interactions are established. The process of cavity formation is dominated by excluded volume effects and solvent reorganization and does not conform well to the linear response formalism. When using a hard-sphere cavity interaction potential, the solute−solvent interaction energy is zero when the solute and the solvent do not overlap, and it is infinite when overlaps occur. In this limit the average solute−solvent interaction energy is identically zero because conformations in which solute cavity-solvent overlaps exist do not appear in the ensemble. Computational studies of cavity formation have generally been conducted using a continuous but sharp solute−solvent repulsive interaction potential.[23] In these cases, however, due to strong nonlinearities in the response of the solvent to the introduction of the solute cavity, the cavity hydration free energy is poorly correlated with the average repulsive cavity interaction potential. For cavity hydration free energy estimation a term proportional to the solute surface has been shown in some cases to be a reasonably good estimator for the free energy of cavity formation in water.[24,25] The proportionality coefficient γ between the cavity hydration free energy and the surface area can be interpreted as a surface tension coefficient. Indeed molecular simulations have obtained values of this proportionality coefficient similar to the value of the experimental air−water surface tension coefficient.[23,26]

Assuming linear response for the processes of introducing van der Waals and electrostatic solute−solvent interactions, the free energy change at each stage is proportional to the appropriate solute−solvent interaction energy (the solute−solvent van der Waals interaction energy and the solute−solvent electrostatic interaction energy, respectively) averaged in the state corresponding to the end of each stage (the uncharged solute and the fully interacting solute, respectively). The linear response coefficients are given by eq 12 and are, in general, different for each stage. Finally, making the approximation that all averages (denoted by $\langle \cdots \rangle_w$) are calculated when the solute fully interacts with the solvent, a LIE model for the hydration free energy is obtained[27]

$$\Delta F_h \simeq \alpha \langle V_{vdw} \rangle_w + \beta \langle V_{el} \rangle_w + \gamma \langle A \rangle_w \tag{14}$$

where α and β are the linear response proportionality coefficients for the van der Waals and electrostatic stages of the hydration process, V_{vdw} and V_{el} are the solute−solvent van der Waals and electrostatic interaction energies, respectively, γ is an empirical surface tension coefficient, and A is the solvent accessible surface area of the solute. As we show below, the form of the cavity hydration term can be justified in terms of linear response when the solvent is modeled implicitly.

Regression equations based on eq 14 have been parametrized against known experimental hydration free energies of small molecules.[27] However, linear response theory provides a way to estimate some of these parameters from first principles. It has been observed that the charging free energy of ionic and polar solutes in water is approximately proportional to the average solute−solvent electrostatic interaction energy with a linear response proportionality coefficient β of 1/2. According to eq 13 this occurs when the solvent reaction field in the absence of solute charges is much smaller than the fluctuations of the solvent reaction field. These conditions have been indeed verified by numerical studies, confirming that in general water behaves as a good linear dielectric medium.[6,28,29] These observations have constituted the basis for electrostatic linear response free energy models for solutions[2,6,7] as well as for the success of continuum dielectric models of water.[28,30−35]

It is well-known that the process of adding solute−water van der Waals interactions has different characteristics than the process of adding electrostatic interactions. It has been shown that the free energy change for adding attractive van der Waals interactions can be well approximated by the average solute−water van der Waals interaction energy.[23,36,37]

This implies that in this case $\Delta F/\langle V_{vdw}\rangle_1 \simeq 1$, which, under the assumption of linear response occurs (see eq 13) when the mean solute−solvent attractive van der Waals interaction energy in the absence of solute−solvent van der Waals interactions, is larger than the variance of the same quantity divided by kT. This is due to the fact that, although the solvent responds linearly to the solute perturbation, this response is smaller than the attractive van der Waals solvent field that exists at the solute location in the absence of solute−solvent van der Waals interactions. Contrary to dipolar fields in water that tend to cancel each other in the absence of a polarization source, van der Waals interactions are always additive. Based on this analysis we conclude that the α coefficient in the LIE regression equation should assume a value near 1. Explicit solvent simulations have generally confirmed this theoretical prediction.[23] Carlson and Jorgensen[27] have instead reported that a value of α significantly smaller than 1 is obtained by fitting the LIE equation to experimental hydration free energies of small molecules. We believe that the small value of α obtained by Carlson and Jorgensen is caused by compensation between the van der Waals and surface area fitting coefficients. The correct relative magnitude of these two coefficients is difficult to pinpoint because they correspond to highly correlated descriptors (the van der Waals solute−solvent interaction energy and the solute surface area). Indeed both the α and γ coefficients obtained by Carlson and Jorgensen are smaller than well established theoretical predictions.[24,36,38] We have reanalyzed the data from Tables 2 and 3 of ref 27. By setting $\alpha = 1$ and allowing for a nonzero intercept (to fit relative hydration free energies rather than the absolute ones) we achieved a nearly equivalent fit to the experimental hydration free energies. Specifically, using the Coulombic+van der Waals+solvent-accessible surface area model of Carlson and Jorgensen we reproduce the parameters reported previously,[27] $\alpha = 0.49$, $\beta = 0.42$, and $\gamma = 20$ cal/mol Å^2, with a cross-validated correlation of $R_{pred}^2 = 0.83$, whereas our model with α set to 1 gives $\beta = 0.49$ and $\gamma = 62$ cal/mol Å^2 with $R_{pred}^2 = 0.81$. The RMSD from the experimental free energies of hydration of the two models are also similar, 0.88 and 0.98 kcal/mol, respectively. Furthermore, the value of γ (62 cal/mol Å^2) we obtained from the analysis of the data of Carlson and Jorgensen is closer to the experimental value of the macroscopic vacuum-water surface tension (104 cal/mol Å^2)[39] and is similar to the value of the microscopic surface tension parameter obtained from explicit solvent estimates of the work of cavity formation in water (73 cal/mol Å^2).[23] These observations indicate that the simulation data obtained by Carlson and Jorgensen is consistent with the linear response behavior for these solutes.

In conclusion, this analysis shows that eq 14 should provide a good approximation of hydration free energies with the following choice of LIE coefficients: $\alpha \simeq 1$, $\beta \simeq 1/2$, and $\gamma \simeq 73$ cal/mol Å^2, the previously reported explicit solvent estimate.[23]

2.2.2. Binding Free Energy Estimation − Explicit Solvent. The binding free energy ΔF_b of a ligand to a receptor is taken as the difference of the work ΔF_c of creating the ligand in the receptor and the work ΔF_h of creating the ligand in solution. For the process of creating the ligand in the receptor an expression similar to eq 14 has been proposed[3,9] where the interaction energies include ligand−water interactions as well as ligand−receptor interactions, and A is taken as the solvent accessible surface area of the ligand in the receptor−ligand complex. By taking the difference between the LIE estimates of the insertion free energies in the receptor and in solution and assuming that the same values of the LIE coefficients α, β, and γ are appropriate for both, the following binding free energy LIE model is obtained

$$\Delta F_b = \Delta F_c - \Delta F_h \simeq \alpha(\langle V_{vdw}^c\rangle_c - \langle V_{vdw}^f\rangle_w) + \beta(\langle V_{el}^c\rangle_c - \langle V_{el}^f\rangle_w) + \gamma(\langle A\rangle_c - \langle A\rangle_w) \quad (15)$$

where V^c represents an interaction energy between the ligand and the environment (receptor and solvent), and V^f is the corresponding ligand−solvent interaction energy in the absence of the receptor (free ligand), A is the solvent accessible surface area of the ligand, $\langle\cdots\rangle_c$ represents an ensemble average with the ligand in the receptor pocket, and $\langle\cdots\rangle_w$ represents the corresponding average in solution.

Equation 15 and its variations are widely used in binding free energy prediction applications.[40−42] In these studies the LIE coefficients α, β, and γ are obtained by fitting eq 15 to known experimental binding free energies. In principle, linear response theory arguments could be applied, as for the case of hydration free energy estimation, to gain insights into the expected values of these coefficients. It is less clear, however, to what extent the receptor/solvent environment can be considered an ideal linear dielectric medium[10,43] and what value of the proportionality coefficient to use to estimate the electrostatic charging free energy from the ligand-environment electrostatic interaction energy. The LIE equation eq 15 implicitly assumes that the same electrostatic proportionality coefficient, β, applies to both the charging process in solution and in the protein receptor. However, contrary to the water environment, many proteins produce strongly anisotropic electrostatic fields. It is therefore reasonable to assume that a non-negligible electrostatic field exists at the binding site even in the absence of ligand charges. Under the assumption of linear response, eq 12 indicates that the presence of a residual average electrostatic potential at the ligand charge sites in the absence of ligand charges (that is $\langle V_{el}\rangle_0 \neq 0$) will cause the ratio $\Delta F/\langle V_{el}\rangle_1$ to deviate from the ideal solution value of 1/2. This analysis predicts that, although assuming $\beta = 1/2$ is a reasonable first guess, in general it would be advantageous to adopt a LIE regression equation in which the electrostatic estimator is split into a receptor environment component and a solution environment component each multiplied by an independent LIE coefficient.[44] Similarly, the LIE eq 15 implicitly assumes that the work of cavity formation in the protein receptor can be also estimated by the ligand surface area using a single surface tension coefficient applicable to both the water and receptor environments. However, due to the complex reorganization of the receptor binding pocket induced by ligand binding, the solute surface area is likely a poor descriptor for the free energy of ligand cavity formation in protein receptors. In this work we explore the alternative approach

of modeling the work of inserting the ligand in the receptor independently from the work of inserting the ligand in solution.

Most of the limitations of the LIE regression equation underlined here are mitigated in practice by cancellation of errors. The main goal of LIE studies is to obtain relative binding free energies within a family of related ligands, rather than absolute binding free energies. The overall deviation of estimated absolute binding free energies from the experimental affinities is accounted for by an adjustable intercept parameter that is often added to the LIE regression equation (see eq 15).[44] This intercept parameter obviously does not affect relative binding free energies estimated from the LIE equation. For example, the differences of cavity formation free energies between ligands of similar shape could be well represented by a surface area descriptor even though the individual absolute values are not. In this context the limits of the LIE regression equation are manifested in the limited range of applicability of a particular parametrization rather than in the accuracy of the LIE parametrization for a particular ligand set. A better understanding of the properties of the LIE equation could therefore lead to improved LIE ligand coverage and to rules that, based on properties of the ligand, associate a particular parametrization to a particular class of ligands.

2.3. The AGBNP Implicit Solvent Model. The Analytical Generalized Born plus Non-Polar (AGBNP) implicit solvent model[35] is based on an analytical pairwise descreening implementation of the Generalized Born model and a nonpolar hydration free energy model consisting of an estimator for the solute−solvent van der Waals dispersion energy and a surface area term corresponding to the work of cavity formation.

In the Generalized Born (GB) model[33] the electrostatic component of the hydration free energy is estimated as

$$G_{\mathrm{el}} \simeq G_{\mathrm{GB}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{\mathrm{in}}} - \frac{1}{\epsilon_{\mathrm{w}}}\right)\sum_{ij}\frac{q_i q_j}{f_{ij}} \tag{16}$$

where ϵ_{in} is the dielectric constant of the interior of the solute, ϵ_{w} is the dielectric constant of the solvent (in this work $\epsilon_{\mathrm{in}} = 1$ and $\epsilon_{\mathrm{w}} = 80$), q_i and q_j are the charges of atom i and j, and

$$f_{ij} = \sqrt{r_{ij}^{\,2} + B_i B_j \exp(-r_{ij}^{\,2}/4B_iB_j)} \tag{17}$$

where r_{ij} is the distance between atoms i and j, and B_i and B_j are the Born radii of atoms i and j defined below. The summation in eq 16 runs for all atom pairs (i, j) including $i = j$. The diagonal $i = j$ terms can be separated from off-diagonal terms $i \neq j$ yielding the equivalent expression

$$G_{\mathrm{GB}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{\mathrm{in}}} - \frac{1}{\epsilon_{\mathrm{w}}}\right)\left(\sum_i \frac{q_i^{\,2}}{B_i} + 2\sum_{i<j}\frac{q_i q_j}{f_{ij}}\right) \tag{18}$$

The first summation at the right-hand side of eq 18 is the sum of the GB self-energies of the atoms of the molecule, and the second term is the sum of the GB pair-energies. The self-energy of atom i corresponds to the solvation energy of the solute when only the charge of atom i is nonzero. It measures the energy of atom i in the reaction field due to the polarization of the solvent induced by the partial charge of atom i in the solute cavity. The self-energy is largest for the atoms that are most exposed to the solvent because they are capable of inducing stronger polarization fields. This effect is captured by the GB model in that atoms exposed to the solvent have smaller Born radii, whereas buried atoms tend to have larger Born radii. The pair-energy term corresponds to the dampening of electrostatic interactions in a high dielectric medium due to the screening of the solute charges. The GB equation (eq 18) can be shown to be an exact representation of the electrostatic charging free energy of the solute in a continuum dielectric in the two limiting cases of infinite atomic separation and complete atomic overlap.[45]

The Born radius of atom i is defined as the radius of the monatomic solute with partial charge q_i whose continuum dielectric hydration free energy is equal to the self-energy of atom i. In the Coulomb field approximation,[46,47] the Born radius is expressed as an integral centered on the position $\mathbf{r}_i$ of atom i

$$\frac{1}{B_i} = \frac{1}{R_i} - \frac{1}{4\pi}\int_{\Omega i} d^3\mathbf{r}\,\frac{1}{(\mathbf{r} - \mathbf{r}_i)^4} \tag{19}$$

where Ω_i is the bounded region corresponding to the solute volume excluding the atomic sphere corresponding to atom i, and R_i is the van der Waals radius of atom i. $1/R_i$ is the inverse Born radius of atom i in the absence of all the other solute atoms. The second term on the right-hand side of eq 19 takes into account the displacement of the solvent dielectric due to the other solute atoms. In pairwise solute descreening schemes this term is approximated by a pairwise sum[48,49] over the volumes of the neighboring atoms, which are traditionally empirically adjusted to account for atomic overlaps. The AGBNP model instead makes use of a parameter-free geometrical algorithm to calculate the volume scaling coefficients used in the pairwise descreening scheme. The same algorithm is also used to calculate atomic surface areas. This feature is particularly advantageous in ligand binding applications when parametrizations of volume scaling coefficients are not available for some chemical groups. It has been shown that AGBNP gives excellent agreement for the GB self-energies and surface areas in comparison to accurate, but much more expensive, numerical evaluations.[35]

The nonpolar model adopted in this work differs from most other implicit hydration free energy models in that the nonpolar component G_{np} of the hydration free energy is subdivided into cavity and solute−solvent van der Waals interaction terms

$$G_{\mathrm{np}} = G_{\mathrm{cav}} + G_{\mathrm{vdw}} \tag{20}$$

rather than estimated as a whole using a surface area model. The cavity component is described by a surface area model[23−26]

$$G_{\mathrm{cav}} = \sum_i \gamma_i A_i \tag{21}$$

where the summation runs over solute atoms, A_i is the van der Waals surface area of atom i, and γ_i is the surface tension parameter assigned to atom i.[35] The solute−solvent van der Waals free energy term is modeled by the expression

$$G_{vdw} = \sum_i \alpha_i \frac{a_i}{(B_i + R_w)^3} \quad (22)$$

where α_i is an adjustable dimensionless atomic parameter on the order of 1,[35] B_i is the Born radius of atom i, $R_w = 1.4$ Å is the radius of a water molecule, and

$$a_i = -\frac{16}{3}\pi\rho_w\epsilon_{iw}\sigma_{iw}^6 \quad (23)$$

where $\rho_w = 0.033428$ Å^{-3} is the number density of water at standard conditions, and σ_{iw} and ϵ_{iw} are the OPLS force field[50] Lennard-Jones interaction parameters for the interaction of solute atom i with the oxygen atom of the TIP4P water model.[51] Equation 22 is derived by integrating the attractive component of the ligand−water Lennard-Jones potential over the solvent volume assuming homogeneous solvent density.[35]

2.4. LIE Models with Implicit Solvation. *2.4.1. Hydration Free Energy − Implicit Solvent.* If the solvent is described explicitly, the energetic descriptors in eq 14 are evaluated simply by averaging the sum of pair interaction energies between ligand atoms and solvent atoms. It is of interest to derive the expressions for the corresponding estimators when the solvent is treated implicitly. In this case the expression for the canonical configurational partition function of the system includes explicitly only solute degrees of freedom, r^N, and the solute−solvent and solvent−solvent potential energies are replaced by the solvent potential of mean force $W(r^N)$[32,52]

$$Z(\lambda) = \int dr^N \exp[-u(r^N)/kT] \exp[-\lambda W(r^N)/kT] \quad (24)$$

where $u(r^N)$ is the intramolecular potential energy of the solute. The original system is replaced by an equivalent reduced system characterized by the effective potential energy function $U_{eff} = u(r^N) + W(r^N)$. In this section we derive expressions for the LIE estimators in implicit solvent that, based on linear response theory, best correspond to their explicit solvent counterparts. The main difference between the explicit and implicit solvent representations is that the latter lacks fluctuations due to solvent motion. We will show that the response of the implicit solvent environment is significantly different from the response of the explicit solvent environment.

In this study we employ a solvent potential of mean force for water of the form[35]

$$W(r^N) = G_{el} + G_{vdw} + G_{cav} \quad (25)$$

where, as described in the previous section, G_{el} is the electrostatic component (modeled in this work using the Generalized Born model), G_{vdw} corresponds to solute−solvent attractive van der Waals interactions, and G_{cav} is the work for creating the solute cavity in the solvent estimated using a model based on the solute surface area. The process of inserting the solute into the solution can be thought of, starting with $W(r^N) = 0$, as first turning on the cavity component G_{cav}, then the van der Waals component G_{vdw}, and finally the electrostatic component G_{el}. This is analogous to the process of turning on explicit solute−solvent interactions described in the previous section. For example the process of adding the electrostatic component G_{el} after having already added the van der Waals and cavity components (G_{vdw} and G_{cav}) is described by the λ-dependent effective potential energy of eq 2 where

$$U_0 = u + G_{vdw} + G_{cav} \quad (26)$$

is the reference potential, and

$$V = G_{el} \quad (27)$$

is the perturbation. It follows that the linear response estimator for the process of adding solute−implicit solvent interactions is $\langle G_{el}\rangle_w$, the average of the electrostatic implicit solvent term in the ensemble of solute conformations generated when G_{el} is turned on.

Assuming linear response, the ratio between the free energy change for this process and the estimator $\langle G_{el}\rangle_w$, which determines the expected ideal value for the corresponding LIE coefficient, is given by eq 12. In this case, unlike the corresponding electrostatic explicit solvent Coulombic term, the average of G_{el}, $\langle G_{el}\rangle_0$, over the ensemble of conformations obtained in absence of G_{el} in general is not small. Moreover, due to the lack of contributions from solvent motion, the variance, $\langle(\delta G_{el})^2\rangle_0$, of G_{el} is expected to be smaller in relation to the variance of the solute−solvent electrostatic interaction energy in explicit solvent. Equation 12 indicates that, due to both of these effects, the expected value of the electrostatic LIE regression coefficient β when the solvent is treated implicitly should be closer to 1 rather than 1/2. This conclusion applies to the cavity and van der Waals implicit solvent estimators as well.

This is best appreciated in the limiting case when the solute is treated rigidly. In this case, because there are no variable degrees of freedom, $\langle(\delta G_{el})^2\rangle_0 = 0$, and the ensemble average of G_{el} reduces to the value of G_{el} for the given solute conformation. This is true for any of the solvation energy components (cavity, van der Waals, and electrostatic). Therefore, insofar as the solute can be treated as a rigid molecule and the implicit solvent model gives an accurate estimate of the electrostatic hydration free energy of the solute, we have

$$\Delta F_h \simeq G_{el} + G_{vdw} + G_{cav} \quad (28)$$

which is the LIE regression equation for the hydration free energy with all of the adjustable coefficients set to 1. In practice due to solute flexibility and limitations of the implicit solvent model, a LIE regression equation which includes adjustable LIE coefficients fit to experimental hydration free energies

$$\Delta F_h \simeq \alpha\langle G_{vdw}\rangle_w + \beta\langle G_{el}\rangle_w + \gamma\langle G_{cav}\rangle_w \quad (29)$$

which is equivalent to eq 14 when the free energy of cavity formation is modeled using the solute surface area, is

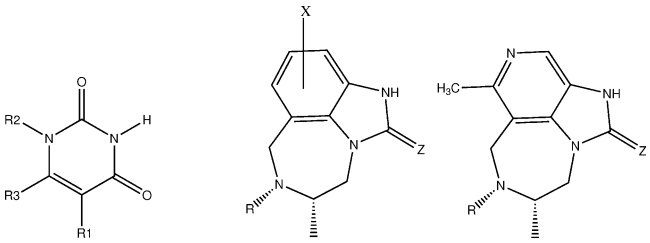


Figure 1. The core structures of the HEPT and TIBO analogues. The TIBO core to the right corresponds to the TIBO pyridyl analogues 27a, 27b, 27c, and 27d.

expected to yield more accurate predictions. Nevertheless the values of the adjustable coefficients α, β, and γ in eq 29 are expected to assume values near 1.

2.4.2. Ligand Binding Free Energy − Implicit Solvent. As pointed out in the previous section, the binding free energy of a ligand to a receptor is calculated as the difference between the work of inserting the ligand in the receptor site solvated by water and the work of inserting the ligand molecule in solution in the absence of the receptor. The LIE equation for the work of creating the ligand in the receptor site takes the same form as eq 29, where now the expressions for the estimators should take into account the fact that the ligand interacts implicitly with the solvent, as modeled by the solvent potential of mean force, as well as with the receptor atoms which are treated explicitly.[10,12] Here we determine, based on linear response theory, the relationship between the free energy changes and the corresponding average effective potential energy changes when part of the system is treated explicitly and part is treated implicitly using the AGBNP implicit solvent model. This is complicated by the fact that implicit solvent models are in general nonpair decomposable. That is it is not possible to define a solute−implicit solvent interaction energy based on pairwise sums between explicit atoms. In the Appendix we derive the following expression for the LIE regression equation for the free energy for creating the ligand in the receptor site

$$\Delta F_c \simeq \alpha \langle V_{LJ} + G^c_{vdw} - G^p_{vdw}\rangle_c + \beta \langle V_{el} + 2(G^c_{el} - G^p_{el})\rangle_c + \gamma \langle G^c_{cav} - G^p_{cav}\rangle_c \quad (30)$$

where all of the averages $\langle\cdots\rangle_c$ are taken for the complex with the ligand fully interacting with the receptor atoms and the solvent continuum. V_{LJ} is the ligand−receptor Lennard-Jones interaction energy, V_{el} is the ligand−receptor electrostatic interaction energy, and G^c and G^p represent implicit solvent free energy terms of the complex and receptor, respectively. Quantities such as $G^c_{el} - G^p_{el}$ are the difference between each implicit solvent energy term evaluated for the complex conformation and the same conformation without the ligand.

The form of the electrostatic LIE estimator for the insertion of the ligand in the complex, $\langle V_{el} + 2(G^c_{el} - G^p_{el})\rangle$, that emerges from our derivation is similar to that proposed by Carlsson et al.[12] Ours includes ligand−receptor interactions (ligand−receptor Coulomb interactions and the pair GB interaction energy between ligand atoms and receptor atoms) and ligand properties (the GB self-energy of the ligand and the GB pair energy between ligand atoms) as well as receptor properties (the change of the GB self-energy of receptor atoms and GB pair energy between receptor atoms due to the displacement of the solvent dielectric by the ligand), which are not included in the electrostatic estimator of Carlsson et al.[12] The implicit solvent LIE van der Waals estimator $\langle V_{LJ} + (G^c_{vdw} - G^p_{vdw})\rangle$ includes ligand−receptor Lennard-Jones interactions as well as two new terms: the ligand−implicit solvent van der Waals interaction energy and the change in the receptor−implicit solvent van der Waals interaction energy upon ligand complexation. This is in contrast to the corresponding estimator in explicit solvent which includes the interactions of the ligand with the receptor and the solvent. This difference can be understood in terms of the additional coupling between ligand atoms and receptor atoms introduced by the partial averaging inherent in the definition of the solvent potential of mean force. In the explicit solvent case the LIE estimator includes explicitly ligand−solvent interactions. In the implicit solvent case the mean effect of the solvent is replaced by effective ligand−receptor as well as ligand−ligand and receptor−receptor interactions.

As previously noted,[12] the implicit LIE solvent electrostatic estimator used earlier by Zhou et al.[4] lacks receptor desolvation contributions. They proposed an electrostatic implicit solvent LIE estimator composed of the GB self-energies of the ligand atoms, the GB pair energies between ligand atoms, and half the GB pair energies between ligand atoms and receptor atoms. The electrostatic estimator derived here based on linear response analysis includes the full amount of the GB pair energies between ligand atoms and receptor atoms and, in addition, the change of self-energies and GB pair energies of the receptor atoms due to the introduction of the ligand. The major difference between the two implementations is the absence of receptor desolvation contributions in the model of Zhou et al. Furthermore, our van der Waals estimator, $\langle V_{LJ} + (G^c_{vdw} - G^p_{vdw})\rangle$, includes estimates of the loss of ligand−solvent and receptor−solvent van der Waals interactions that are not explicitly included in the implicit solvent model adopted by Zhou et al.[4] and Carlsson et al.[12]

To obtain a LIE regression equation for the binding free energy ΔF_b, eqs 30 and 29 can be combined in at least two ways. The first follows previously proposed LIE models in both explicit[3] and implicit[4] solvents. In these models a LIE regression equation is obtained by subtracting eq 29 from eq 30, assuming that the values of LIE coefficients α, β, and γ are the same for the processes of inserting the ligand in solution and in the receptor environment

$$\Delta F_b \simeq \alpha[\langle V_{LJ} + (G^c_{vdw} - G^p_{vdw})\rangle_c - \langle G^f_{vdw}\rangle_w] + \beta[\langle V_{el} + 2(G^c_{el} - G^p_{el})\rangle_c - \langle G^f_{el}\rangle_w] + \gamma[\langle G^c_{cav} - G^p_{cav}\rangle_c - \langle G^f_{cav}\rangle_w] \quad (31)$$

where G^f denotes implicit solvent terms for the ligand free in solution. Alternatively the processes of ligand formation in the receptor and in solution can be decoupled by considering the LIE coefficients for the process of insertion of the ligand in the receptor (eq 30) as indepndent adjustable parameters, whereas the hydration free energy is assumed proportional to the implicit solvent estimate (eq 28) with a

Table 1: Molecular Structures and Experimental IC_{50} and Resulting Binding Free Energies, ΔG_b of the HEPT Analogs

compd	IC_{50}[a]	ΔG_b[b]	R_1	R_2	R_3
H06	0.0027	−12.16	i-Pr	CH_2OCH_2Ph	SPh
H17	0.0027	−12.16	i-Pr	$CH_2OCH_2CH_2OH$	SPh-3,5-di-Me
H11	0.004	−11.89	i-Pr	$CH_2OCH_2CH_3$	CH_2Ph
H18	0.0059	−11.68	Et	CH_2OCH_2Ph	SPh
H10	0.012	−11.24	i-Pr	$CH_2OCH_2CH_3$	SPh
H16	0.013	−11.19	Et	$CH_2OCH_2CH_2OH$	SPh-3,5-di-Me
H09	0.019	−10.96	Et	$CH_2OCH_2CH_3$	SPh
H05	0.088	−10.01	Me	CH_2OCH_2Ph	SPh
H12	0.1	−9.93	c-Pr	$CH_2OCH_2CH_3$	SPh
H15	0.26	−9.35	Me	$CH_2OCH_2CH_2OH$	SPh-3,5-di-Me
H03	0.33	−9.20	Me	$CH_2OCH_2CH_3$	SPh
H20	1.2	−8.40	Me	Bu	SPh
H04	2.1	−8.06	Me	CH_2OCH_3	SPh
H07	2.2	−8.03	Me	Et	SPh
H02	3.6	−7.73	Me	$CH_2OCH_2CH_2CH_3$	SPh
H01	7.0	−7.32	Me	$CH_2OCH_2CH_2OH$	SPh
H13	23.0	−6.52	Me	$CH_2OCH_2CH_2OH$	CH_2Ph
H14	85.6	−5.78	Me	$CH_2OCH_2CH_2OH$	OPh
H08	150[c]	−5.43	Me	Me	SPh
H19	250[c]	−5.11	Me	H	SPh

[a] From ref 53, in μM units. [b] $kT \ln IC_{50}$ at $T = 310$ K, in kcal/mol. [c] Largest concentration tested.

proportionality coefficient ω:

$$\Delta F_b \simeq \alpha\langle V_{LJ} + (G^c_{vdw} - G^p_{vdw})\rangle_c + \beta\langle V_{el} + 2(G^c_{el} - G^p_{el})\rangle_c + \gamma\langle G^c_{cav} - G^p_{cav}\rangle_c - \omega(\langle G^f_{vdw}\rangle_w + \langle G_{el}{}^f\rangle_w + \langle G^f_{cav}\rangle_w) \quad (32)$$

In this work both of these approaches are considered.

3. Ligand Sets and Binding Free Energies

Figure 1 and Tables 1 and 2 show the molecular structures of the HEPT and TIBO analogues. The HEPT NNRTI ligand set used here is the same as the one in the LIE studies by Rizzo et al.[53] in explicit solvent and Zhou et al.[4] in implicit solvent. It is included here for comparison with previous studies. The TIBO NNRTI ligand set has been compiled from the work of Ho et al.[54] who reported activities (IC_{50}'s) of 40 TIBO derivatives and pyridyl analogues. We have included in our study all of the compounds reported by Ho et al. (with the exception of those few for which only an activity range was reported rather than a single activity value) plus additional compounds which were part of a TIBO subset studied previously by LIE modeling.[55] The resulting 40 TIBO ligands which were simulated in our study are listed in Table 2. (Based on the values of the energetic descriptors obtained from the simulation, 3 of these 40 ligands were not included in the LIE training set, see below.) HEPT compounds are labeled using the naming scheme in ref 53; for the TIBO compounds we adopted the names in refs 54 and 55. In Table 1 H11 is the reference HEPT compound MKC-442, and in Table 2 the reference compound is 1a (also known as TIBO-9Cl).

4. Simulation Procedure

A model of the NNRTI binding site is constructed as previously described[53] by including in the simulation only the region of HIV-RT closest to the ligand. The same set of 124 residues was included in the model for both the TIBO and HEPT compounds. These 124 residues are organized in 7 protein segments: residues 91−110, 161−205, 222−242, 316−321, 343−349, and 381−383 in the p66 subunit and 134−140 in the p51 subunit. The end of each fragment is capped with acetyl (ACE) or *N*-methylamide (NMA) groups. Residues 95−108, 179−183, 186−191, 198, 225−229, and 318−319 of the p66 subunit and residues 136 and 138 in the p51 subunit are free to move in the MD simulation. The positions of the atoms of residues 94, 109, 178, 184, 197, 199, 224, 230, 231, and 240 in the p66 subunit and 135, 137, and 139 in the p51 subunit are restrained using a harmonic restraining potential with a force constant of 25 kcal/mol Å^2. All other residues are held fixed. Residues LYS101, LYS102, LYS103, LYS104, and LYS238 are protonated, whereas ASP186, ASP192, ASP237, GLU138, and GLU233 are deprotonated. All other ionizable residues are set in their neutral charge state.

We employ the structure of the complex of HIV-1 RT with MKC-442 (PDB id 1rt1)[56] as the template structure from which the initial structures of all the HEPT complexes are constructed. For the TIBO compounds the corresponding template structure is from the complex with 9-Cl TIBO[57] (PDB 1tvr). The template structures were first energy minimized, and then each complex was constructed by modifying the ligand starting from the corresponding reference ligands (H11 and 1a in Tables 1 and 2) by manual editing of the ligand structure using Maestro.[58] Each complex as modified is then energy minimized to relieve steric clashes. One conformation of each ligand was constructed. The ligand conformations thus obtained were also used as starting conformations for the molecular dynamics simulations in the solution phase using the same protonation state and force field parameters as in the receptor-bound simulations.

All molecular dynamics calculations are performed using the program IMPACT[59] with the 2001 parametrization of the OPLS-AA/AGBNP force field.[35,50] The RESPA multiple

Table 2: Molecular Structures and Experimental IC_{50} and Resulting Binding Free Energies, ΔG_b of the TIBO Analogs

compd	IC_{50}[a]	ΔG_b[b]	X	R	Z
10f	0.0030	−12.09	8-Br	DMA	S
6a	0.0043	−11.87	8-Cl	DMA	S
6c	0.0050	−11.78	8-SCH_3	DMA	S
6b	0.0058	−11.69	8-F	DMA	S
10m	0.0136	−11.16	8-CH_3	DMA	S
17	0.0250	−10.79	9-F	DMA	S
20	0.0255	−10.76	9,10-di-Cl	DMA	S
1d	0.0295	−10.69	8-CCH	DMA	S
10j	0.0296	−10.68	8-CH_2CH_3	DMA	S
1a	0.034	−10.6	9-Cl	DMA	S
6d	0.034	−10.6	8-OCH_3	DMA	O
1	0.044	−10.44	H	DMA	S
10h	0.0473	−10.39	8-I	DMA	S
10e	0.0474	−10.39	8-Br	DMA	O
10b	0.0563	−10.29	8-CN	DMA	S
10g	0.088	−10.01	8-I	DMA	O
6e	0.0959	−9.96	8-OCH_2CH_3	DMA	S
10c	0.188	−9.54	8-COH	DMA	S
27c	0.243	−9.39		DEA	S
18c	0.3142	−9.23	8-CH_3	DEA	O
1i	0.4371	−9.02	8-CCH	DMA	O
10i	0.4376	−9.02	8-CH_2CH_3	DMA	O
18b	0.485	−8.96	9-CF_3	DMA	S
10l	0.989	−8.52	8-CH_3	DMA	O
21	1.075	−8.47	10-Br	DMA	S
10a	1.1396	−8.43	8-CN	DMA	O
27b	2.0	−8.09		DEA	O
16	2.45	−7.96	9-NO_2	CPM	S
1l	3.155	−7.81	H	DMA	O
19b	4.725	−7.56	10-OCH_3	DMA	S
18a	5.919	−7.42	9-CF_3	DMA	O
19a	6.63	−7.35	10-OCH_3	DMA	O
14b	6.65	−7.35	8-$N(CH_3)_2$	CPM	O
15b	6.65	−7.35	9-$N(CH_3)_2$	CPM	O
13	33.43	−6.35	9-NO_2	CPM	O
15a	60.55	−5.98	9-NH_2	CPM	O
15c	159	−5.39	9-$NHCOCH_3$	CPM	O
27d	596	−4.58		CPM	O
14a	849	−4.36	8-NH_2	CPM	O
27a	872	−4.34		DMA	O

[a] From ref 54, in μM units. [b] $kT \ln IC_{50}$ at $T = 310$ K, in kcal/mol.

time step MD integrator with an inner time step of 0.25 fs for covalent interactions, and a time step of 1 fs for nonbonded interactions is employed. Temperature is controlled by velocity rescaling. The simulations for the unbound ligand consist of 10 ps of heating from 10 to 310 K and 25 ps of equilibration at 310 K followed by 100 ps of data collection. Simulations of the inhibitor−protein complex consist of 50 ps of heating from 10 to 310 K and 50 ps of equilibration at 310 K and 100 ps of data collection. A residue-based nonbonded neighbor list with a 15 Å distance cutoff was used in the inhibitor−protein complex simulations.

Energetic analysis is conducted on trajectory files collected during the data collection phase of each simulation. Each energetic quantity is averaged to obtain values of the LIE descriptors for each ligand according to eqs 33−40. Convergence was monitored by plotting the running average of each property.

5. Results

5.1. LIE Descriptors. The values of the calculated energetic descriptors for the two ligand sets are listed in Tables3 and 4. In these tables and in the rest of the paper one- and two-letters mnemonics are used to identify the calculated descriptors as follows

$$\mathrm{EC} = \langle V_{\mathrm{el}} \rangle_{\mathrm{c}} \tag{33}$$

$$\mathrm{ES} = \langle G^{\mathrm{c}}_{\mathrm{el}} - G^{\mathrm{p}}_{\mathrm{el}} \rangle_{\mathrm{c}} \tag{34}$$

$$\mathrm{EL} = \langle G^{\mathrm{f}}_{\mathrm{el}} \rangle_{\mathrm{w}} \tag{35}$$

$$\mathrm{LJ} = \langle V_{\mathrm{LJ}} \rangle_{\mathrm{c}} \tag{36}$$

$$\mathrm{V} = \langle G^{\mathrm{c}}_{\mathrm{vdw}} - G^{\mathrm{p}}_{\mathrm{vdw}} \rangle_{\mathrm{c}} \tag{37}$$

$$\mathrm{VL} = \langle G^{\mathrm{f}}_{\mathrm{vdw}} \rangle_{\mathrm{w}} \tag{38}$$

$$\mathrm{C} = \langle G^{\mathrm{c}}_{\mathrm{cav}} - G^{\mathrm{p}}_{\mathrm{cav}} \rangle_{\mathrm{c}} \tag{39}$$

$$\mathrm{CL} = \langle G^{\mathrm{f}}_{\mathrm{cav}} \rangle_{\mathrm{w}} \tag{40}$$

where V_{el} and V_{LJ} are, respectively, the Coulomb and Lennard-Jones interaction energies between the ligand and receptor atoms, G_{el}, G_{vdw}, and G_{cav} refer to, respectively, the electrostatic, van der Waals, and cavity components of the AGBNP implicit solvent model of either the receptor−ligand complex (denoted by "c"), the receptor (denoted by "p"), or the ligand (denoted by "f "). Averages are taken for either the receptor−ligand complex (denoted by $\langle \cdots \rangle_{\mathrm{c}}$) or the ligand free in solution (denoted by $\langle \cdots \rangle_{\mathrm{w}}$). The difference between complex and receptor energies are calculated for each conformation of the complex by calculating the energy of the complex and that of the receptor obtained by removing the ligand without changing the conformation of the receptor. For instance $G^{\mathrm{c}}_{\mathrm{el}} - G^{\mathrm{p}}_{\mathrm{el}}$ is the difference between the GB electrostatic energy of the given complex conformation and the corresponding quantity for the same conformation after removal of the ligand. See the Appendix for additional discussion of this point.

5.2. HEPT and TIBO Binding Modes. As was noted earlier, the HEPT inhibitors adopt a butterfly conformation as shown in Figure 3. The two wings are formed by the R_3 side chain (see Figure 1) and the ligand core. To accommodate the ligand the protein forms a complementary pocket which has been described as acting as a "shrink wrap".[60] Compared with the crystal structure of the complex with MKC-442, the molecular dynamics trajectories of the HEPT complexes show relatively little conformational variation. The aromatic side chain (R_3) interacts favorably with TYR181, TYR188, and TRP229, and the thiothymine ring interacts with LYS101 and LYS103. The mostly hydrophobic R_2 side chain contacts with LEU234, PHE227, and VAL106. The position of the R_3 side chain is well conserved for all ligands, indicating that π stacking between R_3 and TYR181 is a prerequisite for binding. The role of the aliphatic R_1

Table 3: Computed Values of the Energetic Descriptors for the HEPT Compounds[a]

compd	EC	ES	EL	LJ	V	VL	C	CL
H06	−12.59	−0.07	−20.35	−62.77	−2.97	−42.23	−17.93	37.37
H17	−14.05	3.72	−18.51	−58.92	−3.61	−39.82	−13.25	35.04
H11	−11.08	6.27	−10.63	−52.91	−2.95	−35.33	−11.04	31.32
H18	−16.85	6.03	−20.11	−62.79	−3.38	−40.87	−17.31	36.78
H10	−11.54	1.29	−16.55	−53.97	−2.54	−35.50	−11.89	31.51
H16	−12.03	3.02	−18.70	−59.93	−3.75	−38.59	−10.31	34.32
H09	−9.12	0.43	−16.36	−53.55	−3.58	−33.80	−11.20	30.34
H05	−11.94	1.47	−20.76	−59.83	−3.84	−39.59	−15.03	35.12
H12	−10.13	0.49	−15.67	−52.82	−2.63	−34.86	−11.58	31.39
H15	−11.93	3.23	−19.39	−53.99	−4.25	−37.59	−10.40	33.11
H03	−1.17	−6.10	−16.66	−49.15	−3.40	−33.15	−9.69	29.35
H20	−7.04	5.80	−8.38	−49.22	−2.72	−32.32	−8.90	28.19
H04	−13.67	5.39	−15.19	−50.67	−3.82	−33.53	−9.51	29.94
H07	−12.78	6.32	−9.46	−43.68	−2.50	−29.71	−9.00	26.16
H02	−11.21	3.00	−16.81	−50.49	−2.41	−35.08	−10.51	31.00
H01	−19.31	10.15	−18.38	−53.59	−2.25	−33.74	−12.09	29.88
H13	−16.09	5.17	−15.32	−42.51	−3.63	−31.09	−5.79	27.61
H14	−9.36	3.42	−19.14	−49.16	−3.02	−32.68	−11.26	29.04
H08	−14.38	8.75	−10.28	−39.42	−3.24	−28.18	−4.78	25.22
H19	−6.01	4.33	−10.18	−39.57	−2.16	−26.58	−7.58	24.42

[a] Values are in kcal/mol. The list is ordered from strongest to weakest binders.

side chain is to modulate the strength of the binding. Binding is favored by small branched groups in this position (see Table 1). As in previous simulations,[19] for most HEPT complexes we observe a stable hydrogen bond between the NH group of the thymine ring with the CO group of the LYS101 backbone. An additional hydrogen bond between the CO of the thymine ring and the NH group of LYS101 is also observed in some conformations collected from the simulations. Previous studies of HEPT in explicit solvent[53] have shown that the R_2 side chain is partly solvent exposed and forms hydrogen bonds with water molecules. Solvent exposure of this ligand side chain, which helps lower the desolvation penalty of these ligands, is also observed in our simulations in implicit solvent. For reasons that are not immediately apparent, the observed binding mode of the H03 HEPT ligand in the simulation differs from all the other HEPT ligands. Relative to the consensus binding mode the core of H03 is twisted, the hydrogen bond with the lysine residues is absent, and the R_2 side chain is less solvent exposed. This causes the unusually small values of the EC and ES descriptors of H03 (Table 3). Nevertheless most of the LIE models we developed (see below) predict binding free energies for H03 in agreement with the experimental binding free energy.

The binding mode of TIBO compounds (see Figure 3) is similar to that of the HEPT compounds except that more variability is observed. The dimthylallyl (DMA) or cyclopropylmethyl (CPM) substitutions in the R position of the TIBO cores (see Figure 1) replaces the aromatic side chain of the HEPT compounds in the R_3 position. As for the HEPT compounds, this group interacts with TYR181, TYR188, and TRP229 but without the extensive π stacking interactions characteristic of HEPT complexes. The DMA (or CPM) ligand side chain interacts with the protein in a variety of orientations and is not as constrained as the R_3 aromatic side chain of the HEPT compounds. In general we find that TIBO compounds form more hydrogen bond interactions with the receptor than the HEPT compounds. Both the NH group and the oxygen (or sulfur) atoms in the Z position of the TIBO cores form hydrogen bonds with the backbone of LYS101. For some ligands (6d, 18c, 1i, and 27b) we found an additional hydrogen bond between the carbonyl of the TIBO cores and the side chain of LYS103. These additional ligand−receptor hydrogen bonds are consistent with the larger ligand−receptor Coulomb interactions calculated for the TIBO compounds (compare the EC descriptor in Tables 3 and 4). Nonspecific electrostatic interactions, however, seem to play a role as well; although ligands 14a and 14b are the ones with the strongest ligand−receptor interaction energies, these ligands do not present on average more hydrogen bonds than the other TIBO ligands. The unusually small ligand−receptor electrostatic energy of ligand 6c is consistent with the absence for this ligand of one of the hydrogen bonds with LYS101. Ligand 10e also presents a weak electrostatic interaction with the protein. This ligand differs from all the other ligands in that it forms a single hydrogen bond with LYS103 rather than LYS101. The weak electrostatic interaction energies of these two ligands are counterbalanced by the correspondingly small receptor desolvation penalties (the ES descriptor in Table 4). In these complexes the LYS101 backbone remains partly solvent exposed suggesting the possibility of water-mediated interactions with the ligand. Both ligands 6c and 10e are characterized by stronger than average van der Waals interactions with the protein due to the almost parallel contact between the dimethylallyl group and the TYR181. Our simulations as well as similar observations by others[20] indicate that both for the HEPT and TIBO complexes a competition exists between hydrophobic and electrostatic interactions. The stronger the electrostatics interactions, the less the ligand is able to make good contacts with TYR181 and the other hydrophobic residues lining the binding pocket.

Table 4: Computed Values of the Energetic Descriptors for the TIBO Compounds[a]

compd	EC	ES	EL	LJ	V	VL	C	CL
10f	−24.44	13.60	−12.14	−51.75	−3.04	−35.46	−12.28	29.14
6a	−23.86	15.78	−9.10	−50.90	−1.85	−33.84	−14.33	28.36
6c	−12.44	9.32	−13.30	−56.02	−3.16	−36.28	−10.75	31.06
6b	−24.65	18.06	−9.68	−48.73	−1.34	−32.52	−14.43	28.40
10m	−24.69	18.06	−11.55	−51.28	−2.91	−34.00	−12.04	29.31
17	−22.29	13.80	−9.82	−47.91	−2.20	−32.22	−10.62	28.79
20	−29.38	19.59	−9.28	−51.51	−2.54	−36.83	−10.80	31.48
1d	−37.98	25.85	−10.63	−50.71	−2.75	−35.36	−11.94	30.16
10j	−24.22	16.04	−9.92	−49.67	−1.01	−33.53	−14.37	29.96
1a	−25.63	15.85	−9.64	−47.72	−2.73	−34.45	−8.65	29.67
6d	−47.70	28.26	−11.02	−48.21	−2.80	−34.58	−11.40	29.33
1	−23.83	15.04	−10.42	−45.50	−2.57	−31.81	−10.55	27.55
10h	−14.22	−3.20	−37.95	−50.13	2.04	−31.42	−11.13	31.10
10e	−10.86	4.07	−14.88	−48.42	−4.07	−34.78	−10.28	28.37
10b	−28.50	18.55	−11.25	−49.18	−1.79	−33.31	−10.77	29.71
10g	−23.80	3.27	−35.73	−48.53	2.68	−31.82	−13.77	31.71
6e	−35.59	24.21	−11.83	−52.48	−2.85	−36.35	−11.34	30.24
10c	−37.51	23.43	−14.57	−50.61	−2.23	−34.69	−11.10	30.84
27c	−21.95	14.29	−11.82	−54.00	−3.12	−36.90	−11.56	29.70
18c	−41.27	26.35	−12.77	−50.85	−2.96	−36.17	−14.27	30.02
1i	−49.47	29.96	−13.72	−49.11	−2.94	−34.44	−11.95	28.90
10i	−31.26	19.54	−12.87	−47.47	−1.95	−32.62	−11.35	28.84
18b	−29.20	20.78	−9.92	−48.65	−0.98	−34.13	−12.44	30.75
10l	−28.54	13.61	−15.33	−45.67	−2.78	−33.21	−10.20	28.31
21	−27.80	17.16	−13.30	−47.61	−3.81	−35.36	−8.19	29.06
10a	−41.51	24.46	−14.30	−47.44	−1.79	−32.75	−11.57	29.34
27b	−40.90	26.85	−12.48	−51.78	−3.20	−35.97	−11.29	29.60
16	−23.77	19.11	−10.07	−47.88	−1.91	−32.76	−9.20	29.32
1l	−25.13	13.89	−13.75	−45.98	−2.98	−31.04	−10.51	26.77
19b	−27.00	18.11	−11.24	−48.57	−3.50	−35.38	−9.757	30.82
18a	−37.30	24.09	−13.15	−45.89	−0.94	−33.39	−11.95	29.88
19a	−23.90	15.67	−14.68	−47.57	−3.11	−34.01	−12.42	29.13
14b	−46.70	25.80	−13.76	−45.08	−1.82	−33.33	−11.45	28.96
15b	−43.93	26.24	−13.72	−46.18	−2.22	−32.97	−11.03	27.95
13	−35.88	21.78	−14.49	−47.26	−1.46	−31.95	−11.56	28.35
15a	−23.75	16.16	−14.95	−45.27	−3.27	−29.81	−7.03	26.35
15c	−35.30	13.81	−30.85	−50.80	−3.66	−35.31	−9.00	30.61
27d	−38.32	22.49	−12.84	−43.96	−2.70	−30.56	−9.23	25.70
14a	−49.35	29.39	−14.70	−41.03	−1.34	−29.62	−12.22	25.74
27a	−41.09	26.34	−12.69	−47.97	−2.18	−32.58	−14.31	27.31

[a] Values are in kcal/mol. The list is ordered from strongest to weakest binders.

5.3. Comparison of HEPT and TIBO LIE Descriptors. The values of the calculated descriptors in Tables 3 and 4 show clear differences between the HEPT and TIBO data sets. The HEPT estimators based on van der Waals interactions (LJ, LJ + V and LJ + V − VL) exhibit stronger correlation with the experimental affinities than the corresponding TIBO estimators. A similar trend is observed for cavity estimators (C and C − CL). The van der Waals and cavity estimators both generally reflect the amount of surface area of the ligands. Indeed we find a strong linear correlation between the LJ and CL descriptors for the HEPT compounds ($R^2 = 0.94$). This correlation is significantly weaker for the TIBO compounds ($R^2 = 0.52$).

Due to the presence of additional ligand−receptor hydrogen bonds discussed above, the ligand−receptor electrostatic interaction energies (the EC descriptor) of the TIBO compounds are significantly more favorable than those of the

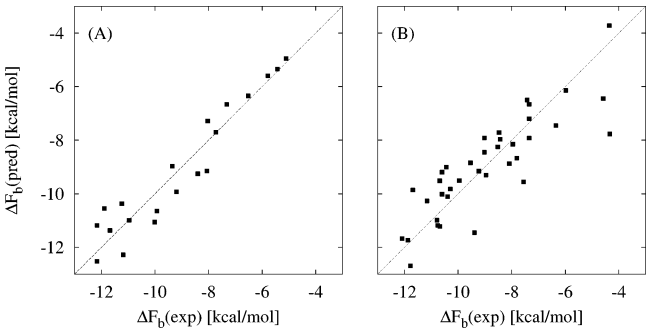


Figure 2. Predicted versus experimental binding free energies of the (A) HEPT and (B) TIBO analogues to HIV-RT for the MDL3 LIE model. The correlation coefficients and RMSDs for this model are 0.90 and 0.71 kcal/mol and 0.73 and 1.08 kcal/mol for the HEPT and TIBO compounds, respectively (see Table 5).

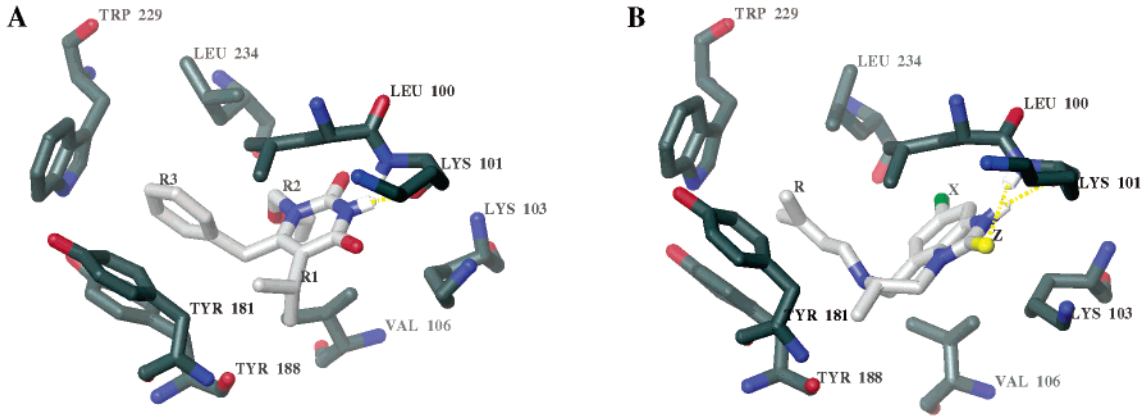


Figure 3. Representative conformations of the H11-RT (A) and 1a-RT (B) complexes extracted from the MD trajectories. Carbon atoms of the ligand are in light gray; those of the protein are in dark gray. Oxygen atoms are red, nitrogen atoms are blue, hydrogen atoms are white, the sulfur atom is yellow, and the chlorine atom is green. Hydrogen bonds are yellow.

HEPT compounds. However this trend is reversed when the estimator EC + 2ES, which takes into account desolvation, is considered, suggesting that, due to desolvation penalties, stronger ligand−receptor electrostatic interactions do not necessarily lead to better binding. Indeed the sign of the EC + 2ES estimator shows that generally this measure of the electrostatic component of the work of inserting the ligand into the receptor favors binding for the HEPT compounds (negative values), whereas it disfavors binding for the TIBO compounds (positive values). The EC + 2ES electrostatic estimator is a better predictor for the HEPT binding energies than for the TIBO binding energies. On the other hand, the EC + 2(ES − EL) estimator, which models the transfer of the ligand from solution into the receptor and also incorporates the loss of ligand−solvent interactions, is a better estimator for the binding energies of the TIBO compounds than the HEPT compounds. Finally the ligand hydration free energy estimator EL + VL + CL alone is rather weakly correlated to the HEPT and TIBO binding affinities.

These observations indicate that the energetic balance that drives binding is very different between these two sets of ligands and suggest that accurate modeling of both sets of ligands using a single regression equation would be difficult. It is also apparent that the HEPT compounds constitute a less challenging data set than the TIBO compounds. The HEPT binding free energies in general can be predicted fairly accurately using a single descriptor (the LJ descriptor), whereas for the TIBO compounds there is no single strong predictor of binding free energies. Successful prediction of the TIBO binding free energies must hinge on the interplay between the various predictors rather than on any single predictor.

Analysis of the calculated predictors for the TIBO compounds (Table 4) reveals a few ligands that stand out from the rest. Ligands 10g, 10h, and 15c have much more favorable estimated electrostatic hydration free energies (the EL descriptor) than all the other ligands. These unusually large ligand solvation free energies disfavor binding and would lead to the prediction of poor binding affinities if not counterbalanced by equally favorable ligand−receptor Coulomb interactions (the EC descriptor) and residual complex−solvent interactions (the ES descriptor). The 10g and 10h ligands are characterized by less positive ES values than the other ligands, but the EC descriptors are of equal magnitude as the other ligands. This results in unusually unfavorable electrostatic contributions to binding (measured for example by the EC + 2(ES − EL) estimator) which are inconsistent with the measured relatively large potency of these ligands. Indeed these two ligands show up as outliers in all of the LIE models we have tried. Both of these ligands have the iodo-substitution in the X position. This indicates a possible deficiency of the OPLS-AA/AGBNP force field for iodo-substitutions. Hence, suspecting inaccurate energetic modeling, we removed ligands 10g and 10h from the LIE training set. Similar observations apply to ligand 15c except that for this ligand the unfavorable electrostatic contribution to binding is consistent with its poor measured inhibition activity. Nevertheless this ligand appears as a clear outlier in some of the LIE models we investigated. The large predicted electrostatic hydration free energies of 15c is related to the presence of the acetoamido group in the X position. The acetoamido group, due to its large dipole moment, is by far the most polar substituent in this class of ligands. It is likely that the binding mode of 15c differs from the binding mode of 1a on which the structural model for the HIV-RT/15c complex is based. For this reason we have decided to remove the 15c ligand from the training set as well. The final TIBO training set consists of 37 ligands.

5.4. LIE Regression Models. The descriptors of eqs 33−40 are combined based on eqs 31 and 32 to construct two classes of models. The first class, based on the assumption that the same linear response proportionality coefficients apply to the work of inserting the ligand in the receptor and in water, is described by the general regression equation

$$\Delta F_b = \beta[\mathrm{EC}] + \beta'[\mathrm{ES} - \mathrm{EL}] + \alpha[\mathrm{LJ} + \mathrm{V} - \mathrm{VL}] + \gamma[\mathrm{C} - \mathrm{CL}] + \delta \quad (41)$$

where ΔF_b is the binding free energy, and the estimators result from the combination of descriptors of the complex (EC, ES, LJ, V, and C) and descriptors of the free ligand (EL, VL, and CL).

The second class of LIE models, given by the general expression

$$\Delta F_b = \beta[\mathrm{EC}] + \beta'[\mathrm{ES}] + \alpha[\mathrm{LJ} + \mathrm{V}] + \gamma[\mathrm{C}] - \omega[\mathrm{EL} + \mathrm{VL} + \mathrm{CL}] + \delta \quad (42)$$

treat the process of inserting the ligand in solution independently from the work of inserting the ligand in the receptor pocket. In this class one of the LIE estimators is the hydration free energy of the ligand, EL + VL + CL, estimated by the implicit solvent model with the LIE proportionality coefficient ω. All the other estimators include descriptors related only to the complex.

In addition to the electrostatic, van der Waals, and cavity LIE proportionality coefficients, the LIE regression equations we have investigated include an adjustable intercept parameter δ. Inclusion of this parameter is, in our view, crucial to the success of LIE paramterizations. The intercept parameter absorbs effects that influence equally absolute binding free energies and are therefore inconsequential for the ultimate goal of estimating relative binding free energies of a family of similar ligands. It is unreasonable and unnecessary to assume that linear response is applicable to the full process of ligand formation in the receptor and in solution. The success of a LIE model which includes an intercept parameter hinges only on the applicability of linear response for the simpler process of mutating one ligand into another. The inclusion of the intercept parameter allows the LIE coefficients to reflect relative binding free energies rather than absolute binding free energies, leading to better accuracy in ligand ranking as well as values of LIE coefficients in better agreement with linear response predictions.

We have investigated several models based on eqs 41 and 42, including those with the maximum number of adjustable parameters allowed by the general expressions (5 for eq 41 and 6 for eq 42). We present results for the following LIE regression models with 3 or fewer adjustable parameters other than the intercept, one from the first class (eq 41) and two from the second class (eq 42):

$$\mathrm{MDL1}: \Delta F_b = \beta[\mathrm{EC} + 2(\mathrm{ES} - \mathrm{EL})] + \alpha[\mathrm{LJ} + \mathrm{V} - \mathrm{VL}] + \gamma[\mathrm{C} - \mathrm{CL}] + \delta \quad (43)$$

$$\mathrm{MDL2}: \Delta F_b = \beta[\mathrm{EC} + 2\mathrm{ES}] + \alpha[\mathrm{LJ} + \mathrm{V}] + \gamma[\mathrm{C}] - \frac{1}{2}[\mathrm{EL} + \mathrm{VL} + \mathrm{CL}] + \delta \quad (44)$$

$$\mathrm{MDL3}: \Delta F_b = \frac{1}{2}\left\{\frac{1}{2}[\mathrm{EC}] + [\mathrm{ES}] + [\mathrm{LJ} + \mathrm{V}] - [\mathrm{EL} + \mathrm{VL} + \mathrm{CL}]\right\} + \gamma[\mathrm{C}] + \delta \quad (45)$$

We have chosen these as the most representative models because the first, MDL1, is comparable to standard models used in previous LIE studies[3,4] and the second, MDL2, is based on the novel LIE formulation (eq 42) with the same number of adjustable parameters as MDL1. The third model, MDL3, is a minimally parametrized version of MDL2. MDL1 is based on eq 41 with $\beta' = 2\beta$, MDL2 is based on eq 42 with $\beta' = 2\beta$ and $\omega = 1/2$, and MDL3 is the same as MDL2 with the additional constraints that $\beta = 1/2$ and $\alpha = 1$. Further motivation for these choices of parameters is provided below and in the Discussion section, where we also discuss some of the results obtained with the other models we tested.

Table 5 shows the results of the LIE regressions for the HEPT and TIBO ligand sets using the models presented above. This table contains the results of fitting MDL1, MDL2, and MDL3 to the HEPT set, the TIBO set, and the two sets combined. It reports the values of the LIE coefficients obtained from multivariate linear least-squares fitting, the square of the correlation coefficient (R^2), and root-mean-square deviation (RMSD) between experimental and estimated LIE binding free energies. R_{pred}^2 and $RMSD_{pred}$ are the corresponding quantities for the jack-knife regression tests. In these tests a LIE model is fit to the ligand set removing each ligand in turn and then using the resulting LIE model to predict the binding free energy of the ligand that was left out. The jack-knife results give a better unbiased representation of the predictive ability and transferability of each model.

MDL1 closely follows corresponding formulas for LIE estimates of ligand binding from explicit solvent simulations. The electrostatic estimator (EC + 2(ES − EL)) is the difference between the bound and solution phases of the average electrostatic interaction energy between the ligand and the environment. In the bound form the electrostatic interaction energy is estimated as the sum of the ligand−receptor average electrostatic interaction energy plus twice the Generalized Born energy of the ligand in the complex as measured by the ES descriptor (eq 34). In the solution phase the average ligand-environment electrostatic energy is measured as twice the EL descriptor (eq 35). As discussed in the previous section and in the Appendix, the assumption that $\beta' = 2\beta$ in eq 41 puts the estimators corresponding to interactions between explicit atoms and the implicit solvent components on equal footing. To validate this choice we have tested two additional models based on eq 41, one setting $\beta' = \beta$ and another one in which the β and β' coefficients are allowed to vary independently. The results of these tests are discussed in the next section.

The second model (MDL2) has the same number of adjustable parameters as MDL1. It is based on eq 42 with $\beta' = 2\beta$ and the observation that setting $\omega = 1/2$ leads to more accurate results than choosing the value $\omega = 1$ suggested based on linear response theory (see Theory and Methods section). Indeed, as further discussed later, we find that, while their ratios agree with linear response predictions, the absolute values of the fitted β, β', and α LIE coefficients are approximately half their theoretical values. The final model (MDL3) exploits this observation by using only one adjustable coefficient (beside the intercept parameter): the surface tension parameter γ for creating the ligand cavity in the receptor. All the other parameters are set to half the values expected from linear response theory.

6. Discussion

6.1. Model Performance. MDL1 uses estimators that are designed to mimic the corresponding estimators commonly used in LIE explicit solvent studies. This model is also comparable to the implicit solvent LIE formulation introduced by Zhou et al.[4] The R^2 and RMSD values obtained in this work (see Table 5) for the HEPT compounds using MDL1 are 0.87 and 0.80 kcal/mol, respectively, compared

Table 5: Fitting Results for MDL1, MDL2, and MDL3 to the HEPT and TIBO Sets, Individually and Combined[b]

model	n_p	β	α	γ	δ	R^2	R_{pred}^2	RMSD	$RMSD_{pred}$
				HEPT Set (20 Ligands)					
MDL1	4	0.22	0.40	0.20	1.56	0.87	0.81	0.80	0.97
MDL2	4	0.21	0.45	−0.05	6.28	0.91	0.87	0.66	0.80
MDL3	2	1/4[a]	1/2[a]	−0.18	7.73	0.90	0.88	0.71	0.77
				TIBO Set (37 Ligands)					
MDL1	4	0.27	0.40	0.36	3.35	0.70	0.64	1.12	1.23
MDL2	4	0.23	0.45	0.19	5.80	0.73	0.66	1.07	1.19
MDL3	2	1/4[a]	1/2[a]	0.19	8.09	0.73	0.68	1.08	1.16
				HEPT+TIBO Set (57 Ligands)					
MDL1	4	0.15	0.11	0.29	0.21	0.55	0.48	1.43	1.54
MDL2	4	0.07	0.39	0.17	4.15	0.67	0.62	1.23	1.31
MDL3	2	1/4[a]	1/2[a]	0.004	7.34	0.07	−0.02	2.06	2.16

[a] Set value not allowed to vary during the fitting. [b] n_p is the number of adjustable parameters (including the intercept parameter), and β, α, γ, and δ are the LIE adjustable coefficients. R^2 is the square of the correlation coefficient of the fit, and RMSD is the root-mean-square deviation between the LIE equation and the experimental binding free energies. R_{pred}^2 and $RMSD_{pred}$ are the corresponding quantities for the jack-knife validations.

to $R^2 = 0.78$ and RMSD = 1.07 kcal/mol reported by Zhou et al. for the same set of ligands. Zhou et al. report that the binding free energies of MKC-442 (compound H11) and H14 were significantly underpredicted by their models. These two ligands are correctly predicted by the MDL1 model (the deviation between predicted and experimental binding free energies for these two compounds are 0.58 and 0.03 kcal/mol, respectively). The superior performance of the MDL1 LIE model for the HEPT compounds relative to the LIE results reported by Zhou et al. is likely due in part to the differences in the expressions for the LIE estimators but may also reflect differences in the conformational ensemble sampled by the two studies. Our MD simulations, including thermalization and equilibration phases, are longer than those performed by Zhou et al., and the implicit solvent models employed in the two studies are different: SGB/SA[61] in the study of Zhou et al. and AGBNP[35] in the present study. The electrostatic LIE fitting coefficients β we obtained with MDL1 (0.22) for the HEPT compounds is in reasonable agreement with the value obtained by Zhou et al., but the van der Waals and cavity parameters α and γ are significantly different. The smaller values for the α parameter obtained by Zhou et al. are most likely due to differences in the corresponding implicit solvent functional form. The large differences in LIE γ fitting coefficients are mainly due to the differences in the values of the cavity surface tension employed in the two studies.

The same set of HEPT compounds were also studied previously using extended LIE models based on descriptors obtained from explicit solvent simulations.[19,53] Based on 3 descriptors (variation in number of hydrogen bonds upon binding, ligand−protein Lennard-Jones interaction energy, and variation in exposed hydrophobic surface area), and excluding H17, the reported correlation coefficient of the best performing extended LIE model[19] is $R^2 = 0.85$ and RMSD = 0.87 kcal/mol similar to our MDL1 results ($R^2 = 0.87$, RMSD = 0.80 kcal/mol).

Subsets of the TIBO compounds studied here have been previously studied by LIE modeling. Smith et al.[55] studied 12 TIBO compounds using explicit solvent LIE models similar to our MDL1 model. Smith et al. obtain RMSDs between 0.9 and 1.0 kcal/mol. Rizzo et al.[19] considered 22 TIBO compounds using two descriptors (variation in number of hydrogen bonds upon binding and ligand−protein Lennard-Jones interaction energy), obtaining, after excluding two outliers, $R^2 = 0.79$ and RMSD = 0.75 kcal/mol. In comparison, results for the same set of ligands using MDL1 are $R^2 = 0.88$ and RMSD = 0.71 kcal/mol. The TIBO set studied here (37 compounds) is larger and more diverse than in the two previous LIE studies; nevertheless the MDL1 LIE model achieves good accuracy for these ligands (see Table 5).

The results of fitting MDL1, MDL2, and MDL3 (eqs 43−45) to the HEPT and TIBO compounds separately and combined are listed in Table 5. MDL3 (see also Figure 2) yields the highest prediction accuracy for the HEPT and TIBO compounds separately. This model is based on eq 42 with only two adjustable parameters (γ and the intercept δ); all the other LIE coefficients are set to half the linear response theory predictions ($\beta = 1/4$, $\beta' = 1/2$, $\alpha = 1/2$, $\omega = 1/2$). R_{pred}^2 of MDL3 for the HEPT compounds is 0.88 and $RMSD_{pred}$ is 0.77 kcal/mol which are the best among all the models tested, including those with up to 6 adjustable parameters. The corresponding values for fitting MDL3 to the TIBO compounds are $R_{pred}^2 = 0.68$ and $RMSD_{pred} = 1.16$ kcal/mol, again the best among all the models tested. This result highlights the benefit of minimizing the number of adjustable parameters in LIE models. Increasing the number of parameters to improve the accuracy of the fit can affect the transferability of the model even among closely related ligands, such as those of the HEPT set.

Based on the jack-knife indicators, the best model for the combined HEPT+TIBO set is MDL2 ($R_{pred}^2 = 0.62$ and $RMSD_{pred} = 1.31$ kcal/mol). This model is slightly superior to the LIE model with 6 adjustable parameters we tested based on eq 42, two more than MDL2, and significantly better than MDL1. MDL3 fails completely in reproducing the binding free energies of the combined HEPT+TIBO set, as expected by the very different values of the γ LIE coefficient obtained by fitting MDL3 to the two ligand sets separately. This issue is discussed further in the next section. The values of the LIE coefficients of MDL3 for the combined

HEPT+TIBO set are similar to those of MDL2 except for β. It appears that MDL2 is able to fit the binding free energies of the combined set by employing a smaller electrostatic LIE coefficient than MDL3.

In general the models that allow for differences in the response of the receptor and water media (MDL2 and MDL3) perform better than MDL1 that does not. For example, for the HEPT compounds the RMSD of MDL2 is 0.66 kcal/mol as compared to 0.80 kcal/mol of MDL1 which has the same number of parameters. The superior performance of MDL2 over MDL1 is particularly noticeable for the combined HEPT+TIBO set. The model with the highest prediction accuracy for the HEPT and TIBO set taken separately is MDL3 which is based on the decoupling of the solution and receptor environment insertion processes. This is a further indication that the decoupling of the receptor and solution insertion processes can lead to LIE models with superior prediction accuracy.

Consistent with the linear response prediction that the electrostatic implicit solvent descriptor should be weighted twice the receptor−ligand Coulomb interaction energy, models we tested with $\beta' = \beta$ in eqs 41 and 42 perform significantly worse than MDL1 and MDL2 for both the HEPT and TIBO compounds. This is further confirmed by the results of models in which the β and β' LIE coefficients are optimized independently. In these models the optimal value of β was found to be roughly half the value of β'.

In principle, because the binding free energy is the sum of the work of inserting the ligand in the protein environment and the work of inserting the ligand in water (the hydration free energy), the LIE coefficient, ω, associated with the estimated hydration free energy should be set to 1. However we find that the accuracy of models that set $\omega = 1$ is rather poor. Much better results are obtained, as in MDL2, when ω is set to half the theoretical value. Indeed it is generally observed that, when the LIE coefficients in eq 42 are allowed to vary, they assume optimal values approximately half their theoretical values; this issue is discussed below. Two models we tested based on eq 42 which allow for optimization of ω and both ω and β' have resulted in modest gains in term of accuracy over MDL2 for which ω is fixed at one-half its theoretical value.

The TIBO set studied here is larger (37 ligands) and more diverse than previous LIE studies. The LIE models for the TIBO class of NNRTI's presented here are, to our knowledge, the best reported in terms of prediction accuracy, as measured by the jack-knife and R^2 and RMSD values. The models have few outliers and reproduce well the main trends in this class of compounds. The generally superior binding affinity of compounds with sulfur at the Z position relative to oxygen is reproduced. Analysis of the predictors of the pairs of TIBO compounds that differ only on the substitution at the Z position reveals that the C predictor (the average cavity hydration free energy of the complex relative to the that of the receptor) is mainly responsible for the ability of the LIE models to distinguish the sulfur and oxygen substituents. This points to hydrophobicity as the main cause of the stronger binding affinity of sulfur-containing TIBO inhibitors; whereas the sulfur atom at the Z position forms electrostatic and hydrogen bonding interactions with the receptor of equal magnitude as oxygen in the same position, the sulfur atom in this position is capable of burying more solvent-exposed receptor surface area. The models also reproduce the higher binding affinity obtained with the 8-Cl substitution relative to the 9-Cl substitution. We ascribe this result to the interaction between the chlorine atom at the 8 position with PHE227 also seen in crystal structures.[57] This contact causes a shift of the position of the ligand leading to increased contacts between the hydrophobic side chain of the ligand and the hydrophobic pocket formed by TYR181, TYR188, and TRP229.

MDL2 is able to fit well the combined HEPT and TIBO sets (57 ligands). Particularly encouraging are the results with MDL3. This model, based on only one adjustable parameter (excluding the intercept), has the highest prediction accuracy for the HEPT and TIBO compounds separately.

Our results suggest that the LIE formalism described here using the OPLS-AA/AGBNP effective potential improves on the results we obtained previously using the standard LIE formalism and the SGB/NP implicit solvent potential[4] and is competitive with corresponding results in explicit solvent.[19,53,55] Although differences in ligand force field parametrizations could exist, the major difference between this and previous LIE studies of HEPT and TIBO HIV-RT inhibitors is the modeling of the solvent. The AGBNP implicit solvent model does not provide a description of hydration forces with the same detail as explicit solvent models. Nevertheless molecular dynamics sampling with implicit solvent is able to explore a more diverse ensemble of conformations than a comparable calculation of the same CPU cost using an explicit solvent representation. A more complete representation of the ensemble of conformations of the ligand and the complex may be the basis for the superior results we obtained relative to similar explicit solvent studies.[19,55]

6.2. Physical Interpretation of LIE Models. The values of the β and α parameters for all models tested and the values of γ for MDL1 are positive. This implies that these models, in accordance with physical intuition, predict that ligands that have stronger interactions with the receptor and smaller desolvation penalties tend to be better binders. The seemingly counterintuitive negative value of γ obtained with MDL2 and MDL3 for the HEPT compounds is discussed below.

The LIE coefficients for MDL1 correspond to the difference between the work of inserting the ligand in the receptor and in water assuming that the same linear response proportionality coefficient applies to both processes. In the MDL2 and MDL3 models, however, the process of insertion of the ligand in the receptor is decoupled from the process of insertion in water, thus allowing for differences in the effective linear response of the receptor and water media. We see from Table 5 that the value of β differs little from one class of models to the other, suggesting that the receptor environment responds linearly to the ligand electrostatic field in a manner similar to water. A similar behavior can be observed for the α coefficient which measures the response of the water and receptor environment to the introduction of ligand−solvent and ligand−receptor van der Waals

interactions. Despite these similarities, as discussed above, the models that allow for differences in the response of the protein receptor and water (MDL2, MDL3, and related models) generally perform better.

In models such as MDL1 based on eq 41 the γ coefficient corresponds to the C−CL estimator, whereas in models based on eq 42, such as MDL2 and MDL3, it corresponds to the C descriptor, the average of the difference of the cavity hydration free energy of the complex with and without the ligand. The C descriptor approximately measures the free energy gain of burying the receptor residues lining the binding site. Thus, we would expect a positive linear correlation ($\gamma > 0$) between the C descriptor, which is always negative (see Tables 3 and 4), and the binding free energy. This is indeed the case for MDL2 and MDL3 for the TIBO compounds. However the same models applied to the HEPT compounds yield a negative γ coefficient (see Table 5). For both ligand sets the value of γ for MDL2 and MDL3 is significantly smaller than for MDL1. It appears therefore that in these models the γ coefficient absorbs physical effects correlated to the size of the binding site which oppose binding and are not directly related to the hydrophobic effect.

A likely candidate in this role is the reorganization free energy of the receptor, that is the work required to deform the binding site region to accommodate the ligand. It is reasonable to assume that the larger the binding site the larger the work required to deform the receptor structure. This contribution effectively reduces the benefit of having a large ligand−receptor contact surface area and could result in the smaller values for the γ LIE coefficients obtained for MDL2 and MDL3 as compared to MDL1, which adopts estimators that combine solution and receptor environments descriptors and is therefore less able to capture these effects. We hypothesize that for the HEPT compounds the conformational strain contribution overcomes the favorable hydrophobic desolvation effect, resulting in a negative value of γ. The prediction that the receptor reorganization free energy plays a more important role for the HEPT compounds than the TIBO compounds could also be a consequence of the wider distribution of ligand sizes of the HEPT compounds as measured by the ligand cavity formation descriptor which is proportional to the solute surface area. The range of variation of the CL descriptor of the HEPT compounds is 13.0 kcal/mol as compared 5.8 kcal/mol of the TIBO compounds (see Tables 3 and 4). The larger variation of receptor reorganization free energies embedded in the experimental binding free energies of the HEPT compounds makes the statistical fit of this term via linear regression more significant.

Although deviations from ideality of LIE parameters can also be caused by the residual average electrostatic potential created by the receptor in absence of solute charges,[42] we do not believe that this is the case in the present system. As eq 12 shows, linear response theory predicts that a nonzero residual electrostatic interaction energy ($\langle V \rangle_0$ in eq 12) tends to increase the value of the corresponding electrostatic LIE coefficient from its ideal value of 1/2. If the electrostatic LIE descriptor EC+2EL is negative (as for most HEPT compounds), this effect would appear, to an LIE model that adopts an ideal LIE electrostatic coefficient of 1/2, as an unaccounted term *favorable* to binding. It seems therefore unlikely that the residual electrostatic field of the receptor is the origin of the observed negative value of γ resulting from an unaccounted effect *unfavorable* to binding, such as the reorganization free energy of the receptor.

The results obtained with LIE models that do not include the factor of 2 for the implicit solvent electrostatic estimators are significantly inferior than the other models with the same number of adjustable parameters. It is predicted based on linear response theory that β should be half the value of β'. This is indeed approximately the case for models in which β (which corresponds to the ligand−receptor electrostatic interaction energy) is allowed to vary independently from β' (the coefficient corresponding to the implicit solvent electrostatic estimator). Contrary to linear response predictions however β is always smaller than the theoretical value of 1/2, and β' is always smaller than the theoretical value of 1. This is the case for the calculated α and ω parameters as well. It appears that the calculated parameters differ from their theoretical values by a constant proportionality factor. This would occur if a linear relationship with a proportionality coefficient different from 1 exists between the effective binding free energies calculated from the IC_{50}'s as $kT \ln IC_{50}$ and the actual binding free energies. We plan to further investigate this issue by studying complexes for which direct measurements of the binding free energies have been reported. The IC_{50}'s used in this work have been obtained from cell survival assays,[53−55] rather than enzymatic rate inhibition assays which are more directly related to the binding affinity. In cell-based assays the IC_{50} is calculated from the number of cells in a colony that remain viable after infection with the HIV-1 virus in the presence of the inhibitor. A number of environmental factors such as cell absorption and metabolism and molecular factors[62] such as stoichiometry of binding and induced RT dimerization could affect the observed effective binding free energies. The results for HIV-1 RT NNRTI systems similar to those studied here[5,55] are in general agreement with our finding that, even though the optimal values of β and α are smaller than expected, their relative magnitude is consistent with linear response predictions.

To support the hypothesis that the binding free energies calculated from the measured IC_{50}'s are proportional to the actual binding free energies, we show in Table 6 the ratios between the β, β', and ω parameters and the value of the α parameter obtained from a series of models (including those in Table 5). The theoretical values of these ratios based on linear response are $\beta/\alpha = 1/2$, $\beta'/\alpha = 1$, and $\omega/\alpha = 1$. It can be clearly seen from Table 6 that the relative magnitudes of the LIE coefficients are in good agreement with the theoretical predictions even though their absolute values do not. Although this result indicates that the effective free energies of binding derived from the IC_{50}'s deviate from the actual binding free energies by a constant proportionality factor, further investigations are needed to confirm this hypothesis.

The constant proportionality factor between the theoretical and calculated LIE coefficients appears to be close to 1/2.

Table 6: Ratios between the LIE Parameter β, β', and ω and the van der Waals LIE Parameter α Extracted from the Fits to the Experimental HEPT and TIBO Binding Data for the Model Listed Compared to Linear Response Theoretical Values

	HEPT set			TIBO set		
model	β/α	β'/α	ω/α	β/α	β'/α	ω/α
theory	0.50	1	1	0.50	1	1
MDL1	0.55			0.67		
MDL1a[a]	0.79	1.18		0.67	1.37	
MDL2	0.47		1.11	0.51		1.11
MDL2a[b]	0.45		1.07	0.55		1.24
MDL2b[c]	0.68	1.06	1.18	0.64	1.18	1.24

[a] Eq 41. [b] Eq 42 with $\beta' = 2\beta$. [c] Eq 42.

The values of the α, β', and ω coefficients are consistently near 0.5 (see Table 5), whereas the theoretical value for these parameters based on linear response is 1. Models in which ω is set to 1 perform significantly worse (data not shown) than MDL2 in which ω is set to 1/2 (Table 5). Similarly, the calculated value of the β coefficients is close to 1/4 as compared to the theoretical value of 1/2. Based on these observations we have attempted to construct a minimally parameterized LIE model for both the HEPT and TIBO compounds by setting the LIE coefficients to their theoretical linear response values reduced by a factor of 1/2. The remaining two parameters for which we do not have a firm theoretical prediction (the parameter γ corresponding to the cavity descriptor C, and the intercept δ) were allowed to vary to fit to the experimental binding free energies. This model (MDL3 in Table 5) is the most successful in predicting the binding free energies of the HEPT and TIBO compounds separately as measured by the jack-knife correlation coefficient R_{pred}^2 and root-mean-square deviation $RMSD_{pred}$.

A noticeable difference between the LIE coefficients of model MDL3 for the HEPT and TIBO compounds is the value of γ, which is negative ($\gamma = -0.18$) for the HEPT compounds and positive ($\gamma = 0.19$) for the TIBO compounds. The high statistical significance of this difference is highlighted by the poor results obtained when fitting MDL3 to the combined HEPT and TIBO set. In this fit the γ coefficient is close to zero, an intermediate value between the value appropriate for the HEPT compounds and that appropriate for the TIBO compounds, and, as a result, the model is unable to fit accurately the experimental binding free energies. As discussed above we believe that this difference is due to the reorganization free energy of the receptor which opposes binding of the HEPT analogues more than the TIBO analogues, whereas hydrophobicity favors association relatively equally. Based on this result we predict that a hypothetical increase of ligand size, which leaves electrostatic and van der Waals receptor−ligand interactions unchanged, would have opposite effects on the binding affinities of the HEPT and TIBO analogues. For the TIBO analogues the hydrophobic gain due to the increase in the amount of receptor surface area buried by a larger ligand overcomes the opposing effect due the increase of receptor reorganization free energy, leading to stronger binding. For the HEPT analogues, instead, the hydrophobic gain is more than offset by the increase of receptor reorganization free energy, leading to weaker binding. In practice, however, because it is not possible to modify the size of the ligand without also affecting electrostatic and van der Waals receptor−ligand interaction energies, all of the energetic components of the LIE model need to be considered to accurately predict the effect of ligand modifications.

The success we obtained with MDL3 in predicting the binding free energies of the HEPT and TIBO compounds based on theoretically derived parameters and only one adjustable LIE coefficient (excluding the intercept) raises the question of whether this result is a direct consequence of the receptor and solution environments obeying linear response or simply a coincidental occurrence for the receptor system and ligand sets we analyzed. The results of previous LIE studies on a variety of ligand binding systems are mainly inconclusive on this issue, partly due to difficulty of interpreting results obtained with different LIE regression expressions.

A LIE explicit solvent study of P450cam complexes by Paulsen and Ornstein[63] has shown that values of LIE coefficients near to their theoretical values of $\alpha = 1$ and $\beta = 1/2$ reproduced well the experimental binding free energies, whereas in a related study Almlöf et al.[44] reproduced the binding free energies of a similar set of inhibitors of P450cam with a significantly smaller value of the α LIE coefficient. As noted by Amlöf et al.[44] the difference between the values of α between the two studies is due to the intercept parameter, considered only in the latter study, whose optimal value was found to depend on the hydrophobicity of the receptor site.[44] Wang et al.[64] have investigated different values of the van der Waals parameter α in conjunction with β set to 1/2 without and intercept parameter. They found that for various ligand−protein complexes the optimal value of α varied from 1 to e 0.1 depending on the hydrophobicity of the binding pocket. In a related study Wang et al.[65] have compared explicit solvent LIE predictions to rigorous free energy thermodynamic integration (TI) calculations for complexes with streptavidin and established that for these systems values of α and β near their theoretical values reproduced well the experimental binding free energies for ligands for which LIE and TI produce consistent results. Cases in which TI produced results in better agreement with the experiments could be rationalized by the inadequacy of the LIE model to properly take into account the free energy cost for reorganizing the receptor pocket.[65] Earlier LIE studies[2,9] reported smaller optimal values for the α LIE parameter when setting β to 1/2. Clearly, much remains to be done to better understand the factors that influence the values of the LIE coefficients obtained on different systems with different LIE regression equations.

It is conceivable that LIE regression models, such as those adopted in the present work, which include an estimator for the work of cavity formation[3,4] are more likely to yield electrostatic and van der Waals LIE coefficients consistent with linear response predictions. The values of electrostatic and especially van der Waals LIE coefficients obtained using LIE models without an explicit cavity formation estimator,[44] and especially those without an intercept parameter,[63,65] are

instead more likely to absorb effects that are not directly related to electrostatic and van der Waals interactions and are thus not as amenable to rationalization and generalization in terms of linear response theory. Furthermore, because solute−solvent van der Waals interaction energies are not perfectly correlated with the solute surface area,[23,66,67] surface area-based cavity estimators provide nonredundant information content in addition to the van der Waals estimator, potentially leading to better descriptive accuracy and transferability. However, although it has been generally recognized that the free energy of cavity formation in water approximately scales as the surface area of the solute,[23,38] the validity of using a surface area estimator to describe the work of ligand cavity formation in the receptor pocket remains to be fully addressed.

7. Conclusions

In this paper we have addressed a number of outstanding issues in regard to the theory and practice of LIE modeling for binding free energy prediction. We have reviewed and clarified the linear response theory on which LIE methods are based. Following linear response formalism, we derived expressions for the LIE estimators when the solvent is treated implicitly. We showed that these estimators include descriptors related to the desolvation of receptor atoms which were not considered in a previously reported implicit solvent LIE model.[4] The form of the estimators we derived are consistent with those proposed previously in the context of the LRA method and the PDLD implicit solvent models[10] and for the LIE electrostatic estimator with a GB model.[12] We have also developed a novel class of LIE models that, contrary to current practice, attempt to model independently the processes of insertion of the ligand in the receptor and in solution. These models are motivated by the fact that, potentially, the receptor and the solvent respond differently to the introduction of the ligand.

We have applied these ideas to the problem of the binding free energy estimation of a series of NNRTI inhibitors of HIV-1 RT. LIE descriptors were collected from 57 molecular dynamics simulations of HIV-1 RT complexed with 20 HEPT inhibitors and 37 TIBO inhibitors. Based on the measured binding affinities and the calculated descriptors we developed a series of LIE models. We presented results for three of these models and tested several others. The first model, MDL1, is comparable to previously reported LIE studies for NNRTI binding in both explicit and implicit solvent. Relative to these studies, the predictive accuracy of our MDL1 model is generally superior when applied to the same ligand sets, suggesting that the AGBNP implicit solvation model provides sufficient accuracy for LIE modeling. This result also indicates that the LIE estimators we derived are more appropriate to describe the energetics of the binding process in implicit solvent than previously reported alternatives. The second and third models, MDL2 and MDL3, are novel models designed to treat the hydration free energy of the ligand and the work of inserting the ligand in the receptor independently. MDL2, which has the same number of parameters as MDL1, is found to be superior to MDL1 and to LIE models developed by others for predicting the binding affinities of the HEPT and TIBO analogues. This result may indicate that LIE models that treat the two insertion processes independently can lead to better prediction accuracy. MDL3, a minimally parametrized version of MDL2 in which some parameters are set to their values predicted by linear response, is found to be superior to MDL1 and MDL2 in terms of predictive ability (jack-knife tests) for the HEPT and TIBO sets individually but not for the two sets combined. We hypothesize that this is caused by the larger sizes of some of the HEPT compounds which induce steric strain of the receptor; an effect which is not taken explicitly into account by the LIE models.

We examined the values of the LIE coefficients obtained from the regression analysis and established that, although they are smaller than expected, their relative magnitudes generally conform to linear response predictions. We are planning calculations to test the applicability of linear response to other protein−ligand binding systems by computing directly the relative free energies of inserting ligands in solution and the protein environment and comparing the results to the corresponding first order cumulant descriptor. The discovery that linear response behavior is generally applicable to protein−ligand binding could provide the basis for the development of minimally parametrized LIE models. Given the substantial reduction of adjustable parameters achievable when assuming linear response for some of the interaction energy contributions and the potential improvement in predictive ability, minimally parametrized models based on linear response, such as the ones presented here, offer a productive route to binding free energy predictions using sparse experimental binding assay data for lead optimization in structure-based drug design.

Acknowledgment. This work has been supported in part by a grant from the National Institutes of Health (GM30580). We thank Dr. Anthony Felts and Dr. Zhiyong Zhou for helpful discussions.

Appendix

In this Appendix we derive eq 30. Let us first consider the process of turning on the ligand charges in the receptor environment. Conceptually we will divide this process into two steps. First the electrostatic ligand−receptor interactions are turned on and then interactions between the ligand and the implicit solvent continuum are turned on. If we assume that the receptor responds to the perturbation as a perfect linear dielectric, the free energy change, $\Delta F_{el}^{(1)}$, corresponding to the first step is approximately

$$\Delta F_{el}^{(1)} \simeq \beta \langle V_{el} \rangle_1 \tag{46}$$

where $\langle V_{el} \rangle_1$ is the average of the ligand−receptor Coulomb interaction energy in the absence of ligand-continuum solvent electrostatic interactions, and $\beta = 1/2$ based on eq 13 under the present assumptions. Based on linear response the free energy corresponding to turning on the ligand−implicit solvent continuum electrostatic interactions is

$$\Delta F_{el}^{(2)} = \beta' \langle G_{el}^{c} - G_{el}^{c'} \rangle_c \tag{47}$$

where G_{el}^{c} is the Generalized Born energy of the receptor−ligand complex and $G_{el}^{c'}$ is the Generalized Born energy of complex when the ligand charges are turned off, and the average is taken in the ensemble in which the ligand is fully charged. To derive eq 47 we write the λ-dependent potential for this process as $U(\lambda) = U_0 + \lambda V$, where $U_0 = U_{expl} + G_{cav} + G_{vdw} + G_{el}^{c'}$ is the reference potential and $V = G_{el}^{c} - G_{el}^{c'}$ is the perturbation. When the ligand and the receptor are both assumed rigid, the free energy change for adding electrostatic solute-continuum solvent interactions reduces to $\Delta F^{(2)}_{el} = G_{el}^{c} - G_{el}^{c'}$, which is the same as eq 47 with $\beta' = 1$ and with the average replaced by the constant value of the argument. Hence, based on the discussion presented above, in the general case when the ligand and the receptor are flexible and linear response applies, we expect the optimal value of the β' coefficient in eq 47 to assume values near 1.

Thus, assuming that in eq 46 the mean ligand−receptor electrostatic interaction energy $\langle V_{el} \rangle$ is the same whether the ligand-continuum solvent electrostatic interactions are present or not and assuming that $\beta' = 2\beta$, eqs 46 and 47 can be combined to give the linear response expression for the free energy of turning on the ligand charges in the receptor environment

$$\Delta F_{el} \simeq \beta \langle V_{el} + 2(G_{el}^{c} - G_{el}^{c'}) \rangle_{c} \quad (48)$$

where the average is taken with the ligand fully interacting with the receptor and solution environments, and c′ corresponds to the state in which the ligand is uncharged. It is straightforward to show using eq 18 that the difference between G_{el}^{c} and $G_{el}^{c'}$ for a given ligand−receptor complex conformation is the sum of the GB self-energies of the ligand atoms plus the sum of the GB pair energies between ligand atoms and between ligand atoms and receptor atoms.

The expression for the LIE estimator for the free energy of adding the van der Waals ligand−receptor and ligand−water interactions to the ligand cavity involves the attractive part of the Lennard-Jones potential and the implicit solvent solute−solvent van der Waals interaction energy G_{vdw}. Following the same derivation as for the electrostatic case and assuming $\alpha \simeq 1$ for both the explicit and implicit contributions, we obtain

$$\Delta F_{vdw} = \alpha \langle V_{LJ}^{vdw} + (G_{vdw}^{c'} - G_{vdw}^{c''}) \rangle_{c'} \quad (49)$$

where V_{LJ}^{vdw} is the sum of the attractive components of the Lennard-Jones interactions[68] between ligand and receptor atoms, $G_{vdw}^{c'}$ is the solute−solvent van der Waals interaction energy of the receptor−ligand complex, $G_{vdw}^{c''}$ is the solute−solvent van der Waals interaction energy of the complex in the absence of van der Waals interactions between the ligand and the solvent, and the average is taken in the state c′ with the uncharged ligand and with full ligand−receptor and ligand−solvent van der Waals interactions.

We now consider the work of creating the ligand cavity within the receptor site. According to the scheme developed above we should consider as an LIE estimator for this process the average of the difference between the solvent potential of mean force in the presence of the solute cavity and in the absence of the solute cavity plus repulsive interactions between the ligand atoms and receptor atoms. As above the average is assumed to be taken over the ensemble of conformations generated when the ligand cavity is present. The solvent potential of mean force difference includes two components: the change in G_{cav} in going from the receptor without the ligand cavity and the receptor with the ligand cavity, and, in addition, the change of the receptor−solvent van der Waals interaction energy and the change of Generalized Born energy of the receptor caused by the increase of the receptor atoms' Born radii due to the introduction of the ligand cavity. The explicit ligand cavity−receptor interactions are modeled using the repulsive component of the Lennard-Jones potential according to the WCA decomposition.[68] Finally, assuming that linear response applies to processes involving V_{LJ}^{rep} and the van der Waals and electrostatic implicit solvent components with the same proportionality coefficients as in eqs 48 and 49, we obtain the following expression for introducing the ligand cavity in the receptor

$$\Delta F_{cav} = \alpha \langle V_{LJ}^{rep} + (G_{vdw}^{c''} - G_{vdw}^{p}) \rangle_{c''} + \beta \langle 2(G_{el}^{c''} - G_{el}^{p}) \rangle_{c''} + \gamma \langle G_{cav}^{c''} - G_{cav}^{p} \rangle_{c''} \quad (50)$$

where V_{LJ}^{rep} is the sum of repulsive Lennard-Jones interactions between the ligand and receptor atoms, $G_{vdw}^{c''}$ and $G_{el}^{c''}$ are, respectively, the solute−solvent van der Waals interaction energy and the Generalized Born energy of the complex with the ligand cavity with the ligand-environment van der Waals and electrostatic interactions turned off. G_{vdw}^{p} and G_{el}^{p} are, respectively, the solute−solvent van der Waals interaction energy and the Generalized Born energy of the receptor conformation obtained from the conformation of the complex after removal of the ligand, $G_{cav}^{c''}$ is the cavity free energy of the receptor−ligand complex, G_{cav}^{p} is the cavity free energy of the complex after removal of the ligand, and $\langle \cdots \rangle_{c''}$ indicates averaging over complex conformations with the ligand cavity.

When considering the three estimators from eqs 48−50 we now make the approximation to evaluate all of the averages in the final state in which the ligand fully interacts with the environment (state *c*). We choose to combine the repulsive and attractive WCA components, V_{LJ}^{rep} and V_{LJ}^{vdw}, of the receptor−ligand Lennard-Jones interaction energies to form the total ligand−receptor Lennard-Jones interaction energy, V_{LJ}. Also, when combining the cavity and electrostatic descriptors the $G_{vdw}^{c''}$ and $G_{el}^{c''}$ terms cancel out. We then obtain the following expression for the LIE regression equation for the free energy for creating the ligand in the receptor site

$$\Delta F_{c} \simeq \alpha \langle V_{LJ} + G_{vdw}^{c} - G_{vdw}^{p} \rangle_{c} + \beta \langle V_{el} + 2(G_{el}^{c} - G_{el}^{p}) \rangle_{c} + \gamma \langle G_{cav}^{c} - G_{cav}^{p} \rangle_{c} \quad (51)$$

which is eq 30.

References

(1) *Free Energy Calculations in Rational Drug Design;* Reddy, M. R., Erion, M. D., Eds.; Springer-Verlag: 2001.

(2) Åqvist, J.; Medina, C.; Samuelsson, J.-E. *Protein Eng.* **1994**, *7*, 385−391.

(3) Jones-Hertzog, D. K.; Jorgensen, W. L. *J. Med. Chem.* **1997**, *40*, 1539−1549.

(4) Zhou, R.; Friesner, R. A.; Ghosh, A.; Rizzo, R. C.; Jorgensen, W. L.; Levy, R. M. *J. Phys. Chem.* **2001**, *105*, 10388−10397.

(5) Marcus, R. A.; Sutin, N. *Biochim. Biophys. Acta* **1985**, *811*, 265−322.

(6) Levy, R. M.; Belhadj, M.; Kitchen, D. B. *J. Chem. Phys.* **1991**, *95*, 3627−3633.

(7) Figueirido, F.; Del Buono, G.; Levy, R. M. *Biophys. Chem.* **1994**, *51*, 235−241.

(8) Lee, F. S.; Chu, Z. T.; Bolger, M. B.; Warshel, A. *Protein Eng.* **1992**, *5*, 215−228.

(9) Hansson, T.; Åqvist, J. *Protein Eng.* **1995**, *8*, 1137−1144.

(10) Sham, Y. Y.; Chu, Z. T.; Tao, H.; Warshel, A. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 393−407.

(11) Chen, X.; Tropsha, A. *J. Chem. Theory Comput.* **2006**, *2*, 1435−1443.

(12) Carlsson, J.; Andér, M.; Nervall, M.; Åqvist, J. *J. Phys. Chem. B* **2006**, *110*, 12034−12041.

(13) De Clercq, E. *Antiviral Res.* **1998**, *38*, 153−179.

(14) Coffin, J. M.; Hughes, S. H.; Varmus, H. E. *Retroviruses at the National Center for Biotechnology Information;* Cold Spring Harbor: Cold Spring Harbor Laboratory: 1997.

(15) Janssen, P. A. J.; Lewi, P. J.; Arnold, E.; Daeyaert, F.; de Jonge, M.; Heeres, J.; Koymans, L.; Vinkers, M.; Guillemont, J.; Pasquier, E.; Kukla, M.; Ludovici, D.; Andries, K.; de Bèthune, M.-P.; Pawels, R.; Das, K.; Clark Jr., A. D.; Frenkel, Y. V.; Hughes, S. H.; Medaer, B.; De Knaep, F.; Bohets, H.; De Clerck, F.; Lampo, A.; Williams, P.; Stoffels, P. *J. Med. Chem.* **2005**, *48*, 1901−1909.

(16) Heeres, J.; de Jonge, M.; Koymans, L. M. H.; Daeyaert, F. F. D.; Vinkers, M.; Van Aken, K. J. A.; Arnold, E.; Das, K.; Kilonda, A.; Hoornaert, G. J.; Compernolle, F.; Cegla, M.; Azzam, R. A.; Andries, K.; de Bèthune, M.-P.; Azjin, H.; Pauwels, R.; Lewi, P. J.; Janssen, P. A. J. *J. Med. Chem.* **2005**, *48*, 1910−1918.

(17) Das, K.; Clark, A. D., Jr.; Lewi, P. J.; Heeres, J.; De Jonge, M. R.; Koymans, L. M.; Vinkers, H. F.; Daeyaert, F.; Ludovici, D. W.; Kukla, M. J.; De Corte, B.; Kavash, R. W.; Ho, C. Y.; Ye, H.; Lichtenstein, M. A.; Andries, K.; Pauwels, R.; De Bethune, M. P.; Boyer, P. L.; Clark, P.; Hughes, S. H.; Janssen, P. A.; Arnold, E. *J. Med. Chem.* **2004**, *47*, 2550−2560.

(18) Das, K.; Lewi, P. J.; Hughes, S. H.; Arnold, E. *Prog. Biophys. Mol. Biol.* **2005**, *88*, 209−231.

(19) Rizzo, R. C.; Udier-Blagović, M.; Wang, D.-P.; Watkins, E. K.; Kroeger Smith, M. B.; Smith, R. H., Jr.; Tirado-Rives, J.; Jorgensen, W. L. *J. Med. Chem.* **2002**, *45*, 2970−2987.

(20) Kroeger Smith, M. B.; Hose, B. M.; Hawkins, A.; Lipchock, J.; Farnsworth, D. W.; Rizzo, R. C.; Tirado-Rives, J.; Arnold, E.; Zhang, W.; Hughes, S. H.; Jorgensen, W. L.; Michejda, C. J.; Smith, R. H., Jr. *J. Med. Chem.* **2003**, *46*, 1940−1947.

(21) Himmel, D. M.; Sarafianos, S. G.; Dharmasena, S.; Hossain, M. M.; McCoy-Simandle, K.; Ilina, T.; Clark Jr., A. D.; Knight, J. L.; Julias, J. G.; Clark, P. K.; Krogh-Jespersen, K.; Levy, R. M.; Hughes, S. H.; Parniak, M. A.; Arnold, E. **2006**, submitted for publication.

(22) Florián, J.; Goodman, M. F.; Warshel, A. *J. Phys. Chem. B* **2002**, *106*, 5739−5753.

(23) Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271−6285.

(24) Pierotti, R. A. *Chem. Rev.* **1976**, *76*, 717−26.

(25) Lum, K.; Chandler, D.; Weeks, J. D. *J. Phys. Chem. B* **1999**, *103*, 4570−4577.

(26) Hummer, G.; Garde, S.; García, A. E.; Paulaitis, M. E.; Pratt, L. R. *Phys. Chem. B* **1998**, *102*, 10469−82.

(27) Carlson, H. A.; Jorgensen, W. L. *J. Phys. Chem.* **1995**, *99*, 10667−10673.

(28) Warshel, A.; Russell, S. T. *Q. Rev. Biophys.* **1984**, *17*, 283−422.

(29) Åqvist, J.; Hansson, T. *J. Phys. Chem.* **1996**, *100*, 9512−9521.

(30) Cortis, C. M.; Friesner, R. A. *J. Comput. Chem.* **1997**, *18*, 1591−1608.

(31) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. *J. Comput. Chem.* **2002**, *23*, 128−137.

(32) Lazaridis, T.; Karplus, M. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139−145.

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

(34) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517−529.

(35) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479−499.

(36) Pratt, L. R.; Chandler, D. *J. Chem. Phys.* **1977**, *67*, 3683−3704.

(37) Huang, D. M.; Chandler, D. *J. Phys. Chem. B* **2002**, *106*, 2047−2053.

(38) Ashbaugh, H. S.; Kaler, E. W.; Paulaitis, M. E. *J. Am. Chem. Soc.* **1999**, *121*, 9243−9244.

(39) *CRC Handbook of Chemistry and Physics*, 86th ed. ed.; Lide, D. E., Ed.; CRC Press: Boca Raton, FL, 2005.

(40) Tounge, B. A.; Reynolds, C. H. *J. Med. Chem.* **2003**, *46*, 2074−2082.

(41) Singh, P.; Mhaka, A. M.; Christensen, S. B.; Gray, J. J.; Denmeade, S. R.; Isaacs, J. T. *J. Med. Chem.* **2005**, *48*, 3005−3014.

(42) Almlöf, M.; Åqvist, J.; Smålas, A. O.; Brandsdal, B. O. *Biophys. J.* **2006**, *90*, 433−442.

(43) Zhang, L.; Gallicchio, E.; Friesner, R. A.; Levy, R. M. *J. Comput. Chem.* **2001**, *22*, 591−607.

(44) Almlöf, M.; Brandsdal, B. O.; Åqvist, J. *J. Comput. Chem.* **2004**, *25*, 1242−1254.

(45) Still, W. C. Tempczyk, A.; Hawley, R. C.; Hendrikson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.

(46) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578−1599.

(47) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712−3720.

(48) Schaefer, M.; Froemmel, C. *J. Mol. Biol.* **1990**, *216*, 1045−1066.

(49) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824−19839.

(50) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(51) Jorgensen, W. L.; Madura, J. D. *Mol. Phys.* **1985**, *56*, 1381.

(52) Simonson, T. *Curr. Opin. Struct. Biol.* **2001**, *11*, 243−252.

(53) Rizzo, R. C.; Tirado-Rives, J.; Jorgensen, W. L. *J. Med. Chem.* **2001**, *44*, 145−154.

(54) Ho, W.; Kukla, M. J.; Breslin, H. J.; Ludovici, D. W.; Grous, P. P.; Diamond, C. J.; Miranda, M.; Rodgers, J. D.; Ho, C. Y.; De Clercq, E.; Pauwels, R.; Andries, K.; Janssen, M. A. C.; Janssen, P. A. J. *J. Med. Chem.* **1995**, *38*, 794−802.

(55) Smith Jr., R. H.; Jorgensen, W. L.; Tirado-Rives, J.; Lamb, M. L.; Janssen, P. A. J.; Michejda, C. J.; Kroeger Smith, M. B. *J. Med. Chem.* **1998**, *41*, 5272−5286.

(56) Hopkins, A. L.; Ren, J.; Esnouf, R. M.; Willcox, B. E.; Jones, E. Y.; Ross, C.; Miyasaka, T.; Walker, R. T.; Tanaka, H.; Stammers, D. K.; Stuart, D. I. *J. Med. Chem.* **1996**, *39*, 1589−1600.

(57) Das, K.; Ding, J.; Hsiou, Y.; Clark A. D., Jr.; Moereels, H.; Koymans, L.; Andries, K.; Pauwels, R.; Janssen, P. A.; Boyer, P. L.; Clark, P.; Smith, R. H., Jr.; Kroeger Smith, M. B.; Michejda, C. J.; Hughes, S. H.; Arnold, E. *J. Mol. Biol.* **1996**, *264*, 1085−1100.

(58) Schrödinger, Inc., Portland, OR.

(59) Banks, J. L.; Beard, J. S.; Cao, Y.; Cho, A. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; Maple, J. R.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Zhang, L. Y.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2005**, *26*, 1752−1780.

(60) Smith, M. B. K.; Rouzer, C. A.; Taneyhill, L. A.; Smith, N. A.; Hughes, S. H.; Boyer, P. L.; Janssen, P. A. J. Moereels, H.; Koymans, L.; Arnold, E.; Ding, J.; Das, K.; Zhang, W.; Michejda, C. J.; Smith, R. H., Jr. *Protein Sci.* **1995**, *4*, 2203−2222.

(61) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983−10990.

(62) Cheng, Y.-C.; Prusoff, W. H. *Biochem. Pharmacol.* **1973**, *22*, 3099−3108.

(63) Paulsen, M. D.; Ornstein, R. L. *Protein Eng.* **1996**, *9*, 567−571.

(64) Wang, W.; Wang, J.; Kollman, P. A. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 395−402.

(65) Wang, J.; Dixon, R.; Kollman, P. A. *Proteins Struct. Funct. Genet.* **1999**, *34*, 69−81.

(66) Su, Y.; Gallicchio, E.; Levy, R. M. *Biophys. Chem.* **2004**, *109*, 251−260.

(67) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. *J. Am. Chem. Soc.* **2003**, *25*, 9523−9530.

(68) Weeks, J. D.; Chandler, D.; Andersen, H. C. *J. Chem. Phys.* **1971**, *54*, 5237−47.

CT600258E

JCTC Journal of Chemical Theory and Computation

Human Cytomegalovirus Protease: Why is the Dimer Required for Catalytic Activity?

César Augusto Fernandes de Oliveira,*,†,§ Cristiano Ruch Werneck Guimarães,‡,∥ Gabriela Barreiro,‡,⊥ and Ricardo Bicca de Alencastro†

Physical Organic Chemistry Group, Departamento de Química Organica, Instituto de Química, Universidade Federal do Rio de Janeiro, Cidade Universitária, CT, Bloco A, lab. 609, Rio de Janeiro, RJ 21949-900, Brazil, and Department of Chemistry, Yale University, 225 Prospect St., New Haven, Connecticut 06520-8107

Received May 18, 2006

Abstract: Human cytomegalovirus (HCMV) is a pathogenic agent responsible for morbidity and mortality in immunocompromised and immunosuppressed individuals. HCMV encodes a serine protease that is essential for the production of infectious virions. In this work, we applied molecular dynamics (MD) simulations on HCMV protease models in order to investigate the experimentally observed (i) catalytic activity of the enzyme homodimer and (ii) induced-fit mechanism upon the binding of substrates and peptidyl inhibitors. Long and stable trajectories were obtained for models of the monomeric and dimeric states, free in solution and bound covalently and noncovalently to a peptidyl-activated carbonyl inhibitor, with very good agreement between theoretical and experimental results. The MD results suggest that HCMV protease indeed operates by an induced-fit mechanism. Also, our analysis indicates that the catalytic activity of the dimer is a result of more favorable interactions between the oxyanion in the covalently bound state and the backbone nitrogen of Arg165, resulting in a reaction that is 7.0 kcal/mol more exergonic and a more significant thermodynamic driving force. The incipient oxyanion in the transition state should also benefit from the stronger interactions with Arg165, reducing in this manner the intrinsic activation barrier for the reaction in the dimeric state.

Introduction

Herpesviruses are responsible for several health problems in humans and in most species in the animal kingdom.[1] All of the herpesviruses are members of one family, the Herpesviridae, and have certain common characteristics such as their ability to establish latency during primary infection.[2] The human herpesviruses have been classified into three subfamilies designated α, β, and γ. The α subfamily includes herpes simplex virus type 1, herpes simplex virus type 2, and the varicella-zoster virus. The β-herpesviruses include human cytomegalovirus (HCMV) and human herpesviruses 6 and 7. The γ subfamily, which is specific for either B or T lymphocytes, includes Epstein−Barr and Kaposi's sarcoma-associated herpesvirus, also known as human herpesvirus 8.[3]

HCMV is a highly species-specific DNA virus infecting up to 80% of the general population, though most of these infections are clinically asymptomatic.[4] HCMV is a leading opportunistic infectious pathogen in individuals with suppressed or compromised immune systems, causing several health problems such as pneumonia, retinitis, and death.[4,5]

* Corresponding author fax: (858) 534-4974; e-mail: cesar@mccammon.ucsd.edu.
† Universidade Federal do Rio de Janeiro.
‡ Yale University.
§ Current address: Department of Chemistry and Biochemistry, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093.
∥ Current address: Amgen, Inc., 1120 Veterans Boulevard, South San Francisco, CA 94080.
⊥ Current address: Department of Pharmaceutical Chemistry, University of California, San Francisco, 600 16th St., San Francisco, CA 94143.

10.1021/ct600175x CCC: $37.00
Published on Web 11/29/2006

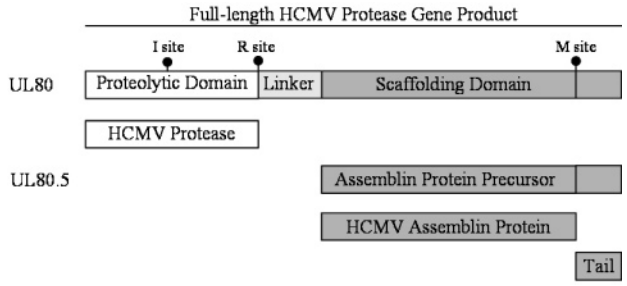


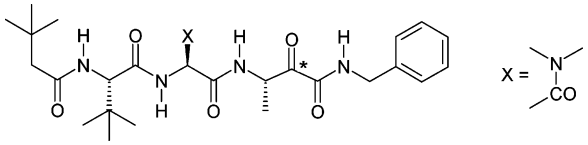
Figure 1. Maturational processing of HCMV protease and the assembly protein.

Because the number of organ transplantations and people infected with HIV has increased considerably in the past decades, the search for effective antiviral treatments is of vital importance.[5,6] HCMV is also very common in neonates, infecting 1% of newborn infants. It is estimated that 10% of these are symptomatic and may experience severe sequelae such as mental retardation and hearing disturbances.[5,7] Current treatment options for HCMV infections in clinical use include ganciclovir, valganciclovir, foscarnet, cidofovir, and formivirsen.[5] Although these HCMV DNA-polymerase inhibitors potently reduce viral replication, they exhibit toxicity and require intravenous administration to attain therapeutic levels. Moreover, viral resistance to current drugs is becoming an increasingly serious problem.[8,14]

All members of the herpesvirus family encode a protease that participates in the maturation of the viral capsid, whose activity is essential for the production of infectious virions.[15−17] Although it has only been demonstrated for herpes simplex virus type 1,[17] on the basis of homology, the protease is assumed to be essential in other herpesviruses. The extensive number of recently published articles aimed at obtaining inhibitors of the HCMV protease shows that this enzyme became an attractive target for the development of anti-herpes agents.[4,7,15,18−26]

The viral genome contains the open reading frame UL80, which encodes the full-length 80 kDa herpesvirus protease and its substrate.[15,27] Full-length HCMV protease is composed of an N-terminal 256-amino-acid proteolytic domain, a linker region, and a C-terminal structural domain, the scaffolding domain (Figure 1). In HCMV, the assemblin protein precursor (UL80.5 gene product) interacts with the major capsid protein and acts as a scaffold in the nucleus around which the capsid assembles.[28,29] To allow for the packaging of viral DNA, the full-length protease UL80 gene product undergoes autoproteolytic cleavage at the maturation site (M-site) and the release site (R-site) of the scaffolding domain. In the case of HCMV, the protease catalyzes an additional cleavage at the I-site.[30,31] The M-site, near the carboxy terminus, removes the carboxy tail from the scaffolding domain and from the assemblin protein precursor, while the R-site releases the N-terminal fragment (proteolytic domain) from the full-length protease gene product. The proteolytic domain fragment retains all the catalytic activity of the full-length protease protein and is generally referred to as HCMV protease.[32] The I-site produces a two-chain form of the protease that is still catalytically active.[30,31]

Chart 1. Peptidyl-Activated Carbonyl Inhibitor Used in This Work[a]



[a] The starred atom corresponds to the carbon attacked by Ser132.

Biochemical and mutagenesis studies have identified the herpesvirus protease as a serine protease, though its sequence does not bear any homology to other serine proteases or other proteins in general.[15,33] Its unusual amino acid sequence and biochemical properties establish the herpesvirus protease as a new class of serine proteases.[34−40] In HCMV, X-ray crystallographic studies revealed that the protease possesses a unique polypeptide backbone fold and catalytic triad;[38−41] the third member is a His residue (Ser132, His63, and His157), whereas classical serine proteases have either an Asp or Glu residue.[42−46] In addition, this enzyme has been shown to operate by an induced-fit mechanism, in contrast to classical serine proteases where substrate binding has generally little impact on the protein conformation.[47−50] Finally, although crystallographic data indicates that each monomeric subunit possesses a well-separated and complete active site, biochemical studies showed that dimerization is required for catalytic activity.[47,51,52]

In this work, we performed long molecular dynamics (MD) simulations on HCMV protease models to study the experimentally observed induced-fit mechanism operated by the enzyme. More specifically, we aimed at investigating if the binding of a peptidyl-activated carbonyl inhibitor (Chart 1),[47] prior to covalent adduct formation, induces conformational changes on the enzyme that would bring it to a catalytically active form. Another goal was to understand the structural and energetic factors responsible for the catalytic activity of the enzyme dimer. The current study is a more complete account of the work previously published by our group.[50]

Computational Details

The crystal structure of HCMV protease in complex with the peptidomimetic inhibitor BILC 821 (PDB code: 2WPO) was used to build our protease models.[48] The crystal structure contains residues 4−46, 53−143, 152−200, and 210−256 for each of four monomers of the protease belonging to two separate dimers with one inhibitor bound covalently to each protease monomer. Because we are currently interested in determining the structural and energetic requisites that govern the activation of the HCMV protease dimeric form and in studying the conformational changes that occur in the protein prior to and after the covalent adduct formation, six different models have been considered: the dimeric and monomeric forms free in solution (**D** and **M**, respectively) and complexed noncovalently (**DI** and **MI**) and covalently (**D-I** and **M-I**) to a peptidyl-activated carbonyl inhibitor (Chart 1), very similar to BILC 821.[48] The 26 residues of the HCMV protease not observed in the monomers of the crystal structure are in the following regions: 1−3 at the N terminus of the enzyme, which form a small helix (named αN), 47−52 at loop L3, 144−151 at loop L9, and 201−209 at loop

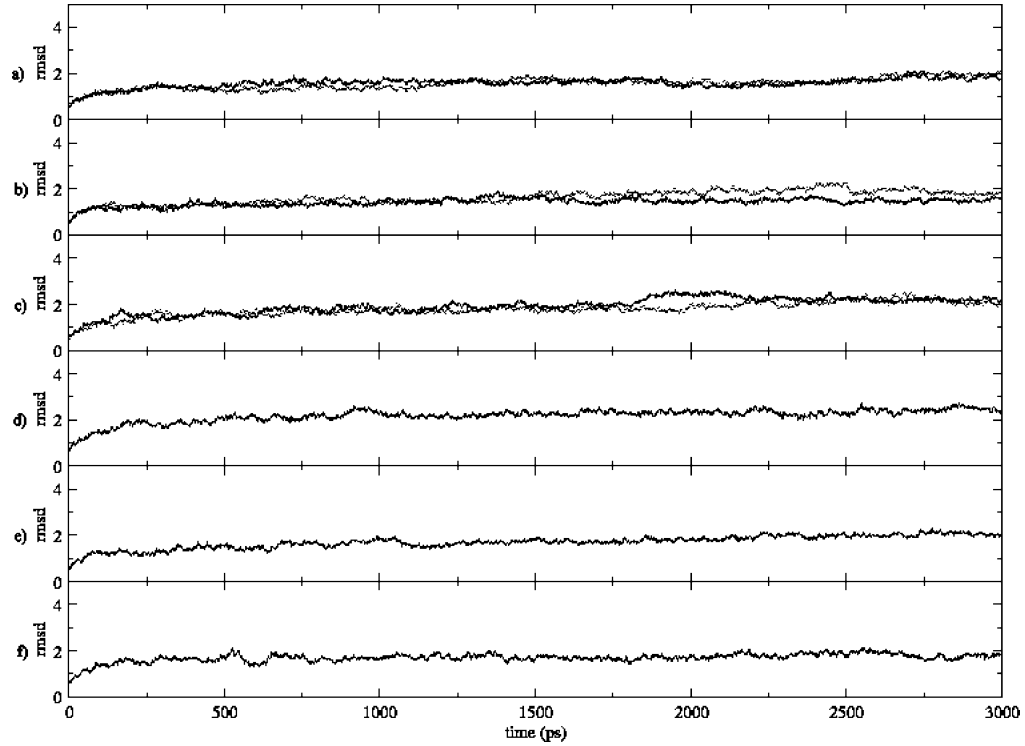


Figure 2. RMSD calculated along the MD trajectories for (a) **D**, (b) **DI**, (c) **D-I**, (d) **M**, (e) **MI**, and (f) **M-I**. In the a, b, and c plots, the black and gray lines correspond to monomers A and B, respectively.

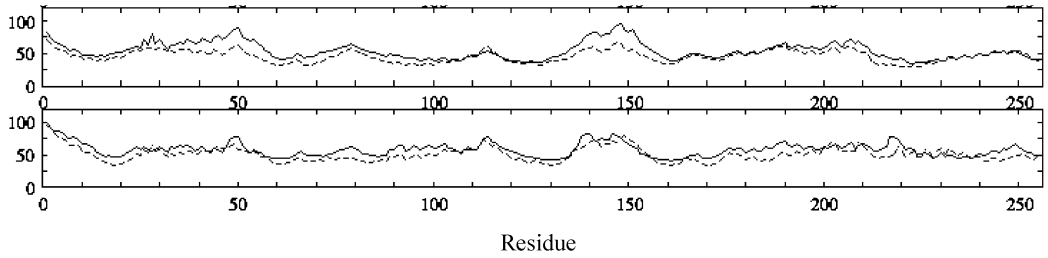


Figure 3. (Top) Calculated B factors for monomer A of **DI** (solid line) and monomer A of **D** (dashed line). (Bottom) Calculated B factors for **MI** (solid line) and **M** (dashed line).

L13. To obtain the complete set of coordinates, we added the missing residues to both monomers of **D-I**; this structure originated the starting points of MD trajectories of 3 ns for each model. Details for the model building and MD simulations can be found in ref 62. All calculations were performed using the AMBER7 package with the Cornell et al. force field.[53,54]

Results and Discussion

The Induced-Fit Mechanism. To assess the stability of the simulations, the root-mean-squared deviation (RMSD) of the Cα atoms for all models with respect to the corresponding starting structure was monitored (Figure 2). The results show that stable trajectories were obtained. Most importantly, the calculated isotropic temperature factor (*B*-factor) from the MD trajectories of **D** and **D-I** agree very well with the experimental data as previously reported.[50]

To analyze the changes in atomic fluctuations between the bound and unbound forms of the dimer and monomer of the protease, the calculated *B*-factors for **DI** and **D** and **MI** and **M** were plotted in Figure 3. In our models, the complex formation increases the mobility of atoms in different regions of the enzyme, but this is much more pronounced for the dimeric state. More specifically, residues 35−90 in αA, L3, β2, L4, β3, L5, and β4 and residues 130−160 in β5, L9, and β6, which are located around the active site, are more ordered in the free enzyme than in the noncovalent complex. The increase in fluctuations observed for these regions is associated with a loss of hydrogen bonds between protein residues. Figure 4 shows the residues 35−90 and 130−160 and the hydrogen bonds (dashed lines) with occupancy greater than 50% that are present in the **D** model but not in **DI**. This demonstrates that the inhibitor binding, before the formation of the covalent bond, has a significant impact on the protein dynamics.

Significant conformational changes were observed experimentally in the dimer upon covalent binding of peptidyl-activated carbonyl inhibitors.[48,49] As shown in our previous work,[50] the comparison between the average structures for the monomers A of **DI** and **D** indicates that the conformational changes are not associated with the adduct formation, but rather with the formation of the noncovalent complex.

Tong and co-workers[48] verified that, if the unbound and covalently bound dimeric states are superimposed using one

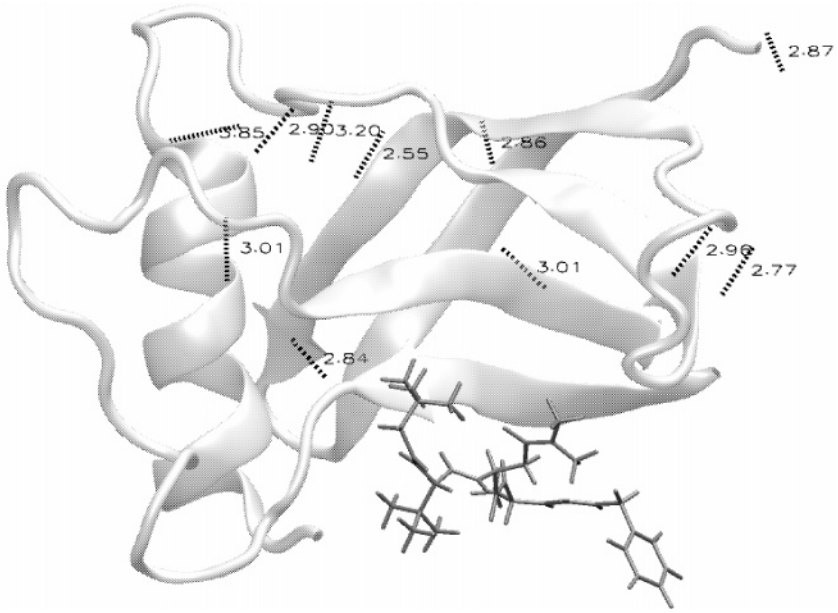


Figure 4. Cartoon representation of segments 35−90 and 130−160 for the **D** model. Dashed lines correspond to hydrogen bonds with occupancy greater than 50% found in the MD trajectory of **D**, but not in **DI**. To better localize the active site, the **D** and **DI** models were superimposed and only the inhibitor in **DI** is displayed.

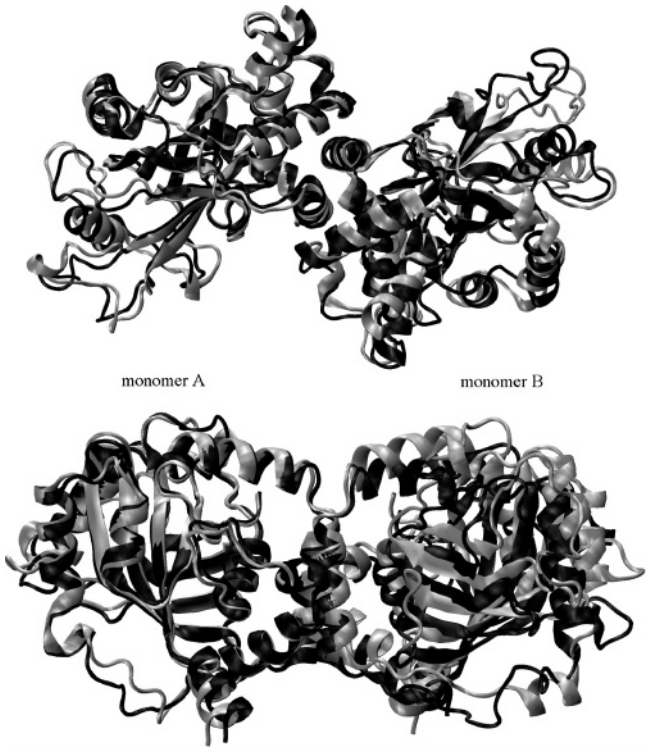


Figure 5. Superposition of the average structures obtained for **DI** (black) and **D** (gray) using monomer A as reference.

of the monomers as a reference, a rotation of 6.5° around an axis mostly perpendicular to the dimer 2-fold axis is needed to bring the second monomer into an overlapping position. This experimental observation was reproduced in our MD simulations of **D** and **DI**. It should be noted that the trajectories for **D** and **DI** were obtained using the crystal structure for the covalent complex as the starting point, which suggests that the sampling was adequate. Figure 5 shows the superposition of the average structures for **D** and **DI** using monomer A as a reference.

The cause of the change in the dimer organization is currently unknown. In order to investigate it, the following geometric parameters were monitored for the **DI** and **D** trajectories: (i) the distance between the center of mass of each monomer and (ii) the angle between the vectors V_1 and V_2, defined by the distance between the center of mass of each monomer in **DI** (V_1) and **D** (V_2) (Figure 6e)—in both models, the vector origins are the center of mass of monomer A; (iii) the angle defined by the centers of mass of monomer A, helices αF, and monomer B (AαFB; Figure 6f, left); and (iv) the dihedral angle defined by the center of mass of monomer A, the Cα atoms of Ser225 of helices αF of both monomers, and the center of mass of monomer B (Figure 6f, right).

Figure 6a shows that both **DI** and **D** have nearly the same distance between the centers of mass. Figure 6b shows that the angle between the vectors V_1 and V_2 oscillates around 6.0°. The noncovalent complex formation induces a rotation of monomer B around an axis mostly perpendicular to the dimer 2-fold axis, as shown in Figure 6c. After ca. 1 ns, the gap between the angles AαFB for **DI** and **D** oscillates around 5.0°, in excellent agreement with the experimental value of 6.5°. Finally, as can be seen in Figure 6d, the dihedral angle as defined above also changes upon noncovalent binding. Interestingly, the changes for this dihedral angle and the angle AαFB seem to be strongly correlated. As the angle AαFB for **D** becomes more linear, its dihedral angle becomes more planar. As the angle AαFB and the dihedral angle get closer to 180°, an increase in the distance between the centers of mass of monomers A and B should occur; this was not observed in the MD simulations. Therefore, a translation of monomer B with respect to monomer A must accompany

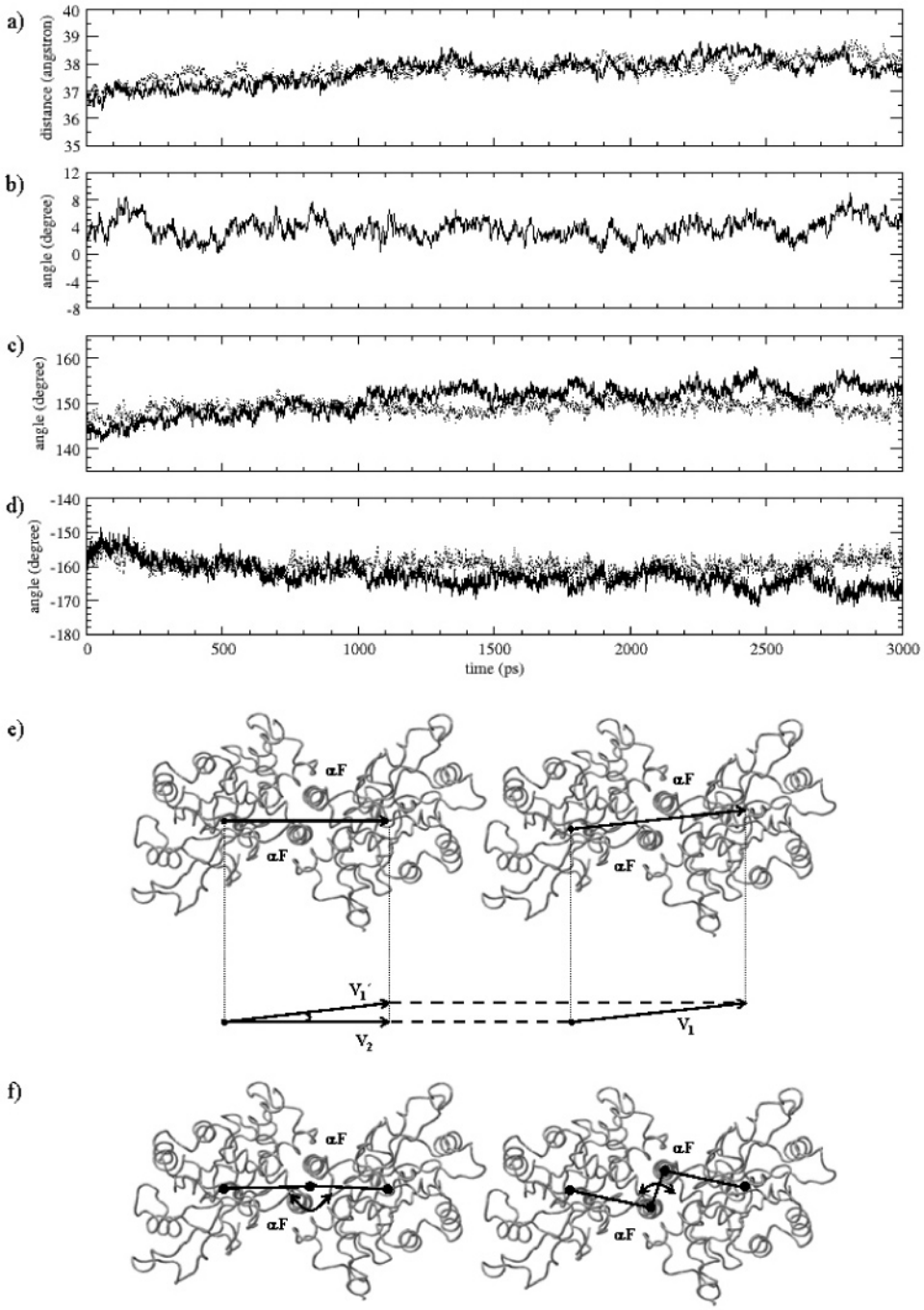


Figure 6. Plots corresponding to the time dependence of the following geometric parameters: (a) distance between the center of mass of each monomer; (b) angle between vectors V_1 and V_2, defined by the distance between the center of mass of each monomer in **DI** (V_1) and **D** (V_2); (c) angle defined by the centers of mass of monomer A, helices αF, and monomer B; (d) dihedral angle defined by the center of mass of monomer A, the Cα atoms of Ser225 of helices αF of both monomers, and the center of mass of monomer B; (e) schematic representation of vectors V_1, V_1', and V_2; (f) left, schematic representation of angle defined in c and, right, schematic representation of the dihedral angle defined in d. In graphics a, c, and d, solid and dashed lines correspond to the **D** and **DI** models, respectively.

the angle AαFB and the dihedral angle motions in order to keep the distance between the centers of mass roughly constant.

The changes in protein organization and conformation observed experimentally for the dimer covalently bound to BILC 821 were well-reproduced by the MD simulations of the noncovalent complex model, before the formation of the tetrahedral intermediate. This suggests that the HCMV protease operates by an induced-fit mechanism.[47]

Why is the Dimer Required for Catalytic Activity? Dynamic cross correlation matrix ($\mathbf{C}_{ij}$) was applied to study the collective motions related to the monomer−monomer interactions.[55−60] $\mathbf{C}_{ij}$ elements were obtained from their respective covariance matrix elements c_{ij} given by eq 1,

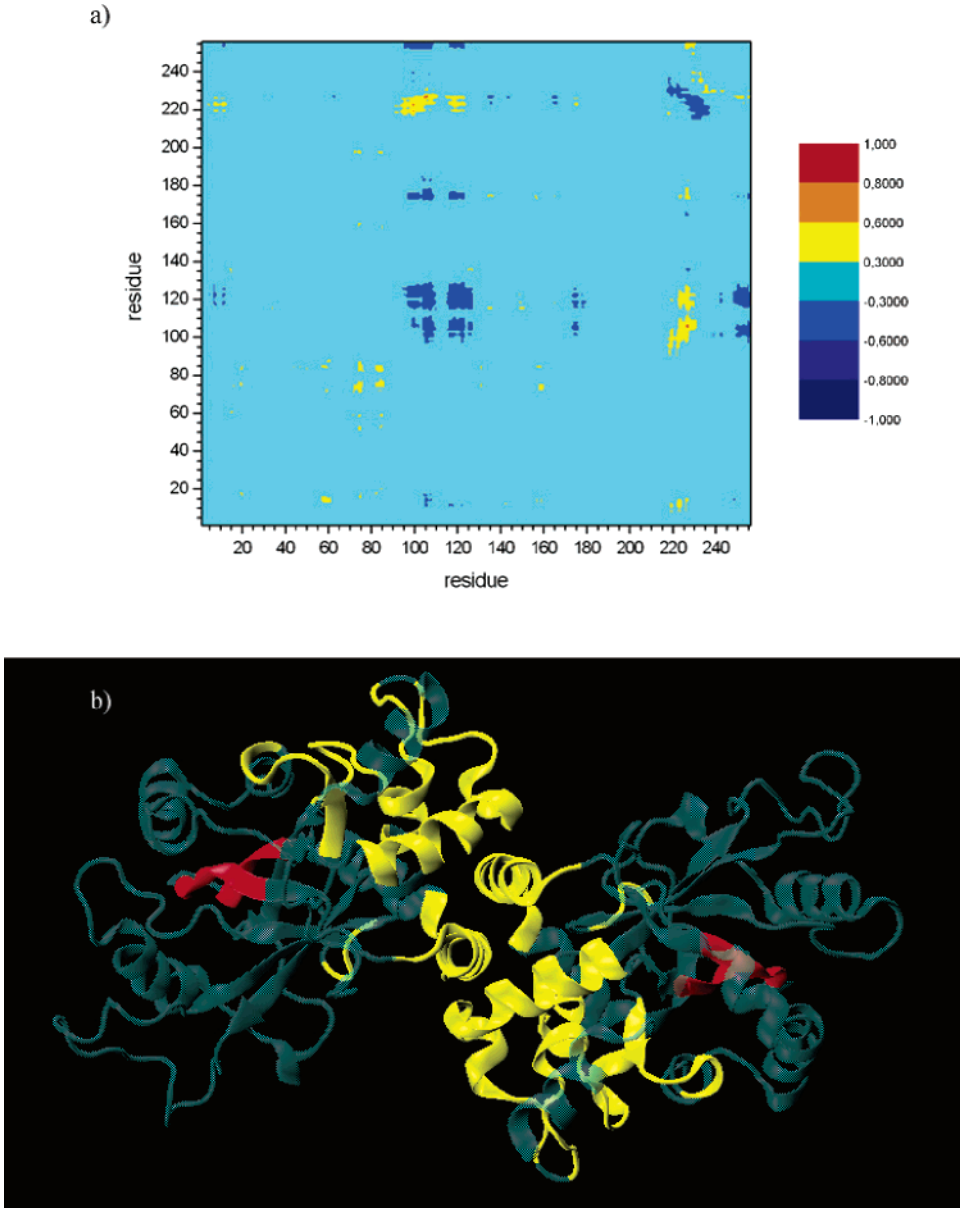


Figure 7. (a) Bidimensional representation of the dynamic cross correlation matrix elements calculated for **D**. Horizontal and vertical axes correspond to residues from monomers A and B, respectively. (b) Protease regions with intermonomer correlated motions are shown in yellow (interface) and red (active site).

where r_i is the atom i Cartesian coordinate vector and $< >$ represents a time average.

$$c_{ij} = \langle (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle) \rangle = \langle r_i \cdot r_j \rangle - \langle r_i \rangle \langle r_j \rangle \qquad (1)$$

When the c_{ij} terms are expanded, where t_{aver} is the time used to calculate the average value and Δt is the time interval between two consecutive configurations, eq 2 is obtained.

$$c_{ij} = \frac{\Delta t}{t_{aver}} \left\{ \sum_{t=0}^{t_{aver}-\Delta t} r_i(t)\, r_j(t) - \frac{\Delta t}{t_{aver}} \left[\sum_{t=0}^{t_{aver}-\Delta t} r_i(t)\, r_j(t) \right] \times \left[\sum_{t=0}^{t_{aver}-\Delta t} r_j(t)\, r_j(t) \right] \right\} \qquad (2)$$

Finally, the elements of the dynamic cross-correlation matrix, or normalized covariance, are defined as in eq 3.

$$\mathbf{C}_{ij} = \frac{c_{ij}}{c_{ii}^{1/2} c_{jj}^{1/2}} = \frac{\langle r_i \cdot r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{[(\langle r_i^2 \rangle - \langle r_i \rangle^2)(\langle r_j^2 \rangle - \langle r_j \rangle^2)]^{1/2}} \qquad (3)$$

$\mathbf{C}_{ij}$ values vary from -1 to $+1$. Positive and negative values represent motions that are correlated and anticorrelated, respectively. The closer the value is to $+1$ or -1 for a pair of residues, the more correlated or anticorrelated their motions are, respectively. The dynamic cross-correlation matrix was calculated by averaging six blocks of 200 ps, with an offset of 50 ps, along the last 1.5 ns of the MD trajectories. To prevent the inclusion of spurious and artificially correlated motions,[55] the following residues were used in the fitting procedure: 13−23, 57−61, 67−78, 81−89, 130−134, 156−161, and 169−174 in the $\beta 1$−$\beta 7$ sheets

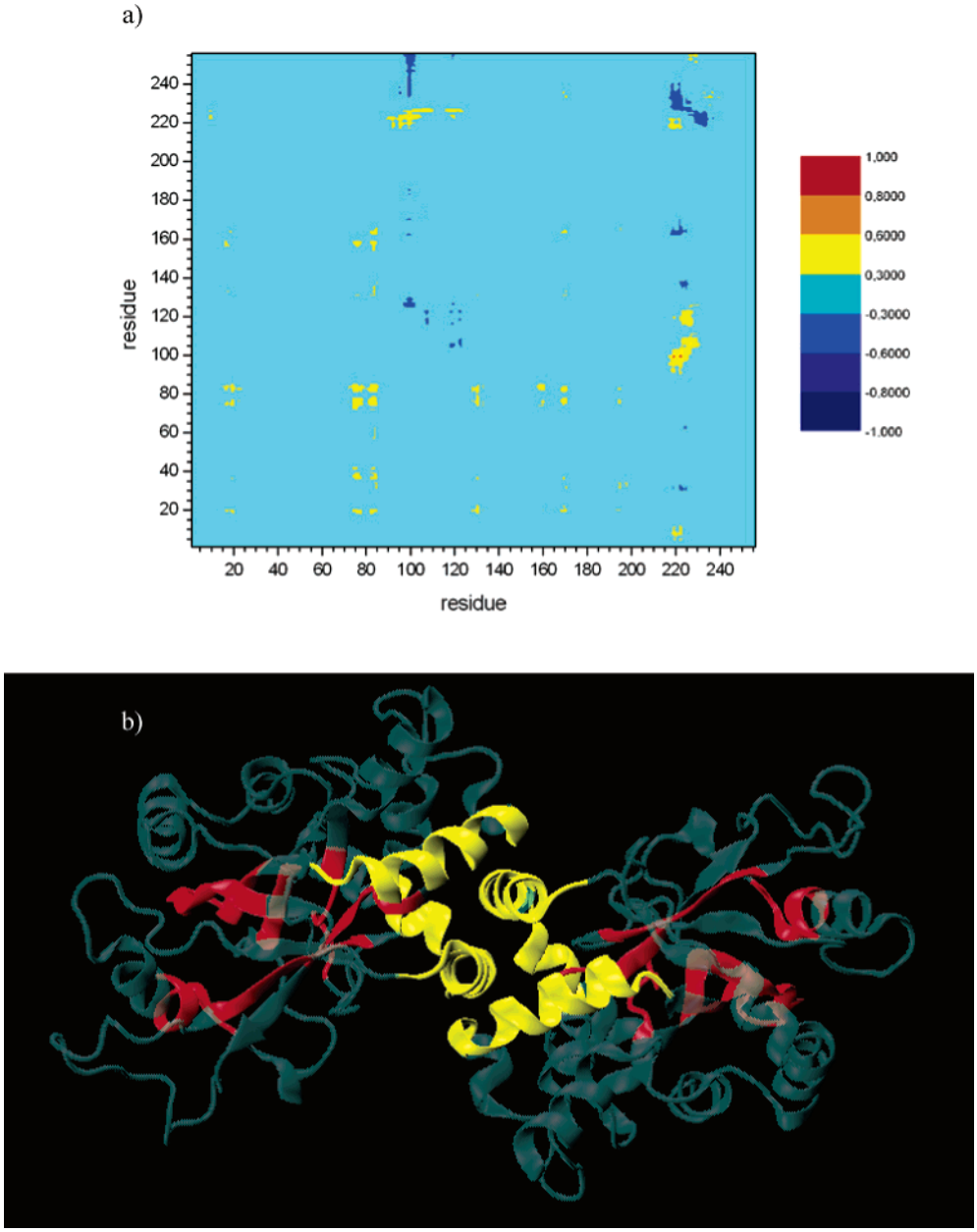


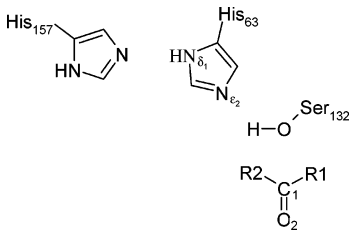
Figure 8. (a) Bidimensional representation of the dynamic cross correlation matrix elements calculated for **DI**. Horizontal and vertical axes correspond to residues from monomers A and B, respectively. (b) Protease regions with intermonomer correlated motions are shown in yellow (interface) and red (active site).

of both monomers. These residues are in the least flexible region of the protease.

Figures 7 and 8 show the dynamic cross-correlation matrix and the most relevant regions with intermonomer correlated motions obtained for the **D** and **DI** models, respectively. As expected, the intermonomer motions for residues close to the interface are strongly correlated (residues in yellow). Moreover, it was possible to identify correlated motions between residues from both active sites, which are separated by more than 30 Å (residues in red). Interestingly, while the correlated motions between residues located near the dimer interface decrease in **DI**, the noncovalent binding increases the number of intermonomer correlated motions between the active sites (Figures 7 and 8). Previous works claim that correlated and anticorrelated motions in the active site region may reduce the activation barrier through an increase in the number of reactive conformations sampled by the noncovalently bound complex.[56,61]

In order to investigate that, we monitored the interactions for the subsystem composed by the inhibitor and the catalytic triad in the noncovalently bound complex models. In both **DI** and **MI**, Ser132 is hydrogen-bonded to His63; the calculated distance between the Ser132 side-chain oxygen and His63 nitrogen $N\epsilon_2$ oscillates around 2.8 Å, while the angle $Ser_{132}O-Ser_{132}H-His_{63}N\epsilon_2$ fluctuates moderately around 170° (Chart 2). For both models, the His63 nitrogen $N\delta_1$ is hydrogen-bonded to His157 with occupancies of 62% for **DI** and 66% for **MI**. To verify whether Ser132 maintains its reactive orientation along the trajectories of **DI** and **MI**, we monitored the distance and the attacking angle between the Ser132 side-chain oxygen and the inhibitor's activated carbonyl group. In both models, the calculated distance

Chart 2. Representation of the Interactions Monitored for the Subsystem Composed by the Inhibitor and the Catalytic Triad in **MI** and **DI**



($Ser_{132}O-C_1$) and the angle ($Ser_{132}O-C_1-O_2$) fluctuate slightly around 2.9 Å and 95°, respectively (Chart 2). This seems to indicate that the presence of monomer−monomer interactions and correlated motions are not crucial to increase the number of reactive conformations and for the catalytic triad to orient properly in order to attack the inhibitor. Although the correlated motions between the active sites might not influence the activation barrier as discussed above, they still could have an impact in the reaction rate for the dimer by increasing its transmission factor. In other words, the presence of a network of coupled motions throughout the dimer would increase the probability of sampling conformations conducive to the catalyzed chemical reaction.[62]

In the crystal structure for the covalent complex between the protease and BILC 821, Arg165 interacts with the resulting oxyanion through its backbone NH. In addition, it has been suggested that the oxyanion is further stabilized by a water-mediated hydrogen bond with the side chain of Arg166.[63] Arg165 and Arg166 are highly conserved among all herpesvirus proteases. Liang and co-workers have shown through site-directed mutagenesis that, while the substitution of Arg166 by alanine has led to ablation of the enzymatic activity without detectable change in the HCMV protease conformation, the substitution of Arg165 by alanine revealed only a 2.7-fold reduction in activity.[63]

The role of these residues in the catalytic activity of the dimer was investigated by monitoring their interactions with the activated carbonyl oxygen of the inhibitor along the trajectories for **DI**, **D-I**, **MI**, and **M-I**. For Arg165, while the average distance between its backbone nitrogen and the carbonyl oxygen is ca. 2.9 Å for all models, the angle for this hydrogen bond changes significantly for **D-I**. Figure 9 shows that the angle distributions for the noncovalently and covalently bound complexes for the monomer are essentially the same. In the dimeric state, however, the angle distribution becomes narrower and its peak is shifted to more linear values upon formation of the oxyanion. A per-residue interaction energy analysis revealed that the more favorable interactions between the oxyanion and Arg165 translate into a reaction that is 7.0 kcal/mol more exergonic for the dimer, indicating a more significant thermodynamic driving force. The incipient oxyanion in the transition state should also benefit from the stronger interactions with Arg165, reducing the intrinsic barrier in the dimeric state. Quantum mechanics/molecular mechanics (QM/MM) simulations performed by our group show that the reaction is 6.6 kcal/mol more exothermic and the barrier is reduced by 8.1 kcal/mol in the dimeric state. The results also show that Arg165 is primarily responsible for it. This study will be published elsewhere.

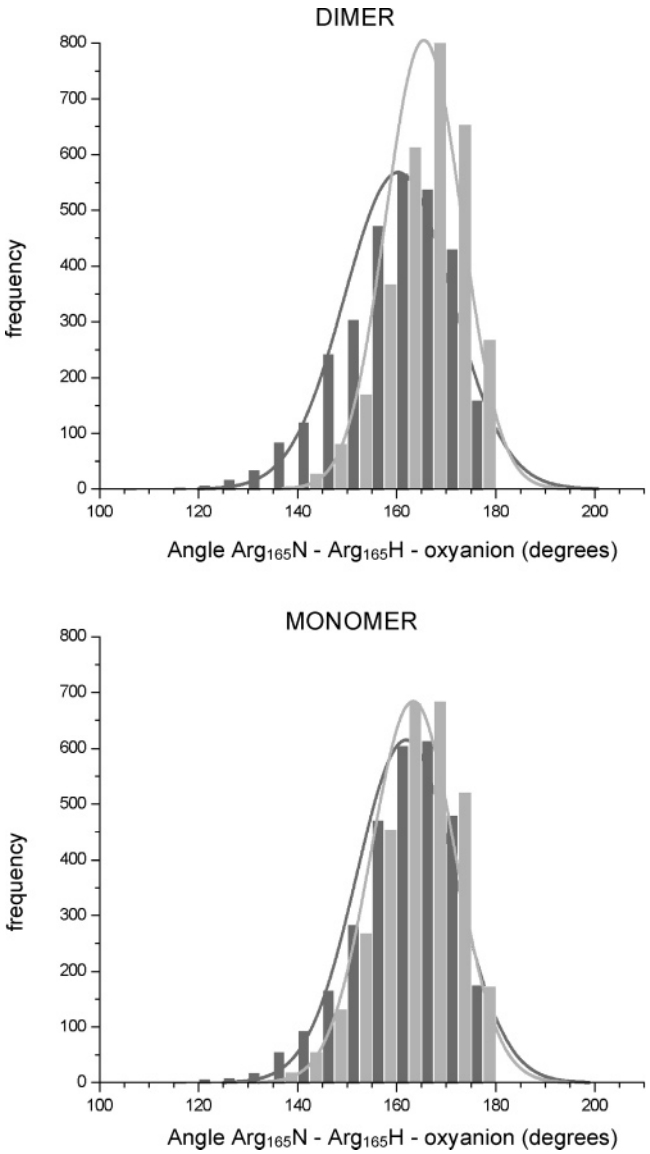


Figure 9. Arg_{165}**N**−Arg_{165}**H**−oxyanion angle distribution for the noncovalent (gray) and covalent complexes (light gray).

For Arg166, the water-mediated hydrogen bond between its side chain and the oxyanion has not been characterized in our MD simulations. In fact, no hydrogen bonds between Arg166 and the oxyanion were formed in any models (Figure 10). Also, the per-residue interaction energy analysis did not show any significant contribution from this residue to the catalytic activity of the dimer. Although our results are at odds with the results from mutagenesis experiments, the latter should be interpreted carefully. For example, the Arg166Ala mutation might abolish the enzymatic activity of the dimer not because important interactions between Arg166 and the oxyanion were lost but because of perturbations introduced by this mutation in the interactions with Arg165. In conclusion, our results suggest that only the dimeric form of the protease is able to reorient the main-chain atoms of Arg165 along the reaction coordinate in order to stabilize more efficiently the oxyanion formed in the reaction pathway.

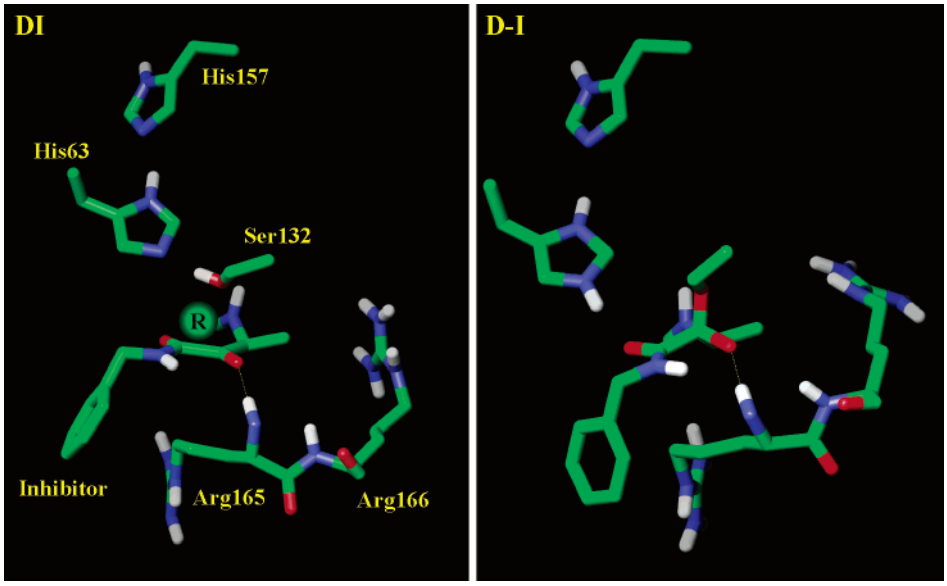


Figure 10. Noncovalently and covalently bound complexes for the dimeric state. The catalytic triad and the interactions between the inhibitor and Arg165 are illustrated. Parts of the inhibitor and other residues were excluded for clarity.

Conclusions

In this work, we used MD simulations to investigate the experimentally observed catalytic activity of the HCMV protease homodimer and the induced-fit mechanism upon the binding of substrates and peptidyl inhibitors. Long and stable trajectories were obtained for models of the monomeric and dimeric states, free in solution and bound covalently and noncovalently to a peptidyl-activated carbonyl inhibitor. Conformational changes observed experimentally for the dimer upon covalent binding were reproduced by the MD simulations of the noncovalent complex model, demonstrating that the HCMV protease operates by an induced-fit mechanism.

Dynamic cross-correlation analysis revealed that the noncovalent binding increases intermonomer correlated motions between residues of both active sites. The presence of a network of coupled motions throughout the enzyme could have a positive impact in the reaction rate for the dimer through the increase of its transmission factor. Finally, our results indicate that the catalytic activity of the dimer is a result of more favorable interactions between the oxyanion in the covalently bound state and the backbone nitrogen of Arg165, resulting in a reaction that is 7.0 kcal/mol more exergonic and a more significant thermodynamic driving force. The intrinsic activation barrier for the reaction in the dimeric state should also be reduced as the incipient oxyanion in the transition state would likely benefit from the stronger interactions with Arg165. This is supported by QM/MM simulations performed by our group, which shows that the activation barrier is reduced by 8.1 kcal/mol in the dimeric state. The results for this study will be published elsewhere.

Acknowledgment. This research has received partial financial support from the Brazilian agencies CNPq and FAPERJ (Grants 570535/1997-2 and E26-152174). C.A.F.O. acknowledges CNPq/Brazil and FAPERJ/Brazil for Ph. D. fellowships.

References

(1) Corey, L.; Spear, P. G. N. Infections with Herpes-Simplex Viruses 1. *N. Engl. J. Med.* **1986**, *314*, 686−691.

(2) Roizman, B.; Desrosiers, R. C.; Fleckenstein, B.; Lopez, C.; Minson, A. C.; Studdert, M. J. The Family Herpesviridae: An Update. *Arch. Virol.* **1992**, *123*, 425−449.

(3) Holwerda, B. C. Herpesvirus Proteases: Targets for Novel Antiviral Drugs. *Antiviral Res.* **1997**, *35*, 1−21.

(4) de Jong, M. D.; Galasso, G. J.; Gazzard, B.; Griffiths, P. D.; Jabs, D. A.; Kern, E. R.; Spector, S. A. Summary of the II International Symposium on Cytomegalovirus. *Antiviral Res.* **1998**, *39*, 141−162.

(5) Nichols, W. G; Boeckh, M. Recent Advances in the Therapy and Prevention of CMV Infections. *J. Clin. Virol.* **2000**, *16*, 25−40.

(6) Bruggeman, C. A.; Marjorie, H. J.; Nelissen-Vrancken, G. Cytomegalovirus and Atherogenesis. *Antiviral Res.* **1999**, *43*, 133−144.

(7) Jong, M. D.; Boucher, C. A. B.; Danner, S. A.; Gazzard, B.; Griffiths, P. D.; Katlama, C.; Lange, J. M. A.; Richman, D. D.; Vella, S. Summary of the International Consensus Symposium on Management of HIV, CMV and Hepatitis Virus Infections. *Antiviral Res.* **1998**, *37*, 1−16.

(8) Gilbert, C.; Boivin, G. Human Cytomegalovirus Resistance to Antiviral Drugs. *Antimicrob. Agents Chemother.* **2005**, *49*, 873−883.

(9) De Clercq, E. Antiviral Drugs in Current Clinical Use. *J. Clin. Virol.* **2004**, *30*, 115−133.

(10) De Clercq, E. New Inhibitors of Human Cytomegalovirus (HCMV) on the Horizon. *J. Antimicrob. Chemother.* **2003**, *51*, 1079−1083.

(11) De Clercq, E. Strategies in the Design of Antiviral Drugs. *Nat. Rev. Drug Discovery* **2002**, *1*, 13−25.

(12) De Clercq, E. Antiviral Drugs: Current State of the Art. *J. Clin. Virol.* **2001**, *22*, 73−89.

(13) Eizuru, Y. Development of New Antivirals for Herpesviruses. *Antiviral Chem. Chemother.* **2003**, *14*, 299−308.

(14) Emery, V. C. Progress in Understanding Cytomegalovirus Drug Resistance. *J. Clin. Virol.* **2001**, *21*, 223−228.

(15) Gibson, W.; Welch, A. R.; Hall, M. R. T. Assembling, a Herpes Virus Serine Maturational Proteinase and New Molecular Target for Antivirals. *Perspect. Drug Discovery Des.* **1994**, *2*, 413−426.

(16) Matusick-Kumar, L.; McCann, P. J.; Robertson, B. J.; Newcomb, W. W.; Brown, J. C.; Gao, M. Release of the Catalytic Domain N(o) from the Herpes Simplex Virus Type 1 Protease Is Required for Viral Growth. *J. Virol.* **1995**, *69*, 7113−7121.

(17) Gao, M.; Matusick-Kumar, L.; Hurlburt, W.; DiTusa, S. F.; Newcomb, W. W.; Brown, J. C.; McCann, P. J.; Deckman, I.; Colonno, R. J. The Protease of Herpes-Simplex Virus Type-1 Is Essential for Functional Capsid Formation and Viral Growth. *J. Virol.* **1994**, *68*, 3702−3712.

(18) Gopalsamy, A.; Lim, K.; Ellingboe, J. W.; Mitsner, B.; Nikitenko, A.; Upeslacis, J.; Mansour, T. S.; Olson, M. W.; Bebernitz, G. A.; Grinberg, D.; Feld, B.; Moy, F. J.; O'Connell, J. Design and Syntheses of 1,6-Naphthalene Derivatives as Selective HCMV Protease Inhibitors. *J. Med. Chem.* **2004**, *47*, 1893−1899.

(19) Borthwick, A. D.; Angier, S. J.; Crame, A. J.; Exall, A. M.; Haley, T. M.; Hart, G. J.; Mason, A. M.; Pennel, A. M. K; Weingarten, G. G. Design and Synthesis of Pyrrolidine-5,5-trans-lactams (5-Oxo-hexahydro-pyrrolo[3,2-b]pyrroles) as Novel Mechanism-Based Inhibitors of Human Cytomegalovirus Protease. 1. The Alpha-methyl-trans-lactam Template. *J. Med. Chem.* **2000**, *43*, 4452−4464.

(20) Dhanak, D.; Burton, G.; Christmann, L. T.; Darcy, M. G.; Elrod, K. C.; Kaura, A.; Keenan, R. M.; Link, J. O.; Peishoff, C. E.; Shah, D. H. Metal Mediated Protease Inhibition: Design and Synthesis of Inhibitors of the Human Cytomegalovirus (hCMV) Protease. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 2279−2282.

(21) Dhanak, D.; Keenan, R. M.; Burton, G.; Kaura, A.; Darcy, M. G.; Shah, D. H.; Ridgers, L. H.; Breen, A; Lavery, P.; Tew, D. G.; West, A. Benzothiopyran-4-one Based Reversible Inhibitors of the Human Cytomegalovirus (HCMV) Protease. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 3677−3682.

(22) Déziel, R.; Malefant, E. Inhibition of Human Cytomegalovirus Protease N-o with Monocyclic Beta-Lactams. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 1437−1442.

(23) Yoakim, C.; Ogilvie, W. W.; Cameron, D. R.; Chabot, C.; Guse, I.; Haché, B.; Naud, J.; O'Meara, J. A.; Plante, R.; Déziel, R. Beta-Lactam Derivatives as Inhibitors of Human Cytomegalovirus Protease. *J. Med. Chem.* **1998**, *41*, 2882−2891.

(24) Pinto, I. L.; West, A.; Debouck, C. M.; DiLella, A. G.; Gorniak, J. G.; O'Donnel, K. C.; O'Shannessy, D. J.; Patel, A.; Jarvest, R. L. Novel, Selective Mechanism-Based Inhibitors of the Herpes Proteases. *Bioorg. Med. Chem. Lett.* **1996**, *6*, 2467−2472.

(25) Patick, A. K.; Potts, K. E. Protease Inhibitors as Antiviral Agents. *Clin. Microbiol. Rev.* **1998**, *11*, 614−627.

(26) Flynn, D. L.; Abood, N. A.; Holwerda, B. C. Recent Advances in Antiviral Research: Identification of Inhibitors of the Herpesvirus Proteases. *Curr. Opin. Chem. Biol.* **1997**, *1*, 190−196.

(27) Baum, E. Z.; Bebernitz, G. A.; Hulmes, J. D.; Muzithras, V. P.; Jones, T. R.; Gluzman, Y. Expression and Analysis of the Human Cytomegalovirus-UL80-Encoded Protease Identification of Autoproteolytic Sites. *J. Virol.* **1993**, *67*, 497−506.

(28) Gibson, W. Action at the Assemblin Dimer Interface. *Nat. Struct. Biol.* **2001**, *8*, 739−741.

(29) LaFemina, R. L.; Bakshi, K.; Long, W. J.; Pramanik, B.; Veloski, C. A.; Wolanski, B. S.; Marcy, A. I.; Hazuda, D. J. Characterization of a Soluble Stable Human Cytomegalovirus Protease and Inhibition by M-Site Peptide Mimics. *J. Virol.* **1996**, *70*, 4819−4824.

(30) Holwerda, B. C.; Wittwer, A. J.; Duffin, K. L.; Smith, C.; Toth, M. V.; Carr, L. S.; Wiegand, R. C.; Bryant, M. L. Activity of 2-Chain Recombinant Human Cytomegalovirus Protease. *J. Biol. Chem.* **1994**, *269*, 25911−25915.

(31) O'Boyle, D. R.; Wager-Smith, K.; Stevens, J. T.; Weinheimer, S. P. The Effect of Internal Autocleavage on Kinetic-Properties of the Human Cytomegalovirus Protease Catalytic Domain. *J. Biol. Chem.* **1995**, *270*, 4753−4758.

(32) Tong, L. Viral Proteases. *Chem. Rev.* **2002**, *102*, 4609−4626.

(33) Welch, A. R.; McNally, L. M.; Hall, M. R. T.; Gibson, W. Herpesvirus Proteinase - Site-Directed Mutagenesis Used to Study Maturational, Release, and Inactivation Cleavage Sites of Precursor and to Identify a Possible Catalytic Site Serine and Histidine. *J. Virol.* **1993**, *67*, 7360−7372.

(34) Reiling, K. K.; Pray, T. R.; Craik, C. S.; Stroud, R. M. Functional Consequences of the Kaposi's Sarcoma-Associated Herpesvirus Protease Structure: Regulation of Activity and Dimerization by Conserved Structural Elements. *Biochemistry* **2000**, *39*, 12796−12803.

(35) Hoog, S. S.; Smith, W. W.; Qiu, X. Y.; Janson, C. A.; Hellmig, B.; McQueney, M. S.; ODonnell, K.; OShannessy, D.; DiLella, A. G.; Debouck, C.; AbdelMeguid, S. S. Active Site Cavity of Herpesvirus Proteases Revealed by the Crystal Structure of Herpes Simplex Virus Protease/Inhibitor Complex. *Biochemistry* **1997**, *36*, 14023−14029.

(36) Qiu, X.; Janson, C. A.; Culp, J. S.; Richardson, S. B.; Debouck, C.; Smith, W. W.; Abdel-Meguid, S. S. Crystal Structure of Varicella-Zoster Virus Protease. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94*, 2874−2879.

(37) Chen, P.; Tsuge, H.; Almassy, R. J.; Gribskov, C. L.; Katoh, S.; Vanderpool, D. L.; Margosiak, S. A.; Pinko, C.; Matthews, D. A.; Kan, C. Structure of the Human Cytomegalovirus Protease Catalytic Domain Reveals a Novel Serine Protease Fold and Catalytic Triad. *Cell* **1996**, *86*, 835−843.

(38) Qiu, X.; Culp, J. S.; DiLella, A. G.; Hellmig, B.; Hoog, S. S.; Janson, C. A.; Smith, W. W.; Abdel-Meguid, S. S. Unique Fold and Active Site in Cytomegalovirus Protease. *Nature* **1996**, *383*, 275−279.

(39) Shieh, H. S.; Kurumbail, R. G.; Stevens, A. M.; Stegeman, R. A.; Sturman, E. J.; Pak, J. Y.; Wittwer, A. J.; Palmier, M. O.; Wiegand, R. C.; Holwerda, B. C.; Stallings, W. C. Three Dimensional Structure of Human Cytomegalovirus Protease. *Nature* **1996**, *383*, 279−282.

(40) Tong, L.; Qian, C.; Massariol, M-J.; Bonneau, P.; Cordingley, M.; Lagacé, L. A New Serine-Protease Fold Revealed by the Crystal Structure of Human Cytomegalovirus Protease. *Nature* **1996**, *383*, 272−275.

(41) Khayat, R.; Batra, R.; Massariol, M.-J.; Lagacé, L.; Tong, L. Investigating the Role of Histidine 157 in the Catalytic Activity of Human Cytomegalovirus Protease. *Biochemistry* **2001**, *40*, 6344−6351.

(42) Hedstrom, L. Serine Protease Mechanism and Specificity. *Chem. Rev.* **2002**, *102*, 4501−4523.

(43) Dereweda, Z. S.; Dereweda, U.; Kobos, P. M. (His)C−Epsilon−H···O═C Hydrogen-Bond in the Active-Sites of Serine Hydrolases. *J. Mol. Biol.* **1994**, *241*, 83−93.

(44) Carter, P.; Wells, J. A. (His)C−Epsilon−H···O═C Hydrogen-Bond in the Active-Sites of Serine Hydrolases. *Nature* **1988**, *332*, 564−568.

(45) Neurath, H. Evolution of Proteolytic-Enzymes. *Science* **1984**, *224*, 350−357.

(46) Kraut, J. Serine Proteases - Structure and Mechanism of Catalysis. *Annu. Rev. Biochem.* **1977**, *46*, 331−358.

(47) LaPlante, S. R.; Bonneau, P. R.; Aubry, N.; Cameron, D. R.; Deziel, R.; Grand-Maitre, E.; Plouffe, C.; Tong, L.; Kawai, S. H. Characterization of the Human Cytomegalovirus Protease as an Induced-Fit Serine Protease and the Implications to the Design of Mechanism-Based Inhibitors. *J. Am. Chem. Soc.* **1999**, *121*, 2974−2986.

(48) Tong, L.; Qian, C.; Massariol, M.-J.; Déziel, R.; Yoakim, C.; Lagacé, L. Conserved Mode of Peptidomimetic Inhibition and Substrate Recognition of Human Cytomegalovirus Protease. *Nat. Struct. Biol.* **1998**, *5*, 819−826.

(49) Bonneau, P. R.; Grand-Maître, C.; Greenwood, D. J.; Lagacé, L.; LaPlante, S. R.; Massariol, M.-J.; Ogilvie, W. W.; O'Meara, J. A.; Kawai, S. H. Evidence of a Conformational Change in the Human Cytomegalovirus Protease upon Binding of Peptidyl-Activated Carbonyl Inhibitors. *Biochemistry* **1997**, *36*, 12644−12652.

(50) de Oliveira, C. A. F.; Guimarães, C. R. W.; Barreiro, G.; Alencastro, R. B. Investigation of the Induced-Fit Mechanism and Catalytic Activity of the Human Cytomegalovirus Protease Homodimer via Molecular Dynamics Simulations. *Proteins* **2003**, *52*, 483−491.

(51) Batra, R.; Khayat, R.; Tong, L. Molecular Mechanism for Dimerization to Regulate the Catalytic Activity of Human Cytomegalovirus Protease. *Nat. Struct. Biol.* **2001**, *8*, 810−817.

(52) Darke, P. L.; Cole, J. L.; Waxman, L.; Hall, D. L.; Sardana, M. K.; Kuo, L. C. Active Human Cytomegalovirus Protease is a Dimer. *J. Biol. Chem.* **1996**, *271*, 7445−7449.

(53) Case, D. A.; Perlman, D. A.; Caldwell, J. W.; Chetham, T. E., III; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Gohlke, H.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *AMBER 7*; University of California: San Francisco, CA, 1995.

(54) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(55) Hunenberger, P. H.; Mark, A. E.; Gunsteren, W. F. Fluctuation and Cross-Correlation Analysis of Protein Motions Observed in Nanosecond Molecular-Dynamics Simulations. *J. Mol. Biol.* **1995**, *252*, 492−503.

(56) Estabrook, R. A.; Luo, J.; Purdy, M. M.; Sharma, V.; Weakliem, P.; Bruice, T. C.; Reich, N. O. Statistical Coevolution Analysis and Molecular Dynamics: Identification of Amino Acid Pairs Essential for Catalysis. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 994−999.

(57) Mazumder-Shivakumar, D.; Bruice, T. C. Computational Study of IAG-Nucleoside Hydrolase: Determination of the Preferred Ground State Conformation and the Role of Active Site Residues. *Biochemistry* **2005**, *44*, 7805−7817.

(58) Zoete, V.; Meuwly, M.; Karplus, M. A Comparison of the Dynamic Behavior of Monomeric and Dimeric Insulin Shows Structural Rearrangements in the Active Monomer. *J. Mol. Biol.* **2004**, *342*, 913−929.

(59) Harte, W. E., Jr.; Swaminathan, S.; Beveridge, D. Molecular Dynamics of HIV-1 Protease. *Proteins* **1992**, *13*, 175−194.

(60) 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.

(61) Luo, J.; Bruice, T. C. Ten-Nanosecond Molecular Dynamics Simulation of the Motions of the Horse Liver Alcohol Dehydrogenase. PhCH2O- complex. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 16597−16600.

(62) Hammes-Schiffer, S.; Benkovic, S. J. Relating Protein Motion to Catalysis. *Annu. Rev. Biochem.* **2006**, *75*, 519−541.

(63) Liang, Po-H.; Brun, K. A.; Feild, J. A.; O'Donnell, K.; Doyle, M. L.; Green, S. M.; Baker, A. E.; Blackburn, M. N.; Abdel-Meguid, S. S. Site-Directed Mutagenesis Probing the Catalytic Role of Arginines 165 and 166 of Human Cytomegalovirus Protease. *Biochemistry* **1998**, *37*, 5923−5929.

CT600175X

JCTC Journal of Chemical Theory and Computation

Density Functionals for Noncovalent Interaction Energies of Biological Importance

Yan Zhao* and Donald G. Truhlar*

Department of Chemistry and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455-0431

Received August 18, 2006

Abstract: Forty density functionals and one wavefunction method are assessed against a recently published database of accurate noncovalent interaction energies of biological importance. The comparison shows that two newly developed density functional theory (DFT) methods, PWB6K and M05-2X, give the best performance for this benchmark database of 22 noncovalent complexes, including both hydrogen-bonding and dispersion-dominated complexes. In contrast, the more popular B3LYP and PBEh functionals fail to describe the interactions in the dispersion-dominated complexes. The local spin density approximation and BHandH functionals give good performance for dispersion-dominated interactions at the expense of a large error for hydrogen bonding. PWB6K and M05-2X constitute a new generation of DFT methods based on simultaneously optimized exchange and correlation functionals that include kinetic energy density in both the exchange and correlation functional, and the present study confirms that they have greatly improved performance for noncovalent interactions as compared to previous DFT methods. We interpret this as being due to an improved treatment of medium-range correlation effects by the exchange-correlation functional. We recommend the PWB6K and M05-2X methods for investigating large biological systems and soft materials.

1. Introduction

Noncovalent interactions[1−10] play very important roles in biological science for problems such as protein folding and nucleobase packing and stacking. An accurate description of noncovalent interactions is also a key to predicting ligand binding and structures in proteins, DNA, and RNA. The accurate description of noncovalent interactions is also one of the challenges in modeling solvation, supramolecular chemistry, and soft materials. However, the available computational methods are not entirely satisfactory. On one hand, state-of-the-art correlated wave function theory [WFT; for example, CCSD(T)[11]] is prohibitively expensive to apply to the complex systems of interest. On the other hand, the more affordable density functional theory (DFT)[12−14] has only been widely validated for its capabilities to treat covalent interactions such as heat of formation and atomization energies, and validations are less complete for noncovalent interactions. Furthermore, until very recently,[15−21] available density functionals were too inaccurate for many demanding problems including noncovalent interactions.[22−26]

The first step toward improving the functionals is to assess the quality of existing functionals to establish a baseline. To accomplish this, we developed[27] a database for noncovalent interactions that contains six hydrogen-bonded complexes (HB6/04), seven charge-transfer complexes (CT7/04), six dipole-interaction complexes (DI6/04), and nine weak-interaction complexes (WI9/04). More recently, we split the WI9/04 data and added more data to create a WI7/05 data set that excludes $\pi-\pi$ stacking plus a separate data set of five $\pi-\pi$ stacking complexes (PPS5/05).[28] Merging the HB6, CT7, DI6, WI7, and PP5 data sets gives a noncovalent database (NC31/05) that is composed of small complexes involving first-row and second-row elements. This has been employed to test DFT[27] and other model chemistry methods.[29] We have also employed this noncovalent database (along with other data including covalent interactions, barrier

* To whom correspondence should be addressed: yzhao@chem.umn.edu (Y.Z.); truhlar@umn.edu (D.G.T.).

10.1021/ct6002719 CCC: $37.00
Published on Web 12/10/2006

heights, and ionization potentials) to develop new DFT functionals, such as PWB6K[17] and M05-2X.[21] In a short communication,[19] we compared the performances of six DFT methods for the prediction of interaction energies of two hydrogen-bonded Watson−Crick base pairs, six stacked nucleobase pairs, and five amino acid residue pairs, and we found that PWB6K gives very good performance. In the present study, we expand the work in that communication, and we assess the capability of 40 density functionals and one level of WFT, namely Møller−Plesset second-order perturbation theory (MP2),[30] for predicting interaction energies against a benchmark noncovalent data set recently proposed by Jurecka et al.[31] Although the 40 functionals considered here represent a wide variety of functional types [local spin density,[32] generalized gradient approximation (GGA),[33−42] hybrid,[39−41,43−49] meta GGA,[50−52] and hybrid meta GGA[17,21,48,50,51,53−56]], none of them model the asymptotic dipolar nature of dispersion interactions explicitly. Thus, these functionals can be accurate at the distances of van der Waals minima but not in the long-range limit that would be important, for example, for small-angle scattering of rare gases in molecular beams.[57] Another class of functionals that considers the long-range functional forms explicitly is also being developed,[58−61] and these functionals are very promising. Nevertheless, they are beyond our scope in the present study.

Since 2003, the research groups of Hobza and Sponer have published several papers on accurate stabilization energies of hydrogen-bonded and stacked DNA and RNA base pairs[9,62−66] as well as amino acid residue pairs.[67] Recently, they merged all these data into a data set, called JCSH-2005.[31] The JSCH-2005 set consists of more than 100 DNA base pairs, amino acid residue pairs, and model complexes which are of biological importance. They also proposed a smaller screening set (S22) of 22 model complexes to quickly assess the quality of theoretical models. In the present study, the S22 database is employed to test 40 DFT functionals and one level of WFT: MP2. It is particularly appropriate to include MP2 in this study because it is widely used to study noncovalent interactions, when affordable (for moderate-size and large systems, it is considerably more expensive than DFT).

Section 2 describes the S22 database and the computational methods used in the present work. Section 3 presents results and discussion, and section 4 has concluding remarks.

2. Database and Computational Methods

2.1. S22 Database. The S22 database is a data set of 22 weakly bonded molecular complexes of biological importance. This database was developed by Jurecka et al., who divided the S22 set into three subsets, namely, seven hydrogen-bonded complexes, eight dispersion-dominated complexes (it might have been preferable to call the interactions in these complexes dispersion-like because some theorists define dispersion interactions only in the long-range limit, but we will use their label for consistency), and seven mixed complexes. The reference interaction energies for the S22 data set were calculated by the following scheme:

$$\Delta E^{\mathrm{CCSD(T)}}\,\mathrm{CBS} = \Delta E^{\mathrm{MP2}}\,\mathrm{CBS} + (\Delta E^{\mathrm{CCSD(T)}} - \Delta E^{\mathrm{MP2}})_{\text{small basis}} \quad (1)$$

where a complete basis set (CBS) limit CCSD(T) interaction energy is approximated by a CBS MP2 interaction energy plus a difference between CCSD(T) and MP2 interaction energies ($\Delta E^{\mathrm{CCSD(T)}} - \Delta E^{\mathrm{MP2}}$) evaluated with a relatively small basis set that was specifically designed[1,62] for this purpose. This is based on the assumption that the ($\Delta E^{\mathrm{CCSD(T)}} - \Delta E^{\mathrm{MP2}}$) term has faster convergence than $\Delta E^{\mathrm{CCSD(T)}}$ with respect to the basis set, and thus this difference can be evaluated with a small or mid-sized basis set. This assumption has been validated for some model hydrogen-bonded[68] and stacked model complexes.[62] The best estimates of the interaction energies in the S22 database were taken from the paper by Jurecka et al.,[31] and they are listed in the Supporting Information (Table S1). The structures of these noncovalent complexes are shown in Figures 1−3.

2.2. Computational Methods. All DFT calculations were carried out using a locally modified *Gaussian 03*[69] program. The tested density functionals are detailed in Table 1. In each case, we specify the year that the functional was first published, the functional forms used for dependence on spin density (ρ_σ, where σ is the component of spin angular momentum along an axis) and the spin density gradient ($\nabla\rho_\sigma$), whether or not the functional includes spin kinetic energy density (τ_σ) in the exchange and correlation functionals, and whether the correlation functional is self-correlation-free (SCorF). Table 2 also contains two columns (one for the exchange functional and one for the correlation functional) that tell whether or not the functional reduces to the correct uniform electron gas (UEG) limit when $\nabla\rho_\sigma \rightarrow 0$ and $\tau_\sigma \rightarrow \tau_\sigma^{\mathrm{LSDA}}$ (where $\tau_\sigma^{\mathrm{LSDA}}$ is the value assumed by τ_σ in the UEG limit; LSDA stands for local spin density approximation) and another column that tells the percentage X of Hartree−Fock (HF) exchange in the functional.

Because many conventional DFT functionals (e.g., B3LYP and X3LYP) fail[23] to predict the minima of the stacked complexes in the S22 database, we test all 40 density functionals with the best estimated geometries specified in Table S1, and these geometries were obtained from the Supporting Information of the paper by Jurecka et al.[31] We also performed geometry optimization with the best performing functional, M05-2X, and these will be discussed separately (in section 3.6).

We used an augmented polarized valence double-ζ basis set labeled DIDZ [which is denoted 6-31+G(d,p)][70] for most of the calculations (all calculations except those discussed in section 3.5) We use the 6-31+G(d,p) basis set because the goal of the present paper is not to obtain benchmark interaction energies for these noncovalent complexes or study the prediction of the functionals in the theoretically interesting infinite-basis limit, but rather to assess the performance of DFT methods for the calculation of noncovalent interactions with a moderate basis set suitable for practical calculations on complex systems. Csonka et al.[71−73] reported that the diffuse functions are very important to obtain reasonable DFT results, but they also showed that the 6-31+G(d) basis set gives a systematic overbinding for weak

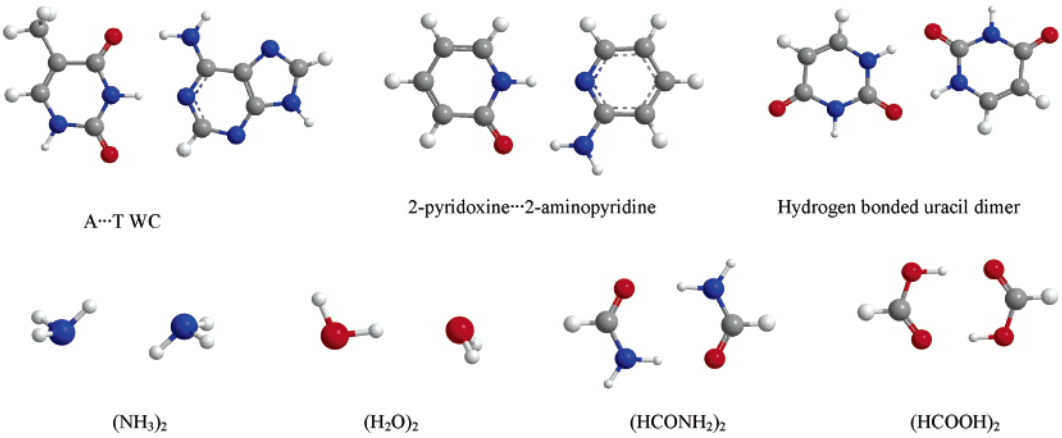


Figure 1. Structures of hydrogen-bonded complexes.

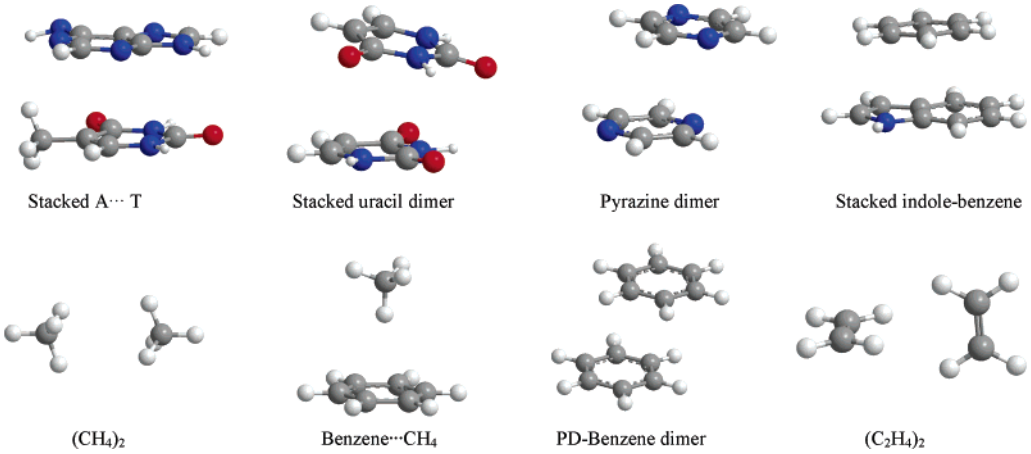


Figure 2. Structures of dispersion-dominated complexes.

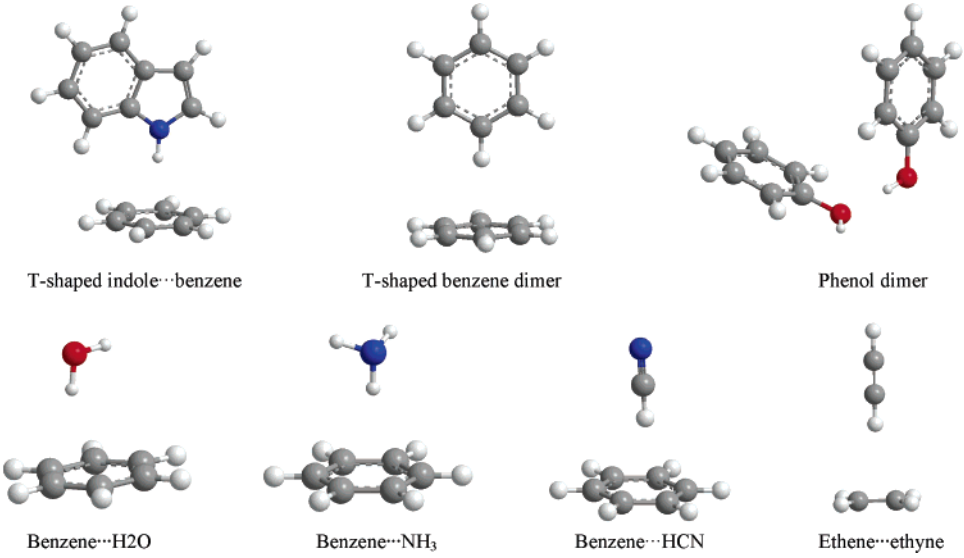


Figure 3. Structures of mixed complexes.

interactions, so we performed calculations both with and without counterpoise (CP) corrections[74,75] for basis set superposition error (BSSE). To investigate the effect of basis set size and confirm that the methods that give good results with the 6-31+G(d,p) basis set do not do so because of some accidental cancellation of functional-related errors and basis-set-related errors, we also performed calculations with a triple-ζ basis set, namely, 6-311+G(2df,2p).

For further analyses, and following the work by Shibasaki et al.,[76] we calculated the intermolecular potential of the $C_6H_6-CH_4$ complex with the 6-311+G(2df,2p) basis set.

3. Results and Discussion

The statistical errors for 15 methods (the four best-performing hybrid meta functionals, three best-performing hybrid GGAs, three best-performing meta-GGAs, three best-performing

Table 1. Tested DFT Methods

			exchange				correlation			
method	year	ref(s)	$\rho_{,\sigma}\ \nabla\rho_\sigma$	X	τ?	UEG?	ρ_σ, $\nabla\rho_\sigma$	τ?	SCorF?	UEG?
BP86	1988	33, 34	B88	0	no	yes	P86	no	no	yes
BLYP	1988	34, 35	B88	0	no	yes	LYP	no	yes	no
PW91	1991	36	PW91	0	no	yes	PW91	no	no	yes
LSDA[a]	1992	32, 103	Slater	0	no	yes	PW91-L	no	no	yes
BHandH[b]	1993	34, 35	Slater	50	no	yes	LYP	no	yes	no
BHandHLYP[b]	1993	34, 35	B88	50	no	yes	LYP	no	yes	no
B3LYP	1994	34, 35, 43	B88	20	no	yes	LYP	no	yes	no
BB95	1996	34, 50	B88	0	no	yes	B95	yes	yes	yes
B1B95	1996	34, 50	B88	28	no	yes	B95	yes	yes	yes
G96LYP	1996	35, 37	G96	0	no	yes	LYP	no	yes	no
PBE	1996	38	PBE	0	no	yes	PBE	no	no	yes
mPWPW[c]	1998	36, 39	mPW	0	no	yes	PW91	no	no	yes
MPWLYP	1998	35, 39	mPW	0	no	yes	LYP	no	yes	no
mPWB95	1998	50, 39	mPW	0	no	yes	B95	yes	yes	yes
mPW1PW[d]	1998	36, 39	mPW	25	no	yes	PW91	no	no	yes
B98	1998	44	B98	21.98	no	no	B98	no	no	no
VSXC	1998	104	VSXC	0	yes	no	VSXC	yes	yes	no
HCTH	1998	40	HCTH	0	no	no	HCTH	no	no	no
B97−1	1998	40	B97-1	21	no	no	B97-1	no	no	no
PBEh[e]	1999	45	PBE	25	no	yes	PBE	no	no	yes
MPW1K	2000	46	mPW	42.8	no	yes	PW91	no	no	yes
B97-2	2001	40	B97-2	21	no	no	B97-2	no	no	no
OLYP	2001	41	OPTX	0	no	no	LYP	no	yes	no
O3LYP	2001	41	OPTX	11.61	no	no	LYP	no	yes	no
τ-HCTH	2002	51	τ-HCTH	0	yes	no	τ-HCTH	no	no	no
τ-HCTHh	2002	51	τ-HCTHh	15	yes	no	τ-HCTHh	no	no	no
TPSS	2003	52	TPSS	0	yes	yes	TPSS	yes	yes	yes
TPSSh	2003	53	TPSS	10	yes	yes	TPSS	yes	yes	yes
X3LYP	2004	35, 47	X	21.8	no	yes	LYP	no	yes	no
BB1K	2004	34, 50, 54	B88	42	no	yes	B95	yes	yes	yes
BMK	2004	55	BMK	42	yes	no	BMK	no	no	no
MPW3LYP	2004	50, 48	mPW	20	no	yes	B95	yes	yes	yes
MPW1B95	2004	50, 39, 48	mPW	31	no	yes	B95	yes	yes	yes
MPWB1K	2004	50, 39, 48	mPW	44	no	yes	B95	yes	yes	yes
PW6B95	2005	17	PW6B95	28	no	yes	PW6B95	yes	yes	yes
PWB6K	2005	17	PWB6K	46	no	yes	PWB6K	yes	yes	yes
PBE1W	2005	42	PBE	0	no	yes	Scaled PBE	no	no	yes
B97-3	2005	49	B97-3	26.93	no	no	B97-3	no	no	no
M05	2005	56	M05	28	yes	yes	M05	yes	yes	yes
M05-2X	2006	21	M05-2X	56	yes	yes	M05-2X	yes	yes	yes

[a] Instead of using VWN, we use PW91-local for the LSDA correlation. [b] Although inspired by Becke's paper,[105] the BHandH and BHandHLYP functionals defined in Gaussian03 are different from the original ones in Becke's paper. (See http://www.gaussian.com/g_ur/k_dft.htm.) [c] Also called mPWPW91. [d] Also called mPW1PW91, mPW0 or MPW25. [e] Also called PBE0 or PBE1PBE.

GGAs, one LSDA, and MP2) are tabulated in Tables 2−4, and the calculated interaction energies for 40 functionals and MP2 are given in Tables S2−S4 of the Supporting Information. Statistical errors comparable to those in Tables 2−4 for all 41 methods are given in Tables S5−S7 of the Supporting Information. In Tables 2−4, we tabulate mean unsigned error (MUE, same as mean absolute deviation), mean signed error (MSE), and MMUE, which is defined as

$$\text{MMUE} = 0.5\text{MUE} + 0.5\text{MUE-CP} \qquad (2)$$

where we use the convention that CP denotes a mean error computed from calculations that include CP corrections for BSSE, and an error specified without "CP", as in the first term on the right side of eq 2, is computed without such corrections.

In general, the use of CP corrections is problematic for several reasons. First, although they usually improve the accuracy for very small basis sets [smaller than 6-31+G-(d,p)], they sometimes make the results less accurate for moderate and large basis sets [such as 6-31+G(d,p) and larger]. Second, for complex systems like biopolymers and soft materials, even for trimers, the CP correction is often ambiguous[77] or impractical[78] or both. Because there are strong proponents of including CP corrections and other workers strongly persuaded by practical reasons for not including it, we base most of our discussion on MMUE, which is a middle ground between the two positions, but the results are in the tables both ways for readers who strongly prefer to include or exclude CP corrections. The

Table 2. Mean Errors (kcal/mol) for the Interaction Energies of Hydrogen-Bonded Complexes[a]

method	MSE-CP	MUE-CP	MSE	MUE	MMUE[b]
PWB6K	0.40	0.83	−0.30	0.68	0.76
PBEh	0.65	0.92	−0.06	0.67	0.79
M05-2X	0.49	0.90	−0.15	0.69	0.80
PW91	0.50	0.94	−0.23	0.68	0.81
MPWB1K	1.08	1.26	0.39	0.87	1.06
PBE	1.05	1.31	0.35	0.94	1.13
B97-1	1.10	1.36	0.43	0.98	1.17
M05	1.08	1.45	0.41	1.06	1.26
TPSS	1.80	1.83	1.08	1.35	1.59
MP2	2.46	2.46	−0.08	0.81	1.64
PBE1W	1.81	1.92	1.13	1.56	1.74
τ-HCTH	2.02	2.02	1.24	1.49	1.76
B3LYP	1.94	1.97	1.31	1.56	1.77
BHandH	−4.89	4.89	−5.54	5.54	5.21
LSDA	−5.21	5.21	−5.88	5.88	5.55

[a] The 6-31+G(d,p) basis is used for all calculations on which this table is based. The signed error is defined as the calculated energy minus the best estimate. [b] MSE denotes mean signed error, and MUE denotes mean unsigned error. MMUE is the average of MUE and MUE-CP (also defined in eq 2).

Table 3. Mean Errors (kcal/mol) for the Interaction Energies of the Dispersion-Dominated Complexes[a]

method	MSE-CP	MUE-CP	MSE	MUE	MMUE[b]
LSDA	−0.08	0.36	−0.75	0.75	0.56
BHandH	0.02	0.60	−0.61	0.84	0.72
M05-2X	1.33	1.33	0.69	0.70	1.01
PWB6K	2.02	2.02	1.37	1.37	1.69
MP2	1.30	1.30	−2.60	2.76	2.03
MPWB1K	2.79	2.79	2.14	2.14	2.46
M05	3.47	3.47	2.85	2.85	3.16
PW91	4.42	4.42	3.73	3.73	4.07
B97-1	4.47	4.47	3.83	3.83	4.15
PBEH	4.59	4.59	3.93	3.93	4.26
PBE	4.86	4.86	4.20	4.20	4.53
PBE1W	5.54	5.54	4.89	4.89	5.21
TPSS	5.95	5.95	5.29	5.29	5.62
B3LYP	6.53	6.53	5.91	5.91	6.22
τ-HCTH	7.09	7.09	6.41	6.41	6.75

[a] The 6-31+G(d,p) basis is used for all calculations on which this table is based. The signed error is defined as the calculated energy minus the best estimate. [b] MSE denotes mean signed error, and MUE denotes mean unsigned error. MMUE is the average of MUE and MUE-CP (also defined in eq 2).

reader will see that our major conclusions are independent of whether or not we include CP corrections.

In Table 5, AMUE is the average of the MUEs in Tables 2 (or S5), 3 (or S6), and 4 (or S7), each weighted one-third, and AMUE-CP is the average of the MUE-CPs in Tables 2−4, again each weighted one-third. MAMUE is the overall mean error defined by

$$\text{MAMUE} = 0.5\text{AMUE} + 0.5\text{AMUE-CP} \qquad (3)$$

In Tables 2−5, the density functionals and MP2 are always arranged in increasing order of the mean error in the last column.

3.1. Hydrogen-bonded Complexes. Table 2 gives the results for the hydrogen-bonded complexes. All tested DFT functionals except LSDA and BHandH underestimate the interaction energies in these seven hydrogen-bonded complexes, whereas LSDA and BHandH severely overestimate the interaction energies. Table 2 also shows the failure of the OLYP and O3LYP functionals for describing hydrogen bonding.

From the statistical errors in Table S5 of the Supporting Information, we can also see the importance of HF exchange in DFT for describing hydrogen bonding. The percentages of HF exchange (see Table 1) in mPWPW91, mPW1PW91, and MPW1K are 0, 25, and 42.8, respectively, and the order of the MMUEs is mPWPW91 > mPW1PW91 > MPW1K. The same trends can be seen for the (mPWB95, MPW1B95, MPWB1K) sequence, the (PBE, PBEh) sequence, the (BLYP, B3LYP, BHandHLYP) sequence, and the (BB95, B1B95, BB1K) sequence. We discuss this point further in section 3.4.

The four best performers, on the basis of their small MMUEs, are PWB6K, PBEh, M05-2X, and PW91.

Table 4. Mean Errors (kcal/mol) for the Interaction Energies of the Mixed Complexes[a]

method	MSE-CP	MUE-CP	MSE	MUE	MMUE[b]
M05-2X	0.34	0.47	−0.11	0.40	0.43
PWB6K	0.66	0.66	0.24	0.44	0.55
MPWB1K	1.22	1.22	0.80	0.80	1.01
M05	1.30	1.30	0.85	0.88	1.09
MP2	1.00	1.00	−1.45	1.45	1.23
LSDA	−1.06	1.06	−1.60	1.60	1.33
BHandH	−1.10	1.10	−1.57	1.57	1.34
PW91	1.60	1.60	1.07	1.11	1.35
B97-1	1.65	1.65	1.17	1.18	1.42
PBEh	1.70	1.70	1.22	1.22	1.46
PBE	1.91	1.91	1.40	1.40	1.66
PBE1W	2.30	2.30	1.79	1.79	2.04
TPSS	2.61	2.61	2.11	2.11	2.36
B3LYP	2.89	2.89	2.40	2.40	2.64
τ-HCTH	2.92	2.92	2.42	2.42	2.67

[a] The 6-31+G(d,p) basis is used for all calculations on which this table is based. The signed error is defined as the calculated energy minus the best estimate. [b] MSE denotes mean signed error, and MUE denotes mean unsigned error. MMUE is the average of MUE and MUE-CP (also defined in eq 2).

3.2. Dispersion-Dominated Complexes. Table 3 shows that LSDA, BHand H, M05-2X, PWB6K, and MP2 are the best performers for the prediction of interaction energies of the dispersion-dominated (or dispersion-like-dominated) complexes. Our previous papers[17,21] show that LSDA gives good predictions for the energetics of the stacked benzene dimers, but LSDA gives large errors for hydrogen bonding, charge-transfer complexes, dipole interactions, and other types of dispersion-like interactions. Kurita et al.[79] showed previously that a post-LSDA method can give reasonable results for stacking. Waller et al.[80] also found that the BHandH functional, which is a hybrid of LSDA exchange and HF exchange (50:50) plus LYP correlation, gives a binding energy for the parallel-displaced benzene dimer in fortuitously good agreement with the best available high-level methods. Nevertheless, BHandH suffers the same problems as LSDA (as shown in Tables 2 and 4).

Among the best five performers in Table 3, M05-2X and PWB6K are also among the best five performers in Table

Table 5. Overall Performance (kcal/mol)[a]

	AMUE-CP	AMUE	MAMUE
M05-2X	**0.90**	**0.60**	**0.75**
PWB6K	**1.17**	**0.83**	**1.00**
MPWB1K	**1.75**	**1.27**	**1.51**
MP2	**1.59**	1.67	**1.63**
PW6B95	**1.96**	**1.46**	**1.71**
M05	2.07	1.60	1.83
MPW1B95	2.08	**1.59**	1.84
PW91	2.32	1.84	2.08
PBEh	2.40	1.94	2.17
B97-1	2.49	2.00	2.24
BB1K	2.63	2.09	2.36
BHandH	2.20	2.65	2.42
PBE	2.69	2.18	2.44
MPW1K	2.71	2.22	2.46
LSDA	2.21	2.75	2.48
mPWB95	2.83	2.30	2.57
MPW3LYP	2.85	2.36	2.60
τ-HCTHh	2.86	2.36	2.61
B98	2.87	2.37	2.62
BMK	2.85	2.39	2.62
BHandHLYP	2.88	2.45	2.66
mPW1PW	3.15	2.61	2.88
B1B95	3.20	2.65	2.93
X3LYP	3.19	2.70	2.95
PBE1W	3.25	2.74	3.00
TPSSh	3.29	2.74	3.01
TPSS	3.46	2.91	3.19
MPWLYP	3.54	3.01	3.28
mPWPW	3.73	3.17	3.45
B97-3	3.77	3.26	3.52
B3LYP	3.80	3.29	3.54
B97-2	3.98	3.41	3.69
τ-HCTH	4.01	3.44	3.73
BP86	4.06	3.51	3.79
BB95	4.27	3.69	3.98
HCTH	4.41	3.82	4.11
BLYP	4.90	4.31	4.60
O3LYP	6.35	5.66	6.00
G96LYP	7.35	6.75	7.05
OLYP	7.48	6.75	7.12
VSXC	7.05	7.33	7.19

[a] AMUE-CP is the average of the MUE-CPs in Tables 2−4 for the 14 featured functionals and MP2 and of those in Tables S5−S7 of the Supporting Information for the other 26 functionals. AMUE is the average of the MUEs (without CP) in the same tables. MAMUE = 0.5 AMUE-CP + 0.5 AMUE. The 6-31+G(d,p) basis set is used for all calculations on which this table is based.

2. It is encouraging that these two fairly new functionals, M05-2X and PWB6K, give equally good performance for calculating interaction energies in both hydrogen-bonded and dispersion-dominated complexes. The second and fourth best performers in Table 2, PBEh and PW91, give much greater MMUEs in Table 3 than M05-2X and PWB6K do. Tables S2 and S6 of the Supporting Information also show that VSXC severely overestimates the dispersion-dominated noncovalent interactions in these complexes.

3.3. Mixed Complexes. Because M05-2X and PWB6K give balanced high accuracy for hydrogen-bonding (Table 2) and dispersion-dominated interactions (Table 3), it is not surprising that they are the best two methods for the predictions of interaction energies in the mixed complexes (Table 4), followed by MPWB1K. All of these leading functionals were published in 2004 or later (see Table 1).

3.4. Overall Performance. Table 5 shows the overall performance for all tested DFT methods and for MP2. With CP correction for BSSE, M05-2X, PWB6K, MP2, and MPWB1K give the best performance. Without CP, M05-2X, PWB6K, MPWB1K, and PW6B95 are the best four performers. Averaging the performance for CP and without CP, the best four methods for describing noncovalent interactions in the S22 database are M05-2X, PWB6K, MPWB1K, and MP2.

Tables S5−S7 of the Supporting Information and Table 5 show that, all other factors being the same, a higher percentage of HF exchange in DFT improves the performance for describing noncovalent interactions in biological systems. However, for some functionals, like BHandHLYP, the predictions for other quantities like covalent bond energies deteriorate badly when X is raised.[46,81] For others, like MPW1K, the predictions for covalent bond energies deteriorate only slightly.[46,81] For M05-2X, the predictions of main-group covalent bond energies improve in quality along with the quality of the predictions for noncovalent interactions.[21] Thus, a key development in recent functional design is that we now have functionals that are more broadly accurate over a range of properties: thermochemistry,[21] barrier heights,[21] torsional potential and proton affinities of conjugated systems,[82] and noncovalent interactions.[20,21,29]

3.5. Dependence on Basis Set. All results in Tables 2−5 are based on the 6-31+G(d,p) basis set. It is also interesting to examine the performance of the DFT functionals for larger basis sets. Table 6 gives the interaction energies and mean errors by the two best hybrid meta-GGAs (M05-2X and PWB6K) and the two best hybrid GGAs (B97-1 and PBEh) with the 6-311+G(2df,2p) basis set. Table 6 also gives the MP2/CBS results from Jurecka et al.[31]

Table 6 shows that MP2/CBS is more accurate than DFT for the hydrogen-bonded complexes; it gives a MUE of 0.15 kcal/mol, and four DFT functionals with the 6-311+G(2df,2p) basis set give errors in the range of 0.7∼1.1 kcal. Comparing Tables 2 and 6, we can see that increasing the basis set size from 6-31+G(d,p) to 6-311+G(2df,2p) reduces the MMUEs for M05-2X, PWB6K, and B97-1 and increases the MMUE for PBEh; in all four cases, the MMUEs change by only 14% or less. Comparing the interaction energies of the hydrogen-bonded complexes in Table S2 of the Supporting Information and Table 6, we can see that the BSSEs are reduced significantly for the larger basis set.

Table 6 also shows that MP2/CBS overestimates the interaction energies by a large margin for the dispersion-dominated complexes, especially for those involving delocalized π systems. In fact, the mean signed error of MP2/CBS is 1.5 kcal/mol. In contrast, all four density functionals underbind these complexes. Encouragingly, M05-2X gives smaller MUEs than MP2/CBS for this type of interaction energy. Comparing Tables 3 and 6, we can see that, when the basis set is increased from 6-31+G(d,p) to 6-311+G(2df,2p), the MMUEs of the M05-2X, PBEh, and B97-1 methods get smaller, whereas the MMUE of PWB6K gets

Table 6. Interaction Energies (kcal/mol) with the 6-311+G(2df,2p) Basis Set[a]

			M05-2X		PWB6K		PBEh		B97-1	
complex	best estimate	MP2/CBS[b]	CP	noCP	CP	noCP	CP	noCP	CP	noCP
Hydrogen-Bonded Complexes										
$(NH_3)_2$	−3.17	−3.20	−3.23	−3.34	−3.20	−3.32	−2.87	−3.00	−2.95	−3.06
$(H_2O)_2$	−5.02	−5.03	−5.09	−5.50	−5.14	−5.58	−4.93	−5.40	−4.89	−5.32
formic acid dimer	−18.61	−18.60	−18.87	−19.49	−18.68	−19.32	−18.49	−19.17	−17.59	−18.23
formamide dimer	−15.96	−15.86	−15.58	−15.96	−15.44	−15.84	−15.14	−15.55	−14.64	−15.02
uracil dimer	−20.65	−20.61	−19.36	−19.89	−19.44	−19.97	−19.02	−19.58	−18.45	−18.95
2-pyridoxine•2-aminopyridine	−16.71	−17.37	−15.22	−15.71	−15.15	−15.63	−15.31	−15.80	−14.77	−15.21
adenine•thymine WC	−16.37	−16.54	−14.69	−15.22	−14.52	−15.05	−14.45	−15.00	−14.01	−14.51
MSE[c]		−0.10	0.64	0.20	0.70	0.26	0.90	0.43	1.31	0.88
MUE[c]		0.15	0.75	0.63	0.77	0.66	0.90	0.70	1.31	0.97
MMUE[c]			0.69	0.69	0.71	0.71	0.80	0.80	1.14	1.14
Dispersion-Dominated Complexes										
$(CH_4)_2$	−0.53	−0.51	−0.50	−0.53	−0.52	−0.54	−0.04	−0.05	−0.23	−0.24
$(C_2H_4)_2$	−1.51	−1.62	−1.40	−1.49	−1.40	−1.47	−0.35	−0.41	−0.62	−0.68
benzene•CH_4	−1.5	−1.86	−1.15	−1.35	−1.00	−1.18	−0.12	−0.31	−0.30	−0.46
benzene dimer	−2.73	−4.95	−1.41	−2.00	−0.42	−1.13	1.70	1.13	1.46	0.99
pyrazine dimer	−4.42	−6.90	−2.99	−3.67	−1.96	−2.54	0.45	−0.17	0.30	−0.25
uracil dimer	−10.12	−11.39	−8.47	−9.52	−6.87	−7.84	−3.38	−4.41	−3.56	−4.46
indole•benzene	−5.22	−8.12	−2.65	−3.55	−1.47	−2.25	1.89	1.05	1.76	1.04
adenine•thymine stack	−12.23	−14.93	−9.58	−10.88	−8.21	−9.35	−2.27	−3.50	−2.19	−3.27
MSE[c]		−1.50	1.26	0.66	2.05	1.49	4.52	3.95	4.36	3.87
MUE[c]		1.51	1.26	0.66	2.05	1.50	4.52	3.95	4.36	3.87
MMUE[c]			0.96	0.96	1.77	1.77	4.23	4.23	4.11	4.11
Mixed Complexes										
ethene•ethyne	−1.53	−1.69	−1.42	−1.49	−1.37	−1.44	−1.16	−1.24	−1.30	−1.37
benzene•H_2O	−3.28	−3.61	−3.50	−3.89	−3.07	−3.48	−2.20	−2.66	−2.29	−2.70
benzene•NH_3	−2.35	−2.72	−2.23	−2.50	−1.95	−2.22	−1.06	−1.35	−1.21	−1.46
benzene• HCN	−4.46	−5.16	−4.80	−5.15	−4.51	−4.83	−3.26	−3.61	−3.18	−3.49
benzene dimer	−2.74	−3.62	−1.95	−2.35	−1.55	−1.91	−0.30	−0.68	−0.40	−0.73
indole•benzene T-shape	−5.73	−7.03	−4.68	−5.20	−3.89	−4.37	−2.32	−2.84	−2.29	−2.75
phenol dimer	−7.05	−7.76	−6.00	−6.62	−5.62	−6.22	−4.17	−4.80	−4.18	−4.76
MSE[c]		−0.64	0.37	−0.01	0.74	0.38	1.81	1.42	1.75	1.41
MUE[c]		0.64	0.53	0.40	0.76	0.54	1.81	1.42	1.75	1.41
MMUE[c]			0.47	0.47	0.65	0.65	1.62	1.62	1.58	1.58
AMUE[d]		0.76	0.84	0.57	1.19	0.90	2.41	2.02	2.48	2.07
MAMUE[d]			0.71	0.71	1.05	1.05	2.22	2.22	2.27	2.27

[a] CP denotes "counterpoise correction", and no-CP denotes "without counterpoise correction". [b] The MP2/CBS results are from Jurecka et al.[31] [c] MSE denotes mean signed error, and MUE denotes mean unsigned error. MMUE is the average of MUE and MUE-CP (also defined in eq 2). [d] AMUE is the average of the three MUEs in each class of complexes. MAMUE = 0.5 AMUE-CP + 0.5 AMUE.

larger, but the MMUEs for all four functionals change by only 5% or less when the basis set size is increased.

For the interaction energies in the seven mixed complexes, MP2/CBS gives an error of 0.64 kcal/mol, whereas M05-2X outperforms MP2 by a small margin. Comparing Tables 4 and 6, we can see that, upon increasing the basis set size from 6-31+G(d,p) to 6-311+G(2df,2p), the performance of all four density functionals for the prediction of this type of interaction energies deteriorates, but again, the change is small, in this case, 18% or less and only 8−11% in three of the cases. Again, the MMUEs change only marginally with the increase of basis set size.

Finally, let us compare AMUEs and MAMUEs in Table 5 to those in Table 6 for these four functionals. The comparison shows that AMUE and MAMUE for M05-2X decrease when the basis set size is increased, whereas the AMUEs and MAMUEs for PWB6K, PBEh, and B97-1 increase. However, the changes in these mean errors are very small. This observation gives us confidence of the validity of the conclusions drawn on the basis of the results in Table 2−6 and Tables S2−S7 of the Supporting Information. One of the reviewers pointed out that MP2 requires large basis sets, and it is a strength of DFT that it does not. It is encouraging that the bottom line accuracy of M05-2X in Table 6 (0.71 kcal/mol) and the comparable number from Table 5 (0.75 kcal/mol) are both better than that of MP2, even when MP2 is taken to the CBS limit.

3.6. Geometry Optimization. In previous sections, we based our discussions on single-point energies calculated with the best estimated geometries for these noncovalent complexes. However, for many applications, it is important that a method can predict good geometries for these noncovalent complexes. To test the quality of the geometry predictions, we performed geometry optimizations for all 22 complexes with the M05-2X method. Encouragingly, M05-2X can locate the optimal structures for all noncovalent complexes in the

Table 7. M05-2X Interaction Energies (kcal/mol) with the M05-2X/6-31+G(d,p) Geometries

		6-31+G(d,p)		6-311+G(2df,2p)	
complex	best estimate	CP	noCP	CP	noCP
		Hydrogen-Bonded Complexes			
$(NH_3)_2$	−3.17	−3.90	−4.14	−3.23	−3.34
$(H_2O)_2$	−5.02	−5.82	−6.61	−5.16	−5.56
formic acid dimer	−18.61	−19.39	−20.22	−19.86	−20.50
formamide dimer	−15.96	−15.64	−16.14	−15.66	−16.03
uracil dimer	−20.65	−19.54	−20.27	−19.43	−19.95
2-pyridoxine·2-aminopyridine	−16.71	−15.92	−16.57	−15.72	−16.21
adenine·thymine WC	−16.37	−15.15	−15.90	−14.94	−15.47
MSE		0.16	−0.48	0.36	−0.08
MUE		0.82	0.76	0.77	0.68
MMUE-HB[a]		0.79	0.79	0.73	0.73
		Dispersion-Dominated Complexes			
$(CH_4)_2$	−0.53	−0.55	−0.56	−0.50	−0.53
$(C_2H_4)_2$	−1.51	−1.33	−1.46	−1.41	−1.49
benzene·CH_4	−1.5	−1.10	−1.24	−1.17	−1.34
benzene dimer	−2.73	−1.52	−2.02	−1.70	−2.20
pyrazine dimer	−4.42	−3.04	−3.66	−3.10	−3.73
uracil dimer	−10.12	−9.07	−10.42	−9.03	−10.19
indole·benzene	−5.22	−3.05	−3.76	−3.15	−3.89
adenine·thymine stack	−12.23	−9.67	−11.20	−9.57	−10.86
MSE		1.12	0.49	1.08	0.50
MUE		1.12	0.57	1.08	0.52
MMUE-D		0.85	0.85	0.80	0.80
		Mixed Complexes			
ethene·ethyne	−1.53	−1.48	−1.67	−1.40	−1.48
benzene·H_2O	−3.28	−3.59	−4.02	−3.52	−3.94
benzene·NH_3	−2.35	−2.18	−2.50	−2.19	−2.45
benzene· HCN	−4.46	−4.61	−4.96	−4.78	−5.14
benzene dimer	−2.74	−2.01	−2.38	−2.02	−2.39
indole·benzene T-shape	−5.73	−4.64	−5.16	−4.72	−5.22
phenol dimer	−7.05	−6.70	−7.57	−6.30	−6.88
MSE		0.28	−0.16	0.32	−0.05
MUE		0.40	0.42	0.48	0.36
MMUE-Mix		0.41	0.41	0.42	0.42
AMUE[b]		0.78	0.59	0.78	0.52
MAMUE[b]		0.69	0.69	0.65	0.65

[a] CP denotes "counterpoise correction", and no-CP denotes "without counterpoise correction". [b] MSE denotes mean signed error, and MUE denotes mean unsigned error. MMUE is the average of MUE and MUE-CP (also defined in eq 2). [c] AMUE is the average of the three MUEs in each class of complexes. MAMUE = 0.5 AMUE(CP) + 0.5 AMUE(noCP).

S22 databases. We note that this is not a trivial triumph for the M05-2X method, because Cerny and Hobza have shown that the X3LYP functional, which was designed to describe noncovalent interactions, fails badly for locating the minima of the dispersion-dominated stacked structures of nucleic acid pairs such as the stacked AT pairs in the present study. The geometries optimized by M05-2X are given in the Supporting Information, and they agree well with the best estimated geometries.

Table 7 shows the M05-2X interaction energies with the geometries optimized at the M05-2X/6-31+G(d,p) level. Comparing the results in Tables 2−7, we can see that the overall performance of M05-2X is improved when we use the M05-2X/6-31+G(d,p) geometries, which is another encouraging result.

3.7. Rationale for the Success of M05-2X for Noncovalent Interactions. There are some common misunderstandings about the performance of DFT for noncovalent interactions. In the literature, one sometimes sees the success of DFT for noncovalent interactions labeled as "fortuitous" or "spurious". It is true that most DFT functionals cannot describe the $-C_6/R^6$ interaction of nonoverlapped densities, where R is the interaction distance of the monomers. Nevertheless, at the equilibrium distance of noncovalent complexes, the lack of explicit R^{-6} terms need not be a serious issue because the higher terms (R^{-8} etc.) in the asymptotic expansion are not negligible,[60,83−88] the dispersion interaction is damped,[86,89,90] the expansion in the inverse power of R is divergent,[84,91] overlap and exchange forces are not negligible,[47,85,86,90,92−94] and the change in intra-atomic correlation energy cannot be neglected.[95,96] In fact, the decomposition of the correlation contribution to the interaction energy into intra-atomic and interatomic (dispersion-like) parts is not unique.[94,95] Thus, DFT is not excluded as a potentially useful theory for the medium-range part of noncovalent interactions, as is shown by our previous papers.[16,19,97]

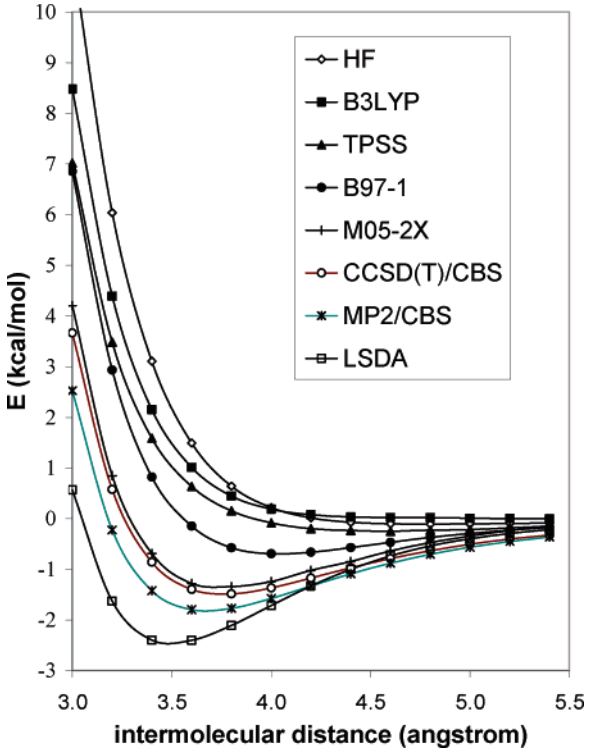


Figure 4. Binding energy curves for the $C_6H_6-CH_4$ complex with the 6-311+G(2df,2p) basis set. The intermolecular distance is defined as the distance between the carbon atom in CH_4 and the C_6H_6 plane. The CCSD/CBS and MP2/CBS results are taken from Shibasaki et al.[76]

Recently, Tao and Perdew[98] and Ruzsinszky et al.[73] have analyzed the performance of several DFT functionals for describing the medium-range part of the weak interactions.

In order to better understand the physical origin of the wells predicted by M05-2X and the range of distances over which M05-2X is suitable for treating noncovalent interactions, we compare the intermolecular potentials of the $C_6H_6-CH_4$ complexes calculated by five density functionals (B3LYP, TPSS, B97−1, M05-2X, and LSDA) and HF in Figure 4, along with the CCSD(T)/CBS and MP2/CBS results by Shibasaki et al.[76] Figure 4 shows that MP2 and LSDA overestimate the strength of the $C_6H_6-CH_4$ complex as compared to the CCSD(T)/CBS results. HF, B3LYP, and TPSS give an almost repulsive potential for this van der Waals complex; B97-1 gives a well with the minimum around 4.0 Å, and M05-2X gives the best agreement with the CCSD(T)/CBS results.

Exchange-only calculations (that is, calculating interaction energies without correlation contributions) employing HF and five density functionals are shown in Figure 5. As can be seen from Figure 5, the exchange-only potentials obtained using M05-2X, B3LYP, TPSS, and B97-1 exchange functionals are repulsive for the $C_6H_6-CH_4$ system, which agrees with the HF results. However, the exchange-only calculation based on the LSDA exchange gives a "spurious" well. (Although this comparison is interesting, one should be careful not overinterpret it because the distinction between exchange and correlation is different in DFT and WFT.)

The correlation contribution to the intermolecular potentials of the $C_6H_6-CH_4$ complexes are plotted in Figure 6.

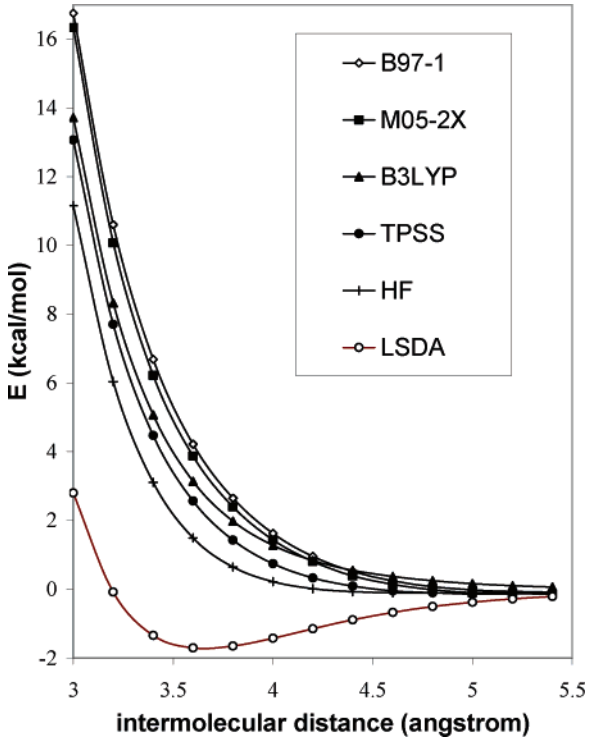


Figure 5. Exchange-only binding energy curves for the $C_6H_6-CH_4$ complex with the 6-311+G(2df,2p) basis set. The intermolecular distance is defined as the distance from the carbon atom in CH_4 to the C_6H_6 plane.

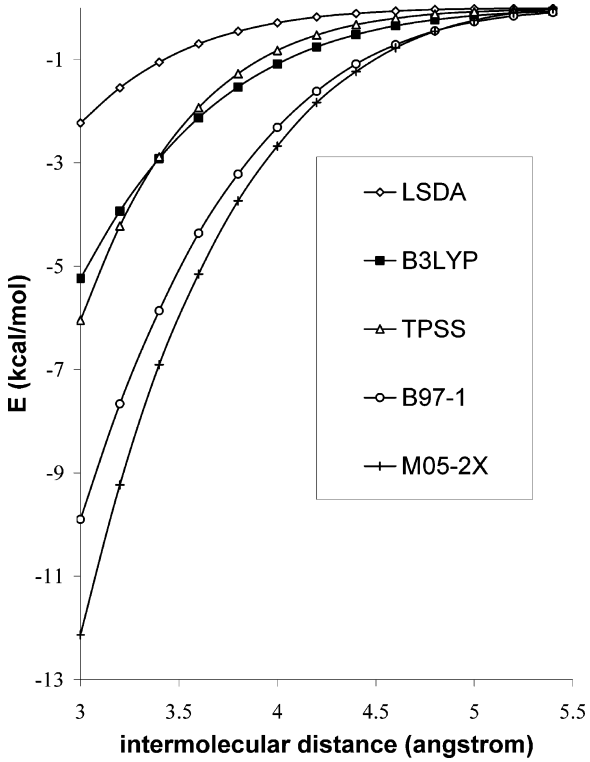


Figure 6. Correlation contribution to binding energy curves for the $C_6H_6-CH_4$ complex with the 6-311+G(2df,2p) basis set.

Figure 6 shows that the good performance of M05-2X comes from its correlation part; it gives the most attractive contribution in the range of 3−5 Å. The LSDA correlation contribu-

tion is the least attractive; this confirms that the good performance of LSDA and BHandH for $\pi-\pi$ stacking is a case of "getting the right answer with the wrong reason". It is interesting to see that TPSS correlation and the B3LYP correlation show very similar behaviors in the range of 3−5 Å for the $C_6H_6-CH_4$ system.

Comparing the results in Figures 4−6, we conclude that the good performance of M05-2X for describing the medium-range part of noncovalent interaction is because M05-2X has a better correlation functional, which gives the most attractive contribution to the potential energy. In fact, the success of M05-2X for medium-range correlation energy is also reflected in its performance for isomerization energies of hydrocarbons.[99]

3.8. Limitation of M05-2X for Noncovalent Interactions. Although M05-2X shows good performance for all noncovalent interactions in the S22 database, M05-2X does not give the asymptotic $-C_6/R^6$ tail for the interaction energy of two systems with no permanent multipole moments; it gives an exponential decay of the interaction energy for such systems at long range.

From the results in section 3.6, M05-2X can safely be applied to describe the interactions of noncovalent complexes with intermolecular distances less than ∼5 Å. For the study of dispersion-dominated noncovalent interactions at long-range (>6 Å), one should probably use wave function theory or functionals[58,60,61,100,101] that build in the correct asymptotic $-C_6/R^6$ dispersion tail.

4. Concluding Remarks

For many years, MP2 [and, for small enough systems, higher-order methods like CCSD(T)] was considered to be the standard method for estimating noncovalent interactions, when affordable. In a previous paper,[16] when the M05-2X and PWB6K functionals had not yet been developed, we concluded that MPWB1K was the best DFT method for noncovalent interactions and that it outperforms MP2 for noncovalent interactions. This general result was confirmed recently by a study by Slanina et al.;[102] they showed that MPWB1K gives very good stabilization energies for the encapsulation of H_2, Ne, and N_2 into C_{60}. Now, on the basis of the assessment in the present study, we see that the M05-2X and PWB6K functionals give even better performance than MPWB1K for noncovalent interactions, and fortunately these functionals, especially M05-2X, also have excellent performance for a broad range of other main-group chemistry.

The PWB6K and M05-2X functionals employ exchange and correlation functionals that include kinetic energy density and that were optimized together. The present study shows that these functionals constitute a new generation of DFT methods that have greatly improved performance for noncovalent interactions as compared to previous DFT methods. In particular, it shows that they give good performance for a benchmark database of noncovalent interactions of biological importance; it also confirms that the widely used B3LYP and PBEh functionals fail to describe the interactions in complexes dominated by dispersion-like interactions. The LSDA and BHandH functionals give good performance for dispersion-dominated interactions at the expense of large errors for covalent interaction, hydrogen bonding, and other types of noncovalent interactions;[16] we show here that the success of these methods is fortuitous. By studying the intermolecular potential of the $C_6H_6-CH_4$ complex, we found that the good performance of the M05-2X functional comes from its improved correlation functional, which gives a better description of the medium-range part of the noncovalent interactions.

We recommend the PWB6K and M05-2X functionals for investigating large biological systems and soft materials.

Acknowledgment. The authors are grateful to Seiji Tsuzuki for sending us the MP2/CBS and CCSD(T)/CBS potential data of the $C_6H_6-CH_4$ complex. This work was supported by the Office of Naval Research under grant number N00014-05-01-0538 (software tools) and by the National Science Foundation under grant number CHE03-49122 (functional development for complex systems).

Supporting Information Available: S22 database, calculated interaction energies and mean errors for 41 methods, $C_6H_6-CH_4$ potential energy data, and the M05-2X/6-31+G(d,p) optimized geometries. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Hobza, P.; Sponer, J. *Chem. Rev.* **1999**, *99*, 3247.

(2) Kannan, N.; Vishveshwara, S. *Protein Eng.* **2000**, *13*, 753.

(3) Müller-Dethlefs, K.; Hobza, P. *Chem. Rev.* **2000**, *100*, 143.

(4) Griffiths-Jones, S. R.; Searle, M. S. *J. Am. Chem. Soc.* **2000**, *122*, 8350.

(5) Kim, K. S.; Tarakeshwar, P.; Lee, J. Y. *Chem. Rev.* **2000**, *100*, 4145.

(6) Cha-lasinski, G.; Szczesniak, M.-l. M. *Chem. Rev.* **2000**, *100*, 4227.

(7) Sponer, J.; Leszczynski, J.; Hobza, P. *Biopolymers* **2001**, *61*, 3.

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

(9) Jurecka, P.; Hobza, P. *J. Am. Chem. Soc.* **2003**, *125*, 15608.

(10) Lamoureux, J. S.; Maynes, J. T.; Glover, J. N. M. *J. Mol. Biol.* **2004**, *335*, 399.

(11) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968.

(12) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, 864.

(13) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, 1133.

(14) Kohn, W.; Becke, A. D.; Parr, R. G. *J. Phys. Chem.* **1996**, *100*, 12974.

(15) Johnson, E. R.; DiLabio, G. A. *Chem. Phys. Lett.* **2006**, *419*, 333.

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

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

(18) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 6624.

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

(20) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 5121.

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

(22) Hobza, P.; Sponer, J.; Reschel, T. *J. Comput. Chem.* **1995**, *16*, 1315.

(23) Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2005**, *7*, 1624.

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

(25) van Mourik, T. *Chem. Phys.* **2004**, *304*, 317.

(26) van Mourik, T.; Gdanitz, R. J. *J. Chem. Phys.* **2002**, *116*, 9620.

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

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

(29) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 1009.

(30) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

(31) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.

(32) Perdew, J. P.; Wang, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *45*, 13244.

(33) Perdew, J. P. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *33*, 8822.

(34) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.

(35) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.

(36) Perdew, J. P. In *Electronic Structure of Solids '91*; Ziesche, P., Eschig, H., Eds.; Akademie Verlag: Berlin, **1991**; p 11.

(37) Gill, P. M. W. *Mol. Phys.* **1996**, *89*, 433.

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

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

(40) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264.

(41) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403.

(42) Dahlke, E. E.; Truhlar, D. G. *J. Phys. Chem. B* **2005**, *109*, 15677.

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

(44) Schmider, H. L.; Becke, A. D. *J. Chem. Phys.* **1998**, *108*, 9624.

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

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

(47) Xu, X.; Goddard, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673.

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

(49) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2005**, *123*, 121103.

(50) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040.

(51) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2002**, *116*, 9559.

(52) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(53) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129.

(54) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715.

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

(56) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103. Note that in this communication we interchanged $c_{C\alpha\beta,i}$ and $c_{C\sigma\sigma,i}$ in Table 1. In addition, "reduced density x_σ" before eq 1 should read "reduced density gradient x_σ".

(57) Ng, C. Y.; Lee, Y. T.; Barker, J. A. *J. Chem. Phys.* **1974**, *61*, 1996.

(58) Sato, T.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2005**, *123*, 104307.

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

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

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

(62) Hobza, P.; Sponer, J. *J. Am. Chem. Soc.* **2002**, *124*, 11802.

(63) Sponer, J.; Jurecka, P.; Hobza, P. *J. Am. Chem. Soc.* **2004**, *126*, 10142.

(64) Jurecka, P.; Sponer, J.; Hobza, P. *J. Phys. Chem. B* **2004**, *108*, 5466.

(65) Dabkowska, I.; Gonzalez, H. V.; Jurecka, P.; Hobza, P. *J. Phys. Chem. A* **2005**, *109*, 1131.

(66) Dabkowska, I.; Jurecka, P.; Hobza, P. *J. Chem. Phys.* **2005**, *122*, 204322.

(67) Vondrásek, J.; Bendová, L.; Klusák, V.; Hobza, P. *J. Am. Chem. Soc.* **2005**, *127*, 2615.

(68) Jurecka, P.; Hobza, P. *Chem. Phys. Lett.* **2002**, *365*, 89.

(69) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision C.01; Gaussian, Inc.: Pittsburgh, PA, 2003.

(70) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*, 1st ed.; Wiley: New York, 1986.

(71) Csonka, G. I. *THEOCHEM* **2002**, 584.

(72) Csonka, G. I.; Ruzsinszky, A.; Perdew, J. P. *J. Phys. Chem. B* **2005**, *109*, 21471.

(73) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I. *J. Phys. Chem. A* **2005**, *109*, 11015.

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

(75) Schwenke, D. W.; Truhlar, D. G. *J. Chem. Phys.* **1985**, *82*, 2418.

(76) Shibasaki, K.; Fujii, A.; Mikami, N.; Tsuzuki, S. *J. Phys. Chem. A* **2006**, *110*, 4397.

(77) Turi, L.; Dannenberg, J. J. *J. Phys. Chem.* **1993**, *97*, 2488.

(78) Van Mourik, T.; Karamertzanis, P. G.; Price, S. L. *J. Phys. Chem. A* **2006**, *110*, 8.

(79) Kurita, N.; Araki, M.; Nakao, K.; Kobayashi, K. *Int. J. Quantum Chem.* **2000**, *76*, 677.

(80) Waller, M. P.; Robertazzi, A.; Platts, J. A.; Hibbs, D. E.; Williams, P. A. *J. Comput. Chem.* **2006**, *27*, 491.

(81) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2001**, *105*, 2936.

(82) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 10478.

(83) Arrighini, G. P.; Biondi, F.; Gvidotti, C. *J. Chem. Phys.* **1971**, *55*, 4090.

(84) Kreek, H.; Meath, W. J. *J. Chem. Phys.* **1969**, *50*, 2289.

(85) Wormer, P. S.; Van der Avoird, A. *J. Chem. Phys.* **1975**, *62*, 3326.

(86) Tang, K. T.; Toennies, J. P. *J. Chem. Phys.* **1977**, *66*, 1496.

(87) Truhlar, D. G. *J. Chem. Phys.* **1993**, *98*, 2491.

(88) Meyer, W.; Hariharan, P. C.; Kutzelnigg, W. *J. Chem. Phys.* **1980**, *73*, 1880.

(89) Schwenke, D. W.; Truhlar, D. G. *Chem. Phys. Lett.* **1983**, *98*, 217; **1987**, *86*, 3760 (E).

(90) Tang, K. T.; Toennies, J. P. *J. Chem. Phys.* **1984**, *80*, 3726.

(91) Alhlriches, R. *Theor. Chim. Acta* **1976**, *41*, 7.

(92) Kochanski, E.; Gouyet, J. F. *Mol. Phys.* **1975**, *29*, 693.

(93) Cha-lasinski, G.; Jeziorski, B. *Theor. Chim. Acta* **1977**, *46*, 277.

(94) Kleinekathöfer, U.; Tang, K. T.; Toennies, J. P.; Yiu, C. L. *J. Chem. Phys.* **1997**, *107*, 9502.

(95) Spiegelmann, F.; Malriev, J.-P. *Mol. Phys.* **1980**, *40*, 1273.

(96) Staemmler, V.; Jaquet, R. *Chem. Phys. Lett.* **1985**, *92*, 141.

(97) Zhao, Y.; Tishchenko, O.; Truhlar, D. G. *J. Phys. Chem. B* **2005**, *109*, 19046.

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

(99) Zhao, Y.; Truhlar, D. G. *Org. Lett.* **2006**, in press, DOI: 10.1021/ol062318n.

(100) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2006**, *124*, 14104.

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

(102) Slanina, Z.; Pulay, P.; Nagase, S. *J. Chem. Theory Comput.* **2006**, *2*, 782.

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

(104) Voorhis, T. V.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400.

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

CT6002719

JCTC Journal of Chemical Theory and Computation

Computational Analysis of the Mechanism and Thermodynamics of Inhibition of Phosphodiesterase 5A by Synthetic Ligands

Bojan Zagrovic* and Wilfred F. van Gunsteren

Laboratory of Physical Chemistry, ETH, Zürich, CH-8093, Switzerland

Received October 30, 2006

Abstract: Phosphodiesterases are a large class of enzymes mediating a number of physiological processes ranging from immune response to platelet aggregation to cardiac and smooth muscle relaxation. In particular, phosphodiesterase 5 (PDE5) plays an important role in mediating sexual arousal, and it is the central molecular target in treatments of erectile dysfunction. In this study, we look at the mechanism and thermodynamics of the binding of selective inhibitors sildenafil (Viagra) and vardenafil (Levitra) to PDE5 using molecular dynamics simulations. Our simulations of PDE5 with and without sildenafil suggest a binding mechanism in which two loops surrounding the binding pocket of the enzyme (H loop, residues 660−683, and M loop, 787−812) execute sizable conformational changes (∼1 nm), clamping the ligand in the pocket. Also, we note significant changes in the coordination pattern of the divalent ions in the active site of the enzyme, as well as marked changes in the collective motions of the enzyme when the ligand is bound. Using the thermodynamic integration approach we calculate the relative free energies of binding of sildenafil, vardenafil, and demethyl-vardenafil, providing a test of the quality of the force field and the ligand parametrization used. Finally, using the single-step perturbation (SSP) technique, we calculate the relative binding free energies of these three ligands as well. In particular, we focus on critical evaluation of the SSP technique and examine the effects of computational parallelization on the efficiency of the technique. As a technical improvement, we demonstrate that an ensemble of relatively short SSP trajectories (10 × 0.5 ns) markedly outperforms a single trajectory of the same total length (1 × 5 ns) when it comes to sampling efficiency, resulting in significant real-time savings.

Introduction

Phosphodiesterases are a large family of enzymes involved in hydrolyzing second messengers cyclic guanosine monophosphate (cGMP) and cyclic adenosine monophosphate (cAMP).[1,2] These two second messengers, produced by enzymes guanylyl and adenylyl cyclase, are key components in a number of different signal transduction cascades providing a link between the activity of extracellular receptors such as the G-protein coupled receptors and downstream intracellular effectors such as different kinases and phosphatases. By decreasing the cellular levels of cGMP and cAMP, phosphodiesterases regulate a multitude of physiological processes including smooth and cardiac muscle relaxation, platelet aggregation, apoptosis, and vision.[3]

There are in total 11 different classes of phosphodiesterases, each class involved in regulating a particular set of physiological functions.[4] Phosphodiesterase 5 (PDE5; Figure 1a), in particular, has been shown to play an important role in mediating sexual response.[5] More specifically, this enzyme is a negative regulator of the signal transduction cascade leading to the relaxation of corpus cavernosum tissue and subsequent penile erection. Inhibition of PDE5 is

* Corresponding author phone: 0041-44-632-5504; e-mail: zagrovic@igc.phys.chem.ethz.ch and igc-sec@igc.phys.chem.ethz.ch.

10.1021/ct600322d CCC: $37.00
Published on Web 12/19/2006

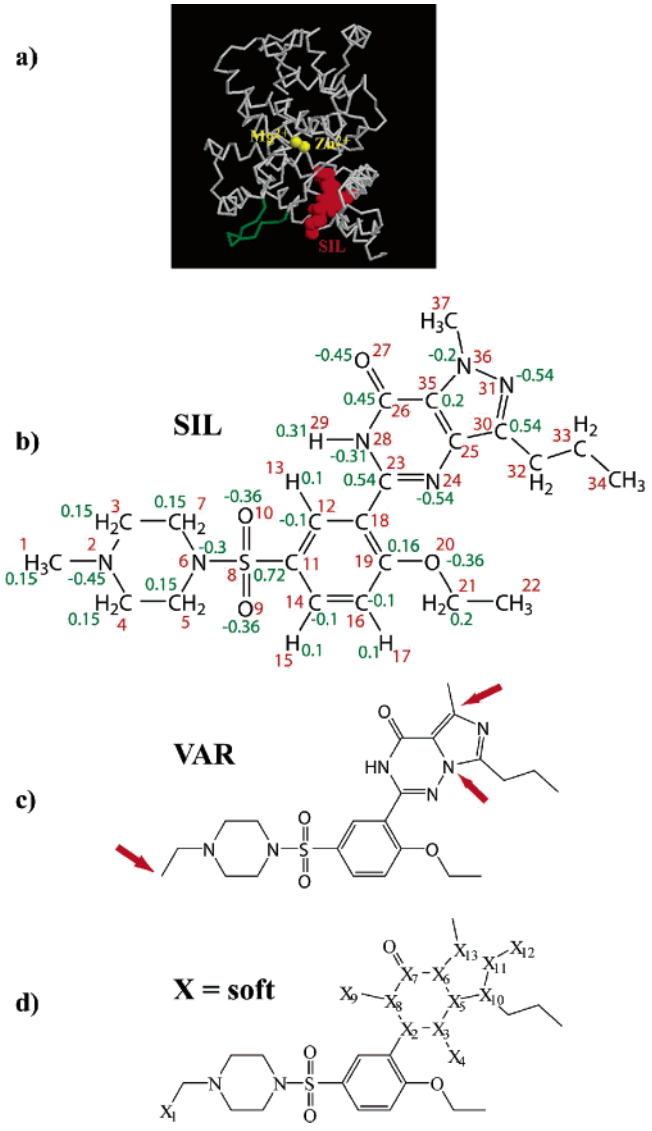


Figure 1. (a) X-ray structure of phosphodiesterase 5A (PDB code: 1UDT) with bound sildenafil (shown in red). The loop that was added (residues 665−675) is shown in green. The catalytically important divalent ions are shown in yellow. (b) Structure of sildenafil with atom numbers (red) and partial charges (green) from our parametrization. Atoms with no charges indicated have zero charge. (c) Structure of vardenafil. Points where it differs from sildenafil are labeled with arrows. (d) Structure of the soft reference state for the SSP technique. Soft atoms are depicted with X.

therefore a natural target in treatments of erectile dysfunction disorder, and several potent synthetic inhibitors of the enzyme have been developed.[6,7] These include the widely used sildenafil (Viagra), vardenafil (Levitra), and tardenafil (Cialis). While these ligands have proved to be very successful, significant efforts have been directed toward improving their pharmacological properties, in particular, decreasing cross-reactivity with other phosphodiesterase types. Namely, some of the side effects of using these drugs, including blurry vision and skin rash, can be attributed to nonselective inhibition of phosphodiesterases not involved in mediating sexual response such as phosphodiesterases 4B and 6.[5]

When it comes to the mechanism of binding of natural ligands as well as synthetic inhibitors to phosphodiesterases, significant progress has been made by detailed X-ray crystallographic analyses of different phosphodiesterase types.[6,8,9] In general, phosphodiesterase ligands bind in a deep, hydrophobic binding pocket, which penetrates all the way to a small water-filled cavity in the very core of the enzyme. This cavity houses two divalent ions, in most cases, magnesium and zinc, which take part in the catalytic stage of the hydrolysis reaction. Regarding the binding of ligands, two well-defined structural features of phosphodiesterases, common to all 11 classes of the enzyme, have been delineated.[6,8] First, a conserved glutamine residue forms a pair of strong bidentate hydrogen bonds with the purine moiety in cGMP, cAMP, and other structurally similar ligands, including sildenafil and vardenafil ("glutamine switch").[8] Second, a conserved valine on one side and an invariant aromatic residue on the other side sandwich the purine ring of the ligand in a fixed orientation with respect to the central cavity ("hydrophobic clamp").[6]

While X-ray analyses have taught us a lot about the mode of binding of different ligands to phosphodiesterases, they provide only an averaged, static picture of the process. Motivated by this, we have undertaken a computational analysis of PDE5A and its ligands sildenafil and vardenafil (Figure 1). By simulating in explicit solvent the enzyme in the presence and in the absence of sildenafil, we have obtained an atomistic picture of the dynamics and the structural mechanism of binding. Our simulations suggest a binding mechanism in which two particular loops surrounding the binding pocket of the enzyme (residues 660−683 and 787−812) undergo sizable conformational changes (∼1 nm), clamping the ligand in the pocket. Second, we observe significant changes in the coordination pattern of the divalent ions in the active site of the enzyme in the presence of a ligand. Finally, we note marked changes in the collective motions of the enzyme when the ligand is bound.

Using computational approaches, we have also analyzed the thermodynamics of ligand binding to PDE5A. Employing the thermodynamic integration technique, we have calculated the relative free energies of binding of sildenafil, vardenafil, and vardenafil with no extra methyl group attached to the piperazine moiety (demethyl-vardenafil).[10] Qualitative rank-order agreement between the calculated and experimental values testified about the adequacy of the force field used and motivated further calculations using the single-step perturbation technique (SSP).[11−13] In its traditional implementation, the SSP technique entails running two trajectories of a carefully chosen soft-core reference ligand, one in the binding pocket of a protein and another one free in solution. From these two trajectories, one should, in principle, be able to derive relative free energies of binding for a large set of real ligands. In addition to a good design of the reference state, the principal challenge for successful application of the SSP technique is adequate sampling of the phase space belonging to the reference state and, indirectly, to all the real ligands whose relative free energies of binding one is interested in.[13] As a consequence, motivation exists to try to improve the sampling of the low-energy states visited in

SSP simulations, while still retaining the method's greatest advantage, computational efficiency. Stimulated by the widespread presence of computer clusters of loosely coupled processors, we have here examined the potential of brute-force parallelization of the SSP approach and its effects on the degree of sampling that can be achieved. Here, by "brute-force parallelization" we mean simple simulation of multiple trajectories of the reference state on a series of fully independent processors. We show how simulating an ensemble of relatively short trajectories results in much better coverage of the configuration space of the protein−ligand complex compared with one single, long trajectory of the same total length. Also, the ensemble approach is shown to be significantly more successful in selecting low-energy states, which dominate thermodynamic averages when it comes to calculation of the relative binding free energies. Most importantly, these improvements are obtained with a linear speedup in the wall-clock time used, resulting in large real-time savings. For example, generating 10 0.5 ns trajectories of the soft state in the protein took only 5 days on a Pentium III cluster, while generating one 5 ns trajectory took approximately 50 days.

Materials and Methods

All simulations were performed using the GROMOS simulation package.[14] The initial structure of PDE5A (324 residues) with sildenafil bound[9] was taken from the Protein Data Bank (code: 1UDT) and was placed in a pre-equilibrated truncated octahedron box filled with SPC[15] water molecules (box size 9 nm, for a total of 10 619 water molecules). Parts of the H loop (residues 665−675; throughout this paper, the amino acid numbering corresponds to the original numbering in the 1UDT structure) missing in the 1UDT structure because of disorder were modeled-in using the Modloop routine in the software package Modeller.[16,17] In all simulations, an equilibration scheme was carried out, which included raising the simulation temperature from 60 to 300 K while simultaneously decreasing the position restraint coupling constant from 25 000 to 0 kJ/mol in five equidistant steps for both the temperature and coupling constant. At each equilibration step, a short 20 ps simulation at a constant volume was carried out. This was followed by another 20 ps at 300 K and 1 atm of pressure and a subsequent production run. Constant temperature and pressure were maintained by a Berendsen thermostat (coupling time of 0.1 ps) and barostat (coupling time 0.5 ps), respectively.[14,18] All simulations were carried out using the GROMOS 45A3 force field,[19] using periodic boundary conditions. The atom numbers and partial charges for the sildenafil ligand are given in Figure 1b. The structure of vardenafil is given in Figure 1c with differences with respect to sildenafil marked with arrows. Complete GROMOS building blocks for sildenafil and vardenafil are given in the Supporting Information. Electrostatics were treated using the reaction field approach and the triple-range cutoff scheme, with cutoffs of 0.8 and 1.4 nm, and a dielectric permittivity of 61. The pair list was updated every five steps. The equations of motion were integrated using the leapfrog scheme and a step size of 2 fs. All bonds were constrained using the SHAKE algorithm[20] with a tolerance of 0.0001. Initial velocities were taken from the Maxwell−Boltzmann distribution at 300 K.

A total of 10 3-ns-long trajectories were simulated each for free PDE5A and PDE5A with sildenafil bound. The starting structure of the free enzyme was generated from the 1UDT structure by removing the ligand. Each of the 10 trajectories in the two cases was initiated with different starting velocities chosen from the Maxwell−Boltzmann distribution at 300 K. For both setups, the 10 trajectories were started from the same starting structure after the equilibration period as described above.

Thermodynamic integration was carried out using the standard λ-dependent approach, with 26 λ points spaced equidistantly between 0 and 1.[21] At each λ point, the system was simulated for 500 ps, but the first 100 ps were considered the equilibration period and were not used for calculating thermodynamic averages. To prevent instabilities, the soft-core approach was followed[22,23] with $\alpha_{ij}^{LJ} = 0.5$. Electrostatic interactions were treated using $\alpha_{ij}^{C} = 0.5$ nm^2. The area underneath the $\langle \partial H/\partial\lambda \rangle$ curves was calculated using trapezoidal integration. The statistical error at each λ point was estimated using the block averaging technique with blocks of different sizes.[24]

Binding free energy differences were also calculated using the SSP technique.[11−13] Here, we outline the basics of the technique and its present implementation: for more details, the reader is referred to one of the existing references.[11−13,25−29] The free energy difference between thermodynamic states A and R can be calculated using the perturbation formula:[30]

$$\Delta G_{AR} = -k_B T \ln\langle e^{-(E_A - E_R)/k_B T} \rangle_R \quad (1)$$

where E_A and E_R are the potential energies of the system in states A and R, respectively, k_B is the Boltzmann constant, and T is the temperature. The ensemble average, denoted by the brackets, is carried out over all the configurations of state R that can be generated in a simulation. In a typical application, states A and R, for example, represent two different ligands bound to the same receptor. In the SSP approach, the state R is a so-called reference state, a potentially unphysical ligand whose configuration space, as sampled in a simulation, exhibits significant overlap with configuration spaces of several different ligands whose relative free energies of binding one is interested in calculating.[13] Post analysis of the molecular dynamics simulations of the reference ligand free in solution and bound to a receptor allows one then to calculate the relative free energy of binding between a real ligand and the reference ligand by using eq 1. By doing this for two different ligands, one can obtain their relative free energy of binding using the following identity:

$$\Delta\Delta G_{BA} = \Delta G_{BR}(\text{bound}) - \Delta G_{BR}(\text{free}) - \Delta G_{AR}(\text{bound}) + \Delta G_{AR}(\text{free}) \quad (2)$$

The chief advantage of the SSP technique is that it allows one to, in principle, calculate relative free energies of binding of a large series of real ligands from a pair of simulations of the reference ligand. In order to expedite sampling and ensure that the sampled configurations of the reference state exhibit

significant overlap with configuration spaces of a number of real ligands, the nonbonded interaction function of the atoms in the reference ligand is typically made soft by removing the singularity at the origin. More precisely, the Lennard-Jones interaction involving atoms of choice is modified in such a way that there is a finite probability that two atoms fully overlap in space. This modification is applied to all those atoms which differ between different real ligands of interest.[22]

The atoms treated as soft are depicted with X in Figure 1d. In our implementation, the softness parameter α_{ij}^{LJ} is set to 1.51 for all soft atoms, while for electrostatic interaction, α_{ij}^{C} is set to 1.11 nm^2 for charged soft atoms.[13] The GROMOS nonbonded interaction atom types for these atoms in the soft state are those of the equivalent atoms in sildenafil. The same goes for bonded interactions. The only exceptions are the soft atoms X4 and X12 (Figure 7a), which were treated as hydrogen atoms (GROMOS nonbonded interaction atom type 18). Other GROMOS force-field types regarding these two atoms are as follows: bonds, 24−X4 and 31−X12 bond type 2; angles, 23−24−X4 and 25−24−X4 bond-angle type 24 and 30−31−X12 and 36−31−X12 bond-angle type 35; improper dihedrals, 24−23−X4−25 and 31−30−X12−36 harmonic dihedral type 1 (here, atom numbers refer to Figure 1b for the real atoms and Figure 1d for the soft atoms). Most soft atoms were assigned no charge except X6 (0.1 *e*), X7 (0.25 *e*), and X8 (0.1 *e*) to balance the negative charge on the nearby carbonyl oxygen. The simulation of the reference ligand in water was carried out in a cubic box with sides of 3.5 nm containing a total of 1585 water molecules. The simulations of PDE5A bound to the reference ligand were carried out in a truncated octahedron box with sides of 9.0 nm containing a total of 10 619 water molecules. For both the reference ligand in water and reference ligand bound to the protein, 10 independent 0.5-ns-long trajectories were run, each initiated from the same pre-equilibrated configuration with different velocity assignments. The pre-equilibration was carried out with the real sildenafil ligand as described above, followed by a 10 ps equilibration with soft parameters at 1 atm of constant pressure and subsequent production runs. In addition, one of the 10 trajectories in both cases was extended to a total length of 5 ns. In all cases, structures were saved for analysis every 0.2 ps.

Results

Structural Analysis and Mechanism of Binding. To enhance the sampling of the configuration space accessible to the molecule, we have run 10 independent 3-ns-long simulations of PDE5A without and with sildenafil bound, each simulation initiated with different velocities. The simulations are stable with respect to both the secondary and tertiary structure of the protein as well as the position of the ligand in the binding pocket. Time traces of the backbone atom-positional root-mean-square deviation (RMSD) from the experimental X-ray structure for the entire protein are shown in Figure 2a for both the unliganded and liganded PDE5A. In both cases, the average RMSD stabilizes around 0.3 nm, with only one out of 10 trajectories in each setup exceeding 0.4 nm at any point. Interestingly, the variance of the RMSD curves is appreciably greater in the presence of the ligand than without it. This fact is also mirrored in the behavior of the two divalent ions, zinc and magnesium, located in the central cavity of the protein. In the presence of sildenafil, not only does the separation between these two ions increase by approximately 0.1 nm but so does the degree of variability in this separation between different trajectories in the ensemble (Figure 2b). Visual analysis of the simulated trajectories reveals that the reason for this lies in the significant change in the coordination of the divalent ions when sildenafil is bound. In the absence of the ligand, the negatively charged side chains of residues Asp 654 and Asp 764 equally coordinate the two ions (Figure 2c on the left), resulting in a stable configuration with small fluctuations in the separation between the ions. In the presence of the ligand, these two residues provide symmetric coordination for the Zn^{2+} ion, while the Mg^{2+} ion is solely coordinated by the water molecules in the cavity (Figure 2c). This allows for its larger mobility and, subsequently, larger separation from the Zn^{2+} ion and larger overall variance of the distance between them.

What stabilizes sildenafil in the binding pocket? We have analyzed both principal contributions to sildenafil binding indicated in the literature, that is, "hydrophobic clamp" and "glutamine switch".[6,8] Both of these proposed mechanisms gain support from our simulations. In Figure 3a, we show the distribution of the angle between the planes of the purine ring in sildenafil and the phenyl ring of Phe 820, key participants in the "hydrophobic clamp" mechanism.[6] These two moieties remain largely parallel with each other, resulting in a stabilizing stacking interaction. Val 782, implicated in stabilizing sildenafil's purine ring in its position, remains in close contact with the ligand throughout the simulations (Figure 3b). Finally, if one analyzes the conserved Gln817, its amide group forms two hydrogen bonds with the purine group of sildenafil 40% of the time, at least one hydrogen bond 93% of the time, and no hydrogen bonds only 7% of the time. Such strong hydrogen bonding is an essential component of the "glutamine switch" mechanism,[8] and it is evident in our simulations as well.

In addition to the previously described "glutamine switch" and "hydrophobic clamp", our simulations suggest a third, novel contribution to ligand stabilization by phosphodiesterases, which we now term "loop clamp". In the original 1UDT X-ray structure of the molecule, parts of the H loop (residues 665−675) were missing in the density, supposedly because of their high mobility, and were modeled-in for our simulations. It is precisely this H loop, together with an opposing M loop (residues 787−812), which participates in a sizable conformational change upon binding of sildenafil (Figure 4). As the ligand binds, the H loop moves toward the M loop by more than 0.7 nm on average, with many configurations exhibiting changes in excess of 1 nm. In effect, the two loops close down on the ligand clamping it in its position. In Figure 4a, we show the average time trace of the distance between residues His 670 and Asn 798: as is evident, these two residues come closer by on average 0.6−0.8 nm in the presence of the ligand. This fact is even more clearly illustrated if one looks at the distribution of

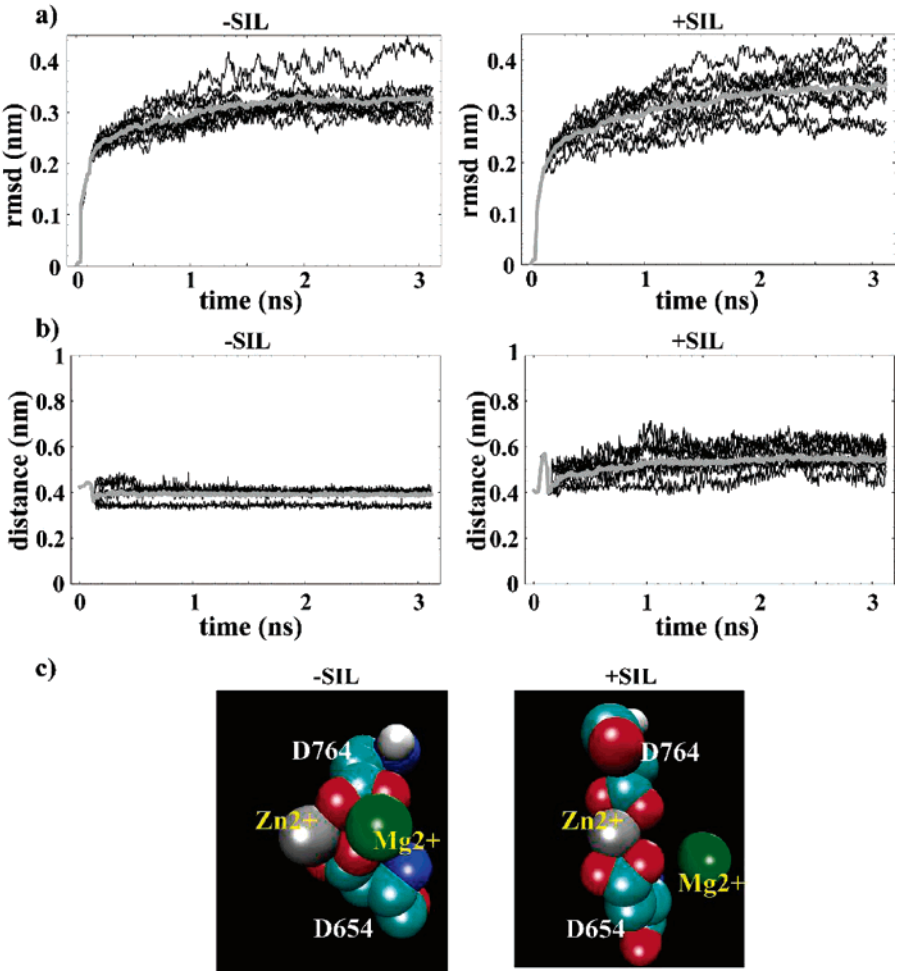


Figure 2. (a) Backbone atom-positional RMSD from the experimental X-ray structure for 10 trajectories without (left) and with (right) sildenafil bound. Backbone atoms were used for both rotational fitting and RMSD calculation. (b) Distance between two divalent ions in the core of the protein for the simulations without (left) and with (right) sildenafil bound. Average traces in a and b are shown in gray. (c) Coordination of the Zn^{2+} and Mg^{2+} ions in the active site of the enzyme without and with sildenafil bound. In the absence of sildenafil, the carboxyl groups of the residues Asp 654 and Asp 764 coordinate both ions equally. With sildenafil bound, these two residues exclusively coordinate the Zn^{2+} ion, while the Mg^{2+} ion moves more freely, being coordinated by water molecules only. This explains the difference in the fluctuations in the distance between the two ions in the two states (Figure 2b). Color code: oxygen, red; zinc ion, gray; magnesium ion, green; nitrogen, blue; carbon, cyan.

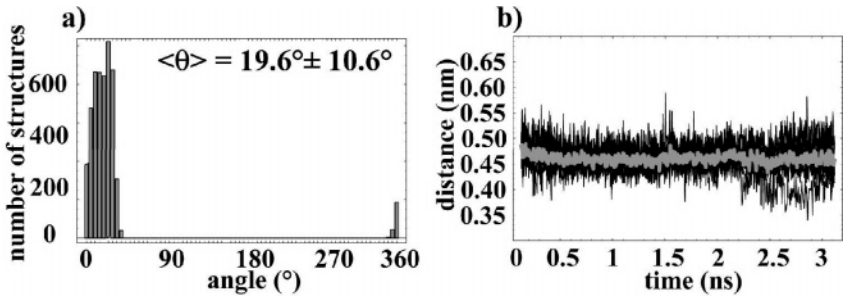


Figure 3. (a) Distribution of the angle between the plane of the purine ring of sildenafil and the phenyl ring of Phe 820. (b) Distance between CB in Val 782 and atom 35 in sildenafil over time for the 10 trajectories with sildenafil bound. The average trace is shown in gray.

this distance over the last 1.5 ns in all 10 trajectories (Figure 4b). The clamping motion of the two loops is exemplified in Figure 4c for two representative members of the ensembles without and with sildenafil bound.

Analysis of the atom-positional root-mean-square fluctuations (RMSF) of the protein around the average configuration (Figure 5a) reveals significant collective changes in the dynamics of the protein in the presence of the ligand compared to a situation with no ligand bound. Looking at the difference in RMSF when the ligand is bound versus when it is not (Figure 5b,c), calculated over the last 1.5 ns of the trajectories, one sees distinct contiguous regions of the protein whose fluctuations collectively increase, that is, decrease when the ligand is bound. In particular, fluctuations

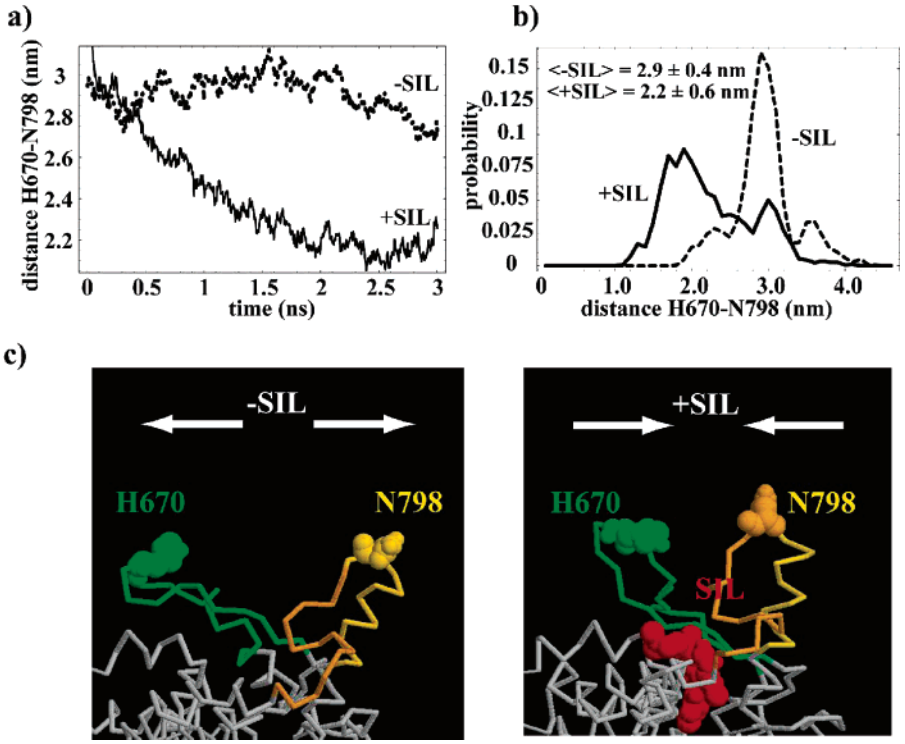


Figure 4. (a) Time traces of the Cα−Cα distance between residues His 670 and Gln 798 without and with sildenafil bound, averaged over the 10 3 ns trajectories in the two states. (b) Distributions of the Cα−Cα distance between residues His 670 and Gln 798 without and with sildenafil bound, for both ensembles over the last 1.5 ns of simulation. Distributions were binned using 0.1 nm bins. (c) Two representative structures from the simulations without and with sildenafil bound. H and M loops are outlined in green and yellow, respectively. Residues His 670 and Gln 798 are shown in an explicit van der Waals representation.

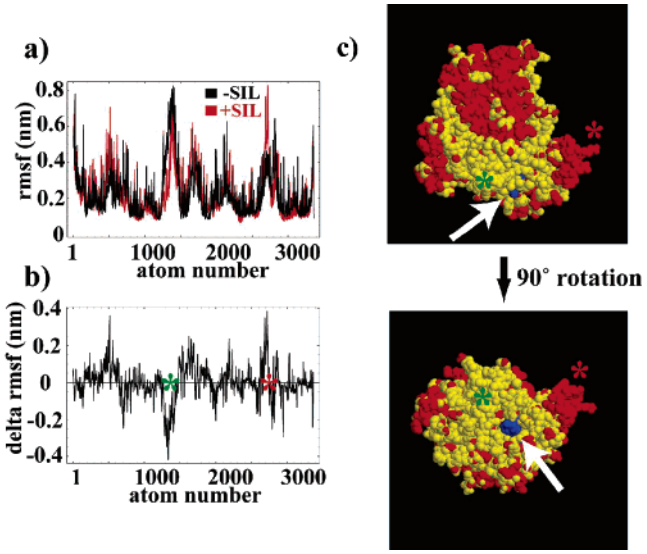


Figure 5. (a) Atom-positional root-mean-square fluctuations (RMSF) calculated with respect to the average structure without (black) and with (red) sildenafil bound. (b) Difference between the RMSF values with and without sildenafil bound, rmsf(+SIL) − rmsf(−SIL). (c) Values from b mapped onto the structure of PDE5A (yellow, RMSF decreases with sildenafil bound; red, RMSF increases with sildenafil bound). Position of the sildenafil binding pocket is indicated with an arrow. In b and c, locations of the H and M loops are marked with green and red stars, respectively.

of the protein around the binding pocket decrease in the presence of the ligand, while the diametrically opposite apical region of the enzyme largely increases in motion (Figure 5c). The behavior of the loops surrounding the binding pocket is particularly noteworthy. The M loop increases in its fluctuations despite getting closer to the H loop. On the other hand, the H loop, which undergoes a much larger relative motion, decreases in fluctuations participating more strongly in relative terms in clamping the ligand in the binding pocket.

Thermodynamics. *Thermodynamic Integration.* Standard λ-dependent thermodynamic integration[21] was carried out to calculate relative free energies of binding of sildenafil and vardenafil. The two molecules differ by the type (N versus C) of only two atoms (25 and 36) of the purine ring, but, in addition, vardenafil has an extra methyl group attached to the piperazine ring (Figure 1b,c). The latter has been shown to make no significant contribution when it comes to the free energy of binding: vardenafil with no methyl group (demethyl-vardenafil) still binds to PDE5A about 10 times more potently than sildenafil.[10] Average $\partial H/\partial\lambda$ curves for the sildenafil-to-vardenafil conversion are shown in Figure 6a (ligand in water and in the binding pocket). Numerical integration of these curves yields a relative free energy of binding of −8.0 kJ/mol, which is in agreement with experimental results. Measurements of the IC_{50} values for PDE5A give a range of 1−7 nM for sildenafil[31,32] and 0.17 nM for vardenafil.[10] This corresponds to a free energy difference of anywhere between −9.3 and −4.4 kJ/mol. Measurements of K_D values, which are a better indication of affinity, yield a range of 8.3−13 nM for sildenafil[10] and 1 nM for vardenafil.[33] This equals a range from −5.8 to −5.3 kJ/mol, when it comes to free energy differences.

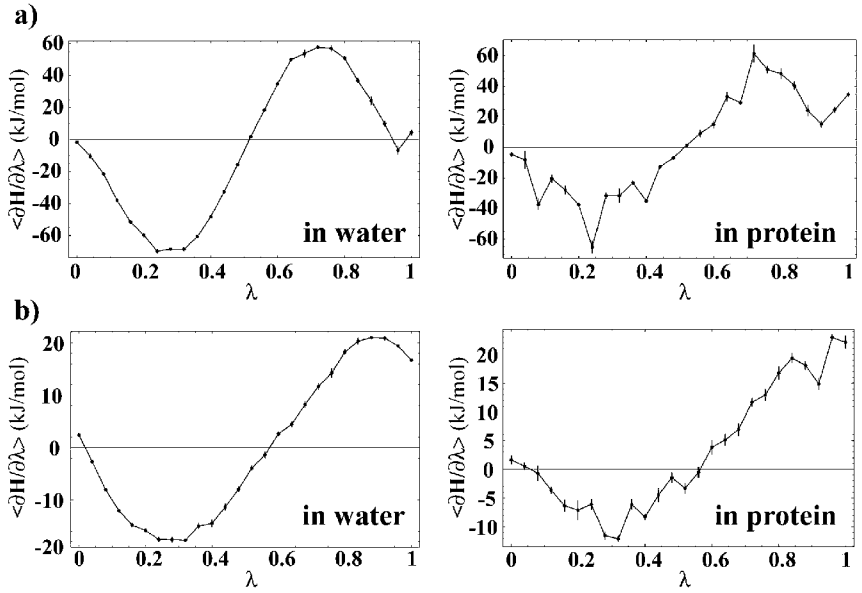


Figure 6. Average thermodynamic integration $\langle \partial H/\partial \lambda \rangle$ curves for simulations of the free ligand ("in water") or ligand bound to the protein ("in protein") for (a) vardenafil ($\lambda = 0$) to sildenafil ($\lambda = 1$) conversion and (b) vardenafil ($\lambda = 0$) to demethyl-vardenafil ($\lambda = 1$) conversion. Vertical bars refer to the statistical error at each λ value (see Methods for details).

As a control, we have also performed another 26-point λ-dependent thermodynamic integration for the conversion from vardenafil to demethyl-vardenafil.[10] Average $\partial H/\partial \lambda$ curves for this conversion are shown in Figure 6b: we calculate a relative free energy of binding of 3.4 kJ/mol. The range of experimental values for this free energy difference is anywhere between -2.1 (based on K_D values) and -0.5 kJ/mol (based on IC_{50} values), indicating essentially no significant effect of the piperazine methyl group on the binding. While the simulations do not ideally match the experimental data, they still provide a qualitatively correct picture of binding: these results give us confidence in the force field used as well as in the parametrization of the two ligands.

Single Step Perturbation. In order to study the effects of computational parallelization in the context of SSP, we have first simulated 10 independent 0.5 ns trajectories of the reference ligand (Figure 1d) in the protein and another 10 free in solution, each trajectory being initiated from the same configuration, but with different velocity assignments. Second, we have extended one of the 10 short simulations of the soft reference ligand, chosen at random, to a total length of 5 ns, once in the protein and once free in solution. As shown in Figure 7a, the backbone atom-positional root-mean-square deviation for the entire protein from the experimental structure for both setups parallels to a large degree the results of our simulations with the real sildenafil ligand bound (Figure 2a). The RMSD in the case of the 5 ns simulation reaches about 0.3 nm on average, while in the case of the 0.5 ns trajectories, it remains around 0.25 nm with only one trajectory exceeding 0.35 nm at one point.

How does the sampling in the two cases compare? Equivalently, how diverse are the structures generated in the 5 ns trajectory as opposed to the ones coming from the composite ensemble consisting of the 10 0.5 ns trajectories? One way to assess this is shown in Figure 7b. We have

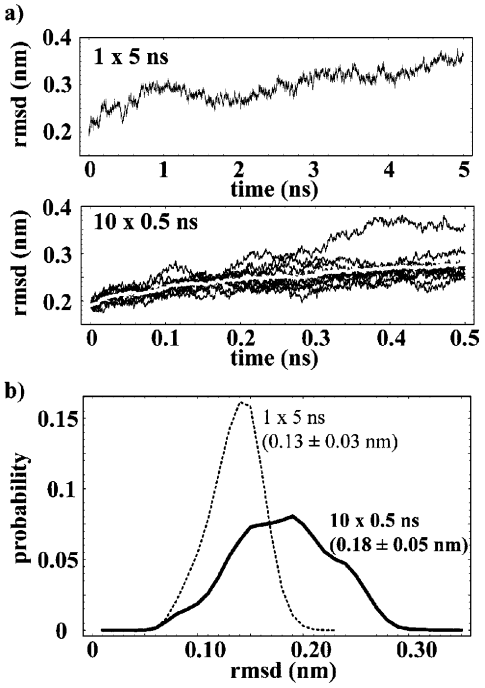


Figure 7. (a) Time traces of the backbone atom-positional RMSD from the experimental X-ray structure for the 5-ns-long trajectory and 10 0.5-ns-long trajectories with the soft ligand bound. The average trace for the ensemble is shown in gray. (b) Distributions of all-against-all atom-positional RMSD values for the structures from the 10 × 0.5 ns ensemble and the structures from the 1 × 5 ns trajectory. Only protein atoms within 0.5 nm from the nearest ligand atom were used for rotational fitting and RMSD calculation. The mean and standard deviation of each distribution is indicated. Distributions were binned using 0.01 nm bins.

calculated an all-against-all atom-positional RMSD matrix for 1000 structures, spaced every 5 ps, for each of the two ensembles, where for rotational fitting and RMSD calculation

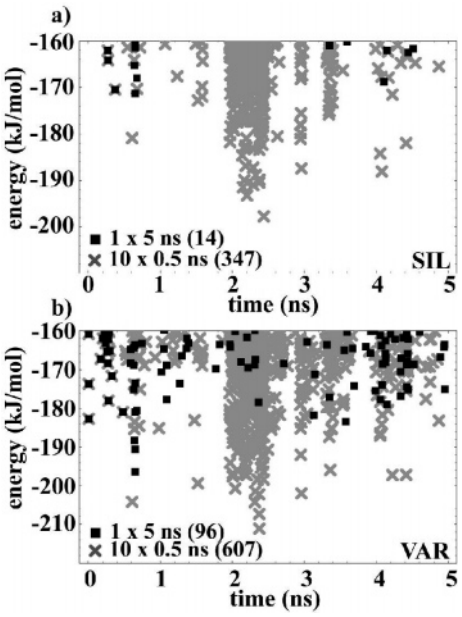


Figure 8. Low-lying energy configurations (<−160 kJ/mol) for an ensemble (10 × 0.5 ns, crosses) or a single trajectory (1 × 5 ns, boxes) for bound (a) sildenafil or (b) vardenafil. Energies are nonbonding energies for all the atoms that were soft in the reference state simulations. The number of individual configurations whose energy is less than −160 kJ/mol is indicated in the graphs. In the case of the 10 × 0.5 ns ensembles, trajectories were concatenated in post analysis to yield one 5-ns-long trajectory.

we have used only those atoms of the protein which are 0.5 nm or less away from the nearest atom of the bound ligand. Interactions between these atoms in the protein and the ligand are supposed to give the largest contribution to the final free energy of binding. In Figure 7b, we show the distribution of the RMSD values generated in this way: strikingly, even though one has used the same amount of processing time in the two cases, the ensemble approach with short trajectories yields a significantly greater diversity of structures according to the RMSD measure than does the single long trajectory. In particular, the average RMSD between the members of the single-trajectory ensemble is 0.13 ± 0.02 nm, while this number climbs to 0.18 ± 0.04 nm in the case of the multiple-trajectory ensemble.

How do the energies differ between the two setups? In Figure 8, we show total nonbonding energy values of the lowest-energy configurations coming from the two setups, calculated for the built-in sildenafil and vardenafil ligands. More precisely, energies were calculated for all the atoms that were simulated using the soft-core interaction, but after replacement by real-atom parameters, as required by a given real ligand. As can be seen, short parallel trajectories result in a significantly larger number of low-energy configurations compared to a long trajectory of the same overall length, and this is true for both ligands. For example, while the 5 ns trajectory results in 14 configurations of bound sildenafil with energy below −160 kJ/mol, this number rises to a total of 347 in the case of the 10 0.5 ps trajectories (Figure 8a). This ratio changes to 96 versus 607 configurations, respectively, in the case of vardenafil (Figure 8b). In general, the entire distribution of energy of the configurations coming from the ensemble of short trajectories is shifted toward lower energy values compared to the long trajectory (Figure 9b,d). However, this is true only for the simulations of the ligand in the binding site of the protein: parallelization results in no advantage when it comes to simulations of the ligand free in solution (Figure 9a,c). This is a direct manifestation of the fact that, in the allotted time, the single long trajectory manages to explore the majority of the relevant phase space volume available to the ligand free in solution. As the system behaves ergodically, there is then no difference between running a number of independent trajectories or just one long one. On the other hand, the barriers between low-energy conformations of the ligand are larger in the presence of the protein, which slows down the sampling and consequently makes the ensemble approach more advantageous. This advantage comes from the fact that, by choosing different starting velocities for the 0.5 ns replicas, one already from the start samples different regions of the phase space. As the barriers between these regions can be sizable, they do not get traversed in the course of the 5 ns trajectory, resulting in that case in exploring only a limited section of the phase space.

Finally, the advantage of the ensemble approach reflects itself in the cumulative curves depicting the buildup of the relative binding free energies of the ligands. Sharp, negative deflections in the ΔG_{AR} (relative free energy of the real ligands with respect to the soft state) curves are indicative of finding low-energy configurations that significantly contribute to the final ΔG_{AR} value. As is evident from Figure 10, there is an appreciably greater number of such configurations coming from the ensemble set as opposed to the single trajectory for both ligands. However, even more important is the fact that the final average relative free energy of binding is significantly more negative in the case of the ensemble approach: running several short trajectories manages to pick out low-energy configurations much more efficiently than does one single, long trajectory.

Discussion

Our observation of significant conformational changes in the H and M loops of PDE5A upon the binding of sildenafil is interesting as it represents a third, novel contribution to the stability of the ligand in the binding pocket, in addition to the previously described "glutamine switch"[8] and "hydrophobic clamp".[6] The two loops cover up the ligand in the binding pocket, clamping it in its position, and hence we term this feature of the enzyme the "loop clamp". Here, it should be emphasized that, in this clamping motion, the H loop undergoes a significantly greater motion compared to the M loop, which remains close to its average unbound position but starts fluctuating around it more significantly (Figure 5c). When our computational study was being initiated, there was no complete structure of the H loop available, and we resorted to modeling techniques to build it into the initial structure. As our study was already completed, a series of new X-ray structures of PDE5A appeared,[34] including, most importantly, structures of the enzyme with and without sildenafil bound, both with the H

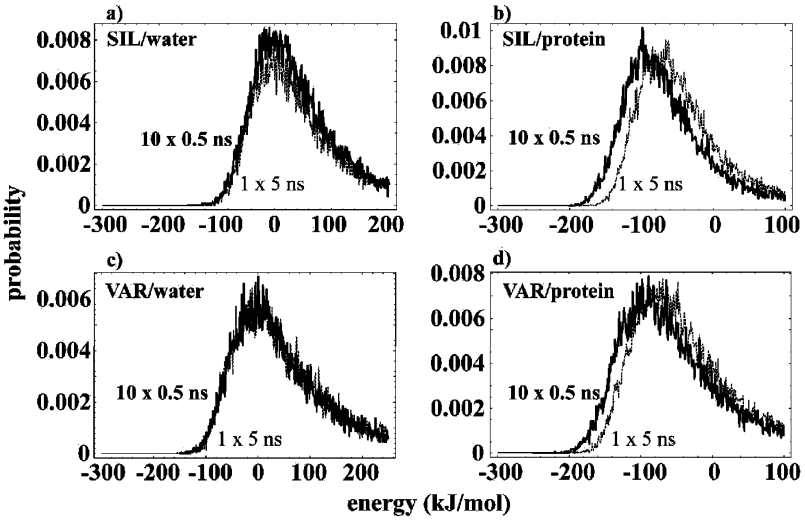


Figure 9. Distributions of nonbonding energies for an ensemble (10 × 0.5 ns, thick lines) or a single trajectory (1 × 5 ns, dashed thin line) for (a) sildenafil or (b) vardenafil in water (left) or in the protein (right). Energies are nonbonding energies for all the atoms that were soft in the reference state simulations. All distributions were binned in 1 kJ/mol bins.

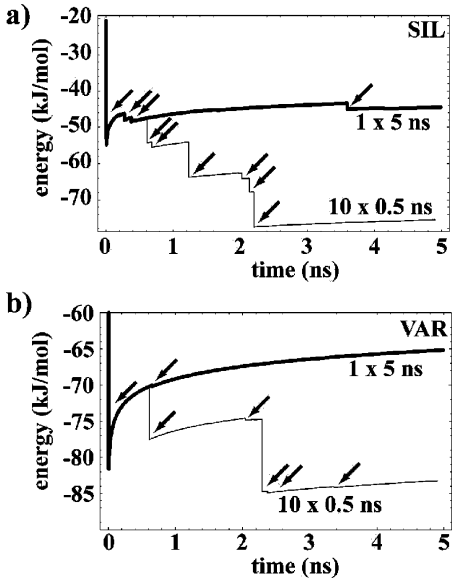


Figure 10. Development of the ensemble average value for ΔG_{AR} (relative free energy of a real ligand with respect to the soft ligand) for (a) sildenafil and (b) vardenafil. Configurations, which contribute significantly to the average, cause discontinuities in the graphs and are indicated by arrows. In the case of the 10 × 0.5 ns ensembles, trajectories were concatenated in post analysis to yield one 5-ns-long trajectory.

loop fully resolved. In this study, Ke and co-workers observed large-scale motions of the loop upon binding of the ligand, which are in many respects in good agreement with our observations. In their structures, the H loop executes an approximately 2 nm motion forming close van der Waals contacts with the piperazine moiety of sildenafil, resulting in almost complete burial of the ligand in the binding pocket. This is highly analogous to what we see, except that in our simulations the two dominant conformations of the H loop are slightly less apart (Figure 4b). Interestingly, Ke and co-workers also note small motions of the M loop toward the H loop, again analogous to what we see. Here, it should be mentioned that, even in the presence of the ligand, a small subset of our simulated structures remains in an unclamped position with H and M loops spread apart. In their measurements of the binding constant of sildenafil to PDE5A, Corbin and co-workers[33] observed two different dissociation constants for the ligand, a slow one and a fast one. It is possible that the fast and the slow dissociation constants reflect dissociation from a heterogeneous ensemble of structures, some of which have the loops spread apart and some of which have them clamped together. In the latter case, for dissociation to occur, the loops would first have to open up, resulting in a slower apparent dissociation rate.

Despite the improvements in sampling achieved through parallelization, the SSP results for the relative free energies of binding of vardenafil, sildenafil, and demethyl-vardenafil are at odds with the experimental findings as well as thermodynamic integration results. In Table 1, we summarize the relative free energies of binding and their components for vardenafil versus sildenafil and vardenafil versus demethyl-vardenafil. In both cases, the SSP results are both quantitatively and qualitatively at odds with the experimental results. While the experiment suggests the ordering of vardenafil, demethyl-vardenafil, and then sildenafil, going from the strongest to the weakest binder, the SSP results give an exactly opposite ordering, with sildenafil being the best binder. If one looks at the single 5 ns trajectory and calculates the free energy estimates over its first and second halves, one gets significantly different values for the ligand in the protein. This is a strong indication that the trajectory is not covering all of the relevant configurations in the time allotted. Interestingly, the first 2.5 ns of this trajectory yield a free energy difference for vardenafil and sildenafil (−3.6 kJ/mol) which is close to the experimentally determined value. This agreement, however, is purely fortuitous. In general, the free energies of the ligand in the protein are likely the ones causing the overall disagreement with experimental results: the convergence of the free energy values for the ligand in water is in general much better.

The fact that the thermodynamic integration results are in a much better agreement with experimental results suggests

Table 1. Free Energy Difference in Water and in the Protein for Vardenafil versus Sildenafil (var-sil) and Vardenafil versus Demethyl-Vardenafil (var-dmvar) Using the Single-Step Perturbation (SSP) Technique[a]

	water (kJ/mol)	protein (kJ/mol)	total (kJ/mol)
$\Delta\Delta G_{var-sil}$ (1×)	−18.3	−20.7	−2.4
$\Delta\Delta G_{var-sil}$ (1×, 0−2.5 ns)	−19.0	−22.6	−3.6
$\Delta\Delta G_{var-sil}$ (1×, 2.5−5 ns)	−16.0	−10.1	5.9
$\Delta\Delta G_{var-sil}$ (10×)	−18.0	−7.8	10.2
$\Delta\Delta G_{var-sil}$ (merged)	**−17.9**	**−7.8**	**10.1**
$\Delta\Delta G_{var-sil}$ (experiment)	**n/a**	**n/a**	**−9.3 to −4.4**
$\Delta\Delta G_{var-dmvar}$ (1×)	−4.1	−9.9	−5.8
$\Delta\Delta G_{var-dmvar}$ (1×, 0−2.5 ns)	−4.6	−11.9	−7.3
$\Delta\Delta G_{var-dmvar}$ (1×, 2.5−5 ns)	−2.3	0.7	3.0
$\Delta\Delta G_{var-dmvar}$ (10×)	−4.3	3.2	7.5
$\Delta\Delta G_{var-dmvar}$ (merged)	**4.3**	**3.2**	**7.5**
$\Delta\Delta G_{var-dmvar}$ (experiment)	**n/a**	**n/a**	**−2.1 to −0.5**

[a] The values are given for the single 5 ns trajectory (1×), the first and the last 2.5 ns of this trajectory (0−2.5 ns and 2.5−5 ns), a composite ensemble consisting of the 10 0.5 ns trajectories (10×), and another composite ensemble in which the 5 ns trajectory was merged with the 10 × 0.5 ns ensemble (merged). Experimental values were based on IC_{50} and K_D values (see text for details).

that the shortcomings of the SSP technique may not be attributed to the deficiencies of the force field used. Rather, the problems are likely to be traced to insufficient sampling as well as to the properties of the reference ligand employed. The reference ligand was designed with mostly neutrally charged soft atoms in the purine ring. The fact that the real ligands vardenafil and sildenafil possess significant partial charges in places where the reference ligand has none, and which are in key positions of opposite polarity, may cause the configurations sampled with the reference ligand to exhibit only little overlap with the configurational space of the real ligand, resulting in poor approximations of the relative binding free energy. This problem has been encountered before, and it represents one of the pressing challenges hindering the further development of the technique. Initially, our aim was to use the SSP technique to calculate the relative binding free energies for a series of different ligands of similar geometry, most of which have not even been synthesized or experimentally tested. This is the reason for using soft atoms in more positions than would be warranted just by the differences between the three ligands that we have analyzed here. However, the unsatisfying agreement with the experiment for these three ligands suggests that such attempts may be premature.

While the results of the SSP calculations may not be fully satisfactory, the demonstration of significant improvements in sampling obtained via brute force parallelization gives hope that further tests and improvements of the method could now be carried out more quickly and with better coverage of the phase space. SSP is essentially a thermodynamic technique, meaning that the exact trajectories of the soft state per se are of no particular interest. What matters is that these trajectories sample well the phase space available to ligands and receptors. In this sense, running multiple short trajectories should pose no problems. One could argue that, if one is interested in kinetics and mechanisms, having single trajectories that are long enough to capture conformational changes of interest, that is, barrier crossing events, may be of advantage. In the case of thermodynamics, using short trajectories should not be a problem, as long as they are long enough to sample well the local free energy minima. We hope that computational parallelization will lead to improvements of the SSP method itself, such that it can reach the levels of accuracy of the much more costly thermodynamic integration.

Acknowledgment. B.Z. acknowledges support from an EMBO postdoctoral fellowship. This work was financially supported by grants from the National Center of Competence in Research (NCCR) in Structural Biology and by grant number 200021-109227 of the Swiss National Science Foundation, which is gratefully acknowledged.

Supporting Information Available: GROMOS 45A3 building blocks for sildenafil and vardenafil are given. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Soderling, S. H.; Beavo, J. A. Regulation of cAMP and cGMP Signaling: New Phosphodiesterases and New Functions. *Curr. Opin. Cell Biol.* **2000**, *12*, 174−179.

(2) Beavo, J. A. Cyclic Nucleotide Phosphodiesterases: Functional Implications of Multiple Isoforms. *Physiol. Rev.* **1995**, *75*, 725−748.

(3) Mehats, C.; Andersen, C. B.; Filopanti, M.; Jin, S. L.; Conti, M. Cyclic Nucleotide Phosphodiesterases and Their Role in Endocrine Cell Signaling. *Trends Endocrinol. Metab.* **2002**, *13*, 29−35.

(4) Torphy, T. J. Phosphodiesterase Isozymes: Molecular Targets for Novel Antiasthma Agents. *Am. J. Respir. Crit. Care Med.* **1998**, *157*, 351−370.

(5) Corbin, J. D.; Francis, S. H. Cyclic GMP Phosphodiesterase-5: Target of Sildenafil. *J. Biol. Chem.* **1999**, *274*, 13729−13732.

(6) Card, G. L.; England, B. P.; Suzuki, Y.; Fong, D.; Powell, B.; Lee, B.; Luu, C.; Tabrizizad, M.; Gillette, S.; Ibrahim, P. N.; Artis, D. R.; Bollag, G.; Milburn, M. V.; Kim, S. H.; Schlessinger, J.; Zhang, K. Y. Structural Basis for the Activity of Drugs that Inhibit Phosphodiesterases. *Structure* **2004**, *12*, 2233−2247.

(7) Manallack, D. T.; Hughes, R. A.; Thompson, P. E. The Next Generation of Phosphodiesterase Inhibitors: Structural Clues to Ligand and Substrate Selectivity of Phosphodiesterases. *J. Med. Chem.* **2005**, *48*, 3449−3462.

(8) Zhang, K. Y.; Card, G. L.; Suzuki, Y.; Artis, D. R.; Fong, D.; Gillette, S.; Hsieh, D.; Neiman, J.; West, B. L.; Zhang, C.; Milburn, M. V.; Kim, S. H.; Schlessinger, J.; Bollag, G. A Glutamine Switch Mechanism for Nucleotide Selectivity by Phosphodiesterases. *Mol. Cell* **2004**, *15*, 279−286.

(9) Sung, B. J.; Hwang, K. Y.; Jeon, Y. H.; Lee, J. I.; Heo, Y. S.; Kim, J. H.; Moon, J.; Yoon, J. M.; Hyun, Y. L.; Kim, E.; Eum, S. J.; Park, S. Y.; Lee, J. O.; Lee, T. G.; Ro, S.; Cho, J. M. Structure of the Catalytic Domain of Human

Phosphodiesterase 5 with Bound Drug Molecules. *Nature* **2003**, *425*, 98−102.

(10) Corbin, J. D.; Beasley, A.; Blount, M. A.; Francis, S. H. Vardenafil: Structural Basis for Higher Potency over Sildenafil in Inhibiting cGMP-Specific Phosphodiesterase-5 (PDE5). *Neurochem. Int.* **2004**, *45*, 859−863.

(11) Liu, H.; Mark, A. E.; van Gunsteren, W. F. Estimating the Relative Free Energy of Different Molecular States with Respect to a Single Reference State. *J. Phys. Chem.* **1996**, *100*, 9485−9494.

(12) Pitera, J. W.; van Gunsteren, W. F. One-Step Perturbation Methods for Solvation Free Energies of Polar Solutes. *J. Phys. Chem. B* **2001**, *105*, 11264−11274.

(13) Oostenbrink, C.; van Gunsteren, W. F. Single-Step Perturbations to Calculate Free Energy Differences from Unphysical Reference States: Limits on Size, Flexibility, and Character. *J. Comput. Chem.* **2003**, *24*, 1730−1739.

(14) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; Biomos: Zurich, Switzerland, 1996.

(15) Berendsen, H. J.; Postma, J. P.; van Gunsteren, W. F.; Hermans, J. *Intermolecular Forces*; Reidel: Dordrecht, The Netherlands, 1981; pp 331−342.

(16) Fiser, A.; Do, R. K.; Sali, A. Modeling of Loops in Protein Structures. *Protein Sci.* **2000**, *9*, 1753−1773.

(17) Fiser, A.; Sali, A. ModLoop: Automated Modeling of Loops in Protein Structures. *Bioinformatics* **2003**, *19*, 2500−2501.

(18) Berendsen, H. J. C.; Postma, J. P.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(19) Schuler, L. D.; Daura, X.; van Gunsteren, W. F. An Improved GROMOS96 Force Field for Aliphatic Hydrocarbons in the Condensed Phase. *J. Comput. Chem.* **2001**, *22*, 1205−1218.

(20) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.

(21) van Gunsteren, W. F.; Beutler, T. C.; Fraternali, F.; King, P. M.; Mark, A. E.; Smith, P. E. *Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications*; ESCOM Science Publishers: Leiden, The Netherlands, 1993; pp 315−348.

(22) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. Avoiding Singularities and Numerical Instabilities in Free Energy Calculations Based on Molecular Simulations. *Chem. Phys. Lett.* **1994**, *222*, 529−539.

(23) Zacharias, M.; Straatsma, T. P.; McCammon, J. A. Separation-Shifted Scaling, a New Scaling Method for Lennard-Jones Interactions in Thermodynamic Integration. *J. Chem. Phys.* **1994**, *100*, 9025−9031.

(24) Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Oxford University Press Inc.: New York, 1989.

(25) Oostenbrink, B. C.; Pitera, J. W.; van Lipzig, M. M.; Meerman, J. H.; van Gunsteren, W. F. Simulations of the Estrogen Receptor Ligand-Binding Domain: Affinity of Natural Ligands and Xenoestrogens. *J. Med. Chem.* **2000**, *43*, 4594−4605.

(26) Oostenbrink, C.; van Gunsteren, W. F. Free Energies of Binding of Polychlorinated Biphenyls to the Estrogen Receptor from a Single Simulation. *Proteins* **2004**, *54*, 237−246.

(27) Oostenbrink, C.; van Gunsteren, W. F. Efficient Calculation of Many Stacking and Pairing Free Energies in DNA from a Few Molecular Dynamics Simulations. *Chem.—Eur. J.* **2005**, *11*, 4340−4348.

(28) Oostenbrink, C.; van Gunsteren, W. F. Free Energies of Ligand Binding for Structurally Diverse Compounds. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6750−6754.

(29) Oostenbrink, C.; Soares, T. A.; van der Vegt, N. F.; van Gunsteren, W. F. Validation of the 53A6 GROMOS Force Field. *Eur. Biophys. J.* **2005**, *34*, 273−284.

(30) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420−1426.

(31) Ballard, S. A.; Gingell, C. J.; Tang, K.; Turner, L. A.; Price, M. E.; Naylor, A. M. Effects of Sildenafil on the Relaxation of Human Corpus Cavernosum Tissue in Vitro and on the Activities of Cyclic Nucleotide Phosphodiesterase Isozymes. *J. Urol.* **1998**, *159*, 2164−2171.

(32) Corbin, J. D.; Turko, I. V.; Beasley, A.; Francis, S. H. Phosphorylation of Phosphodiesterase-5 by Cyclic Nucleotide-Dependent Protein Kinase Alters Its Catalytic and Allosteric cGMP-Binding Activities. *Eur. J. Biochem.* **2000**, *267*, 2760−2767.

(33) Corbin, J. D.; Blount, M. A.; Weeks, J. L., II; Beasley, A.; Kuhn, K. P.; Ho, Y. S.; Saidi, L. F.; Hurley, J. H.; Kotera, J.; Francis, S. H. [3H]Sildenafil Binding to Phosphodiesterase-5 is Specific, Kinetically Heterogeneous, and Stimulated by cGMP. *Mol. Pharmacol.* **2003**, *63*, 1364−1372.

(34) Wang, H.; Liu, Y.; Huai, Q.; Cai, J.; Zoraghi, R.; Francis, S. H.; Corbin, J. D.; Robinson, H.; Xin, Z.; Lin, G.; Ke, H. Multiple Conformations of Phosphodiesterase-5: Implications for Enzyme Function and Drug Development. *J. Biol. Chem.* **2006**, *281*, 21469−21479.

CT600322D

JCTC Journal of Chemical Theory and Computation

Assessment of Detection and Refinement Strategies for de novo Protein Structures Using Force Field and Statistical Potentials

Michael S. Lee*,†,‡ and Mark A. Olson‡

Computational and Information Sciences Directorate, U.S. Army Research Laboratory, Aberdeen Proving Ground, Maryland 21005, and Department of Cell Biology and Biochemistry, U.S. Army Medical Research Institute of Infectious Diseases, Frederick, Maryland 21702

Received June 12, 2006

Abstract: De novo predictions of protein structures at high resolution are plagued by the problem of detecting the native conformation from false energy minima. In this work, we provide an assessment of various detection and refinement protocols on a small subset of the second-generation all-atom Rosetta decoy set (Tsai et al. *Proteins* **2003**, *53*, 76−87) using two potentials: the all-atom CHARMM PARAM22 force field combined with generalized Born/surface-area (GB-SA) implicit solvation and the DFIRE-AA statistical potential. Detection schemes included DFIRE-AA conformational scoring and energy minimization followed by scoring with both GB-SA and DFIRE-AA potentials. Refinement methods included short-time (1-ps) molecular dynamics simulations, temperature-based replica exchange molecular dynamics, and a new computational unfold/refold procedure. Refinement methods include temperature-based replica exchange molecular dynamics and a new computational unfold/refold procedure. Our results indicate that simple detection with only minimization is the best protocol for finding the most nativelike structures in the decoy set. The refinement techniques that we tested are generally unsuccessful in improving detection; however, they provide marginal improvements to some of the decoy structures. Future directions in the development of refinement techniques are discussed in the context of the limitations of the protocols evaluated in this study.

1. Introduction

Protein structure prediction is becoming an increasingly important part of the biologist's toolkit as the number of protein-encoding DNA sequences from genomic studies vastly outnumbers the available experimentally obtained protein structures. Structure prediction has been tackled by a variety of strategies depending on the similarity of a target amino acid sequence to known protein structures. Comparative modeling is used when the target sequence is very close to one or more known protein structures.[1] Fold prediction and threading are employed when the sequence can be matched through profile similarities with one or more known structures.[2] Finally, with little perceived similarity to known folds, de novo algorithms generate protein structures either by united-residue folding simulations[3] or fragment assembly.[4]

The Rosetta program from the Baker group[4] is considered one of the top methods for de novo structure predictions. Traditionally, de novo folding has been used as a last resort for protein structure prediction. The Rosetta protocol has proven to be very powerful for predicting structures where the fold and its subsequent template alignment can be guessed, but the fold prediction is less than certain.[5] Rosetta can generate structures of low to medium resolution in many cases, although detecting such structures as being near-native is frequently difficult.[6−8] Near-native, in the context of this work, refers to structural models whose root-mean-square-

* Corresponding author e-mail: michael.lee@amedd.army.mil.
† U.S. Army Research Laboratory.
‡ U.S. Army Medical Research Institute of Infectious Diseases.

10.1021/ct600195f
Published on Web 12/14/2006

deviation (rmsd) of their alpha-carbon backbone (C_α) are within 2−3 Å of the experimentally determined structure. Often for a given protein target, between 10 000 to 100 000 models must be generated for a few models to be near-native structures. Also, the more near-native structures that are generated, the greater the likelihood an atom-based scoring function will be able to detect one or more of the near-native structures as the best scoring. Two criteria must be satisfied, however, to make successful detection and refinement possible. First, the scoring function should score the native as lowest in energy compared to any misfolded structures. In addition, it is necessary that as the native structure is approached, as can be measured by various native-biased metrics such as rmsd or fraction of native contacts, the scores trend toward the native value. This requirement, which we will call a "scoring funnel" is analogous to the folding funnel, whereby real proteins move on a folding trajectory that take on the native fold in a finite time due to some leaning, however slight, toward the lowest free-energy basin.[9] One caveat in the connection between the scoring funnel and the folding funnel is that the scoring function often lacks some or all of the entropy contributions.[10]

Several refinement protocols have been considered in the literature, although the problem remains largely unsolved.[11] Presently, a grand challenge problem is to consistently refine low- to medium-resolution protein structure predictions (e.g., C_α rmsd > 4 Å) to the accuracy necessary for drug-based design (e.g., C_α rmsd < 2.5 Å.) Recent efforts have included the work of Lu and Skolnick,[12] which evaluated the effect of short simulations (∼50 ps) using force field and knowledge-based potentials. Misura and Baker[7] outlined a scheme of making random perturbations to the original Rosetta models, which works well in tandem with their homology-based enrichment procedure.[8] Fan and Mark[13] investigated the use of long-time molecular dynamic simulations (>10 ns) in improving initial models. In cases, where only small segments of a protein need to be "refined" (i.e., nonconserved regions of a homology model), configurational enumeration techniques can be quite successful.[14−16] Nevertheless, larger nonconserved regions (e.g., number of residues, N_{res} > 11) are still difficult to model because the number of plausible conformations increases exponentially with the number of residues.

A priori knowledge of which protein conformations in a large set of structures are near-native is an unsolved problem because of three reasons. First, the side-chain packing may not be correct, even if the backbone is near-native. In this case, the high-resolution scoring function will often fail. Second, the best structures may not be within the "radius of convergence" of the native basin for a given energy function.[8] Finally, the high-resolution energy function may sometimes assign a lower energy to a non-native structure compared to the native or near-native conformation.

The potential or scoring functions to discriminate and refine protein structures are currently based on three methods: force-field based[10,17,18] and knowledge-based[19] and hybrids of the two.[6,7,20] Force-field based detection functions often employ a standard parameter set such as CHARMM PARAM22[21] or AMBER[22] and an implicit solvent function such as generalized Born(GB)[23] or Poisson−Boltzmann.[24] One of the goals of this work is to compare two different but exemplary scoring functions, PARAM22/GB-SA[17,25] and the all-atom distance-scaled ideal-gas reference state (DFIRE-AA) statistical potential,[19,26] for detection and refinement. The SA denotes a simple solvent-accessible, surface area-based treatment of the hydrophobic effect. PARAM22/GB-SA exhibited one of the best detection capabilities among several force-field based functions in an assessment of CASP4 protein structures, where the specific model of GB was GBMV2.[17] GBMV2 is a molecular-volume dielectric boundary implicit solvent model which does a good job in mimicking the results of more expensive Poisson solvation calculations.[25]

DFIRE-AA, on the other hand, is very good at distinguishing the native structure from non-native conformations for a wide variety of decoy sets.[27] Statistical potential approaches have also been successfully employed in the drug-docking problem to detect native poses and estimate binding affinities.[27,28] Also studied is the ability of such functions to detect near-native structures[3,29,30] or optimal alignments of structural templates in homology models.[31,32] Statistical potentials are developed from the growing database of crystal structures in the Protein Data Bank (PDB).[33] The traditional method involves analyzing the probability distributions and subsequently the potentials of mean force along the distances between pairs of atoms.

In this work, we introduce a hybrid force field for molecular dynamics (MD) simulations that combines a continuous version of the DFIRE-AA statistical potential with the internal energies and van der Waals interactions of a united-atom force field.[12] Interestingly, MD simulations using this hybrid potential quickly condense the protein and trap it in a local minimum. To take advantage of the rapidity of condensing a protein structure, we developed a method that quickly unfolds and refolds a protein model, thereby generating hundreds of new protein models which can be scored by the DFIRE-AA or any other discriminating energy function. The hope is that some of the newly generated protein models will be lower in energy and closer to the native structure.

We first perform a standard comparison between the all-atom PARAM22/GB-SA potential[25] and the DFIRE-AA statistical potentials[34] for detection of native and near-native protein structures using several sets of Rosetta-generated protein conformations. We then perform replica exchange simulations using separately the all-atom potential and an MD-adapted form of the statistical potential. Replica exchange entails running several parallel simulation windows spanning a range of temperatures[35] whereby periodically exchanges of temperature between windows are performed based on a Metropolis Monte Carlo criterion. As a final method, we look at unfolding/refolding of model structures using the hybrid force-field/statistical potential.

2. Theory and Methods

2.1. Potentials. The DFIRE-AA statistical potential is one of several knowledge-based potentials described in the literature.[29,36] It is defined as[19]

$$u_{\mathrm{DFIRE}}(i,j,r) = -\eta k_B T \ln\left[\frac{N_{\mathrm{obs}}(i,j,r)}{\left(\frac{r}{r_{\mathrm{cut}}}\right)^{\alpha}\frac{\Delta r}{\Delta r_{\mathrm{cut}}}N_{\mathrm{obs}}(i,j,r_{\mathrm{cut}})}\right] \quad (1)$$

where i and j are non-hydrogen atom types, r is a pairwise distance, r_{cut} is the cutoff beyond which pairwise interactions are neglected, Δr is the histogram bin size, N_{obs} is a cumulative histogram of the observed occurrence of pairs as a function of the pairwise distance, α is set to 1.61 based on an empirical analysis of hard-sphere protein-like spatial distributions,[37] and k_B and T are the Boltzmann constant and absolute temperature, respectively. The parameter, η, is an arbitrary constant that can be modified either to estimate free-energy differences[27] or to tune the strength of the DFIRE energy term versus other energy terms. The histograms N_{obs} in this work were obtained from analysis of a culled set of 1836 PDB structures from the PISCES server[38] which had better than 1.8-Å resolution and were less than 30% homologous to each other. We deviated from the original DFIRE protocol by assigning $\Delta r = 0.5$ Å at all distances and having r range from 0.25 Å to 14.75 Å, such that $r_{\mathrm{cut}} =$ 15 Å.

Like many statistical potentials, the DFIRE model is not suitable by itself for exploring the energy landscape without some sort of restraints or constraints.[12] In the case of Monte Carlo exploration, one can sample different dihedral rotamers of the backbone and side chains, where each conformation is forced to obey standard bond lengths and bond angles. In our case, where we desire to run molecular dynamics, a further issue is that the DFIRE potential needs to be smoothed out. We employed cubic interpolation[39] to smooth out the potential so that the first derivatives are continuous. An example of this procedure is illustrated in Figure 1. Our complete dynamics potential, denoted here as DFIRE-MD, consists of the standard PARAM19 internal energy and van der Waals attraction/repulsion terms and the smoothed DFIRE-AA potential with η set to 0.25. As compared to DFIRE-AA, DFIRE-MD only incorporates smoothed statistical potential energies from nonbonded list pairs which include intraresidue pairs beyond 1−4 interactions. In minor contrast, typical DFIRE-AA includes all pairs of atoms up to precisely 15 Å excluding all intraresidue pairwise interactions. Electrostatics and solvation were omitted in DFIRE-MD as they were considered analogous to the contributions of DFIRE-AA. The van der Waals term was retained so that short-range steric interactions were properly modeled. Besides the obvious issue of overcounting in this energy model, it is questionable whether a statistical potential that defines a free energy should be used as a potential for molecular dynamics. Nonetheless, we are mainly concerned in this work with the exploration of a scoring-function surface, and not thermodynamics.

We also employ the all-atom PARAM22 force field[21] combined with GBMV2[25] implicit solvent model. A linear surface-area-based hydrophobic term of 30 cal/(mol·Å^2)[17] was also included using the SASA-1 approximation.[25] As indicated in the Introduction, this combination potential, which we will refer to as PARAM22/GB-SA, was one of the best performers in a previous protein structure detection study using the CASP4 predictions as decoy sets.[17] We believe that alternative implicit solvent models might lead to a modest decrease in accuracy but being considerably more computationally efficient may outweigh this.

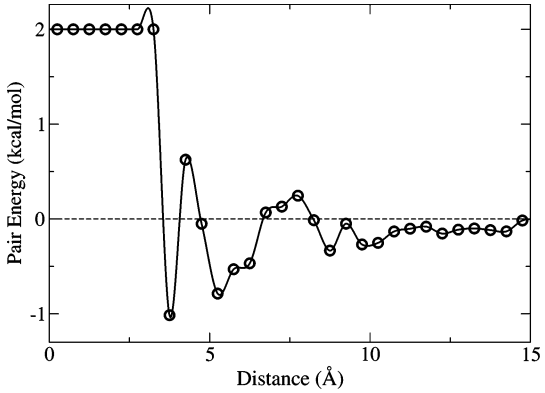


Figure 1. Regular and smoothed DFIRE-AA potential for the pairwise interaction of two alanine C_αs. The circles denote the regular DFIRE-AA potential values at the bin centers. The solid curve is the cubic-interpolated version suitable for MD simulations.

Table 1. Features of the Nine Protein Decoy Sets Used in This Work

PDB ID	N_{res}[a]	% alpha	% beta	no. of decoys in set	best rmsd rmsd	best rmsd % nc[b]	best % nc[b] rmsd	best % nc[b] % nc[b]
1ail	67	85	0	1807	2.0	55	2.0	55
1csp	64	0	53	1809	3.2	43	3.9	46
1ctf	67	52	19	1922	2.7	57	3.5	64
1pgx	57	25	46	1851	1.5	63	1.5	63
1r69	61	64	0	1733	1.4	64	1.4	69
1tif	59	17	37	1849	2.6	56	2.6	56
1utg	62	79	0	1897	3.4	36	5.4	53
1vif	48	0	50	1896	0.4	56	1.2	86
5icb	72	57	6	1870	3.0	58	3.1	59

[a] Number of residues in protein. [b] % nc − percentage of native contacts.

2.2. Protein Model Sets. The specific interest of this work is to assess detection and refinement of de novo-generated protein structure models created by the Baker lab using the Rosetta program.[4] We looked at nine proteins in this study, with the following PDB identifiers:[33] 1ail, 1csp, 1ctf, 1pgx, 1r69, 1tif, 1utg, 1vif, and 5icb (see Table 1). These proteins were chosen based on their diversity of secondary structure, availability of online Rosetta decoy sets (which we call *Rosetta2*, denoting the second generation),[6] availability of X-ray crystal native structures, and overlap with previous detection and refinement studies.[7,8,40] Each one of the decoy sets contains approximately 1800 models, which consist of ∼1000 decoys from the original Rosetta decoy set, ∼400 somewhat near to the native, and ∼400 of the lowest C_α rmsd from an exhaustive 200 000 model Rosetta run.[6] The enrichment of low rmsd structures in these sets is certainly an influence on our results and cannot be fairly compared to a prediction protocol where far less than 200 000 Rosetta models are generated. This issue is considered more in the Discussion section.

For the first statistical potential detection trial, the all-atom decoys were used as-is. For all of the other methods, the Rosetta models were converted to a PARAM19 format using the Multiscale Modeling Tools for Structural Biology (MMTSB) *convpdb.pl* command and minimized modestly to remove steric clashes (MMTSB *minCHARMM.pl* command interfaced with CHARMM[41]): 50 steps with a steepest descent algorithm followed by 100 steps with an adopted basis Newton−Raphson protocol. The energy function for minimization used a distance-based dielectric electrostatic term with a coefficient of 4.[17]

2.3. Clustering. While results may vary for other sets of Rosetta-generated models, low C_α-rmsd structures often show up in the Rosetta2 decoy sets as seen in Table 1. Also, the population of these low C_α-rmsd models may be diminishingly small.[6] In general, we observe that at the collection phase after a Rosetta run, it is imperative not to discard structures solely by score, because they could actually be the best models, i.e., nearest-native. However, some amount of filtering needs to take place before any computationally intensive refinement procedure such as replica exchange or Z-fold (both described below). In this work, we hierarchically cluster Rosetta-generated decoy structures[6] to obtain a diverse set of structures using the MMTSB command *cluster.pl* with the *-jclust* option. Our nondefault clustering parameters included a maximum of four subclusters per parent cluster (*-maxnum* option) and minimum of four elements per subcluster (*-minsize* option.) The clusters were selected from the fourth hierarchical level, such that in each decoy set, at least 16 clusters could be identified in all of the protein sets. The average DFIRE-AA scores from each cluster were ranked, and the lowest-energy conformers from each of the top 16 clusters were defined as the diversity set. Note that the PARAM22/GB-SA scores could have been used instead for ranking.

2.4. Replica Exchange. The replica exchange method (ReX)[42] is a state-of-the-art technique for sampling an energy landscape. It has been used successfully in studies of protein folding,[43] loop structure prediction,[44] and lattice-based protein structure prediction.[45] The concept behind the method is to run multiple simultaneous molecular dynamics or Monte Carlo simulations with a spectrum of biases and/or temperatures. The principle of using ReX in this study is to allow for automatic unfolding of worse scoring structures and refolding of better scoring structures. In this work, a range of temperature windows is used, and we looked at the performance of separately the PARAM22/GB-SA and DFIRE-MD potential. After a specified block simulation time, τ, windows a and b exchange temperatures with a probability, W:[46]

$$W(a \leftrightarrow b) = \begin{cases} 1 & \Delta_{ab} \le 0 \\ \exp(-\Delta_{ab}) & \Delta_{ab} > 0 \end{cases}$$

$$\Delta_{ab} = (\beta_a - \beta_b)(E_a - E_b) \qquad (2)$$

where β is $1/k_B T$ and E is the potential energy of a particular replica. We used 16 temperature windows ranging exponentially from 298 to 650 K for the DFIRE-MD simulations and 298 to 500 K for the PARAM22/GB-SA runs. The different temperature ranges selected for each potential reflected the fact that we tried to ramp up the temperature for the DFIRE-MD simulations as high as possible to counter the strong collapsing propensity of this potential, while retaining some energy overlap between windows. The initial structures placed in each window corresponded to the 16-member diversity set described above. Block simulation times, τ, were set to 0.4 ps. A total of 2500 exchange steps were carried out, for a cumulative simulation time of 1 ns. Molecular dynamics simulations were performed with the CHARMM software package,[41] and the replica exchange method was performed with the MMTSB *aarex.pl* program.[47]

Even though ReX enhances sampling, some accuracy will be lost simply by having to filter out a small number of structures to create a diversity set. Therefore, we decided to also run every minimized decoy with 298 K molecular dynamics for a small amount of simulation time, 1 ps, to compare with the ReX simulations. With such short runs, the relevant question was whether a small amount of refinement could improve detection.

2.5. Z-Fold Method. Noting the strong attractive nature of a pairwise statistical potential during a MD run, we decided to utilize this feature to refold protein structures with the aim of generating a diversity of conformations in the vicinity of a given model structure. The *Z-fold* method starts by temperature unfolding (400 K) a protein model over a short time with secondary structure restraints and only the vdW and internal energy terms turned on. This is followed by refolding with the DFIRE-MD potential retaining the secondary structure restraints. In this work, the unfolding simulations were performed for 10 ps, and refolding simulations were performed for 6 ps. For each starting model, 10 unfolding simulations with different random seeds were performed. For each unfolded structure, there were then 10 refolds performed, for a total of 100 refolded structures per starting model. The secondary structure restraints were obtained via the DSSP[48] program evaluated on the original model. Secondary structure restraints, E_{ss}, of the form

$$E_{ss} = K \times \max\left[0, \mathrm{abs}\left(\theta - \frac{\pi\theta_{\min}}{180^\circ}\right) - w\right]^2 \qquad (3)$$

were used to restrict the backbone dihedral angles of the identified secondary structure elements to plus or minus the width, w, from $\theta_{\min}$. The force constant, K, was set to 100 kcal/mol/rad,[2] w is the width of the potential, and θ corresponds to either the ϕ- or φ-dihedral angles. For the α helix restraints, the parameters were $w = 7^\circ$, $\phi_{\min} = -64^\circ$, and $\varphi_{\min} = -41^\circ$. For the beta-strand restraints, the parameters were $w = 40^\circ$, $\phi_{\min} = -120^\circ$, and $\varphi_{\min} = +120^\circ$. The 16-member diversity sets for each protein were also the starting models in this part of the study. After generation, each refolded structure was minimized and rescored using the PARAM22/GB-SA detection protocol described above.

2.6. Analysis Techniques. A popular measure of the similarity of a model structure with the native conformation is rmsd. In this work, rmsd is defined for the C_α protein backbone versus the native X-ray structure in units of Å. A common evaluation of scoring functions is the Z-score, which normalizes the score of the native, E_{native}, relative to the mean,

$\bar{E}$, and standard deviation, σ, of the scores of the decoy set:

$$Z_{\text{ener}} = \frac{E_{\text{native}} - \bar{E}}{\sigma} \quad (4)$$

The Z-score is a useful measure of the depth of the scoring funnel, whereby greater negative values indicate deeper funnels. Nonetheless, detecting a near-native structure from a set of models can only be reliably achieved when there is some propensity of the scoring function to favor structures as they become more and more nativelike. Therefore, we are concerned as well in this work with other criteria: the rmsd of the lowest scoring structure (excluding the native); the best rmsd of the top five scoring structures; the 15 × 15% enrichment score;[6] and statistical correlation between rmsd and score. The 15 × 15% enrichment score measures the number of structures which are both in the top 15% of scores and top 15% of RMSDs to native divided by the number of structures one would expect by chance to satisfy these two criteria. Summa et al.[36] show that other measures of the usefulness of a scoring potential are correlated significantly with the ones we use here.

While C_α rmsd is a popular measure of similarity of a conformation to the native structure, it is sometimes helpful to look at other similarity measures, such as fraction of native contacts. The definition for fraction of native contacts is as follows. First, for a given native structure, the native contacts are identified as all side-chain center-of-mass pairs, (i,j): $j > i + 1$, whose distances are less than 6.5 Å.[49] Then for each model conformation, the fraction of native contacts is the number of native contacts in the model divided by the total number of native contacts in the X-ray structure. In this work, to conform to the directionality of rmsd scatter plots, we take one minus the result.

3. Results

The following section considers separately detection and refinement using two distinct scoring functions: DFIRE-AA and PARAM22/GB-SA. In the detection subsection, we consider the ability of these two scoring functions to find near-native structures from large decoy sets of de novo-generated conformations. In the refinement subsection, we first ask whether short-time molecular dynamics enhances the detection capabilities of the force field-based score. Then we test the two scoring functions in a standard replica exchange protocol to see whether small subsets of the decoy sets can be induced toward the native state. Finally, noting the collapsing propensities of the DFIRE-AA as a sampling function, we evaluate the above-described unfold/refold method with the same small subsets of decoys.

3.1. Detection. Table 1 outlines some of the features of the decoy sets we chose. The best structures in each set have C_α RMSDs of ~3 Å and below. In contrast, the structures with the best percentage of native contacts have only between 50% and 65% similarity. This means that 35−50% of the native contacts are missing even in the best decoy structures. Therefore, it might be conjectured that scoring functions with atomic resolution may fail to detect the structures that are closer to native, because they are still some distance away in contact space. Finally, in only three of the nine protein sets is the best rmsd structure also the closest to the native in contact space.

Table 2. Summary of Results for Detection of Structures Using the DFIRE-AA Potential Score on the Original Decoy Structures

PDB ID	Z_{ener}	rmsd of top scoring structure	best rmsd of top 5 scoring structures	enrichment (15 × 15%)	av rmsd of top cluster	best av rmsd of top 5 clusters
1ail[a]	−2.3	8.7	4.5	0.69	9.2	7.1
1csp[a]	−3.2	4.3	4.3	2.82	6.0	6.0
1ctf	−3.5	3.3	3.3	1.85	4.8	4.8
1pgx	−4.4	5.9	2.4	2.45	5.5	4.1
1r69	−3.6	2.2	1.5	3.49	3.5	3.5
1tif	−5.1	7.8	3.9	0.96	5.1	5.0
1utg[a]	−1.3	10.7	6.7	0.54	6.2	6.2
1vif	−2.8	0.6	0.6	4.80	3.8	3.8
5icb[a]	−2.2	4.3	4.3	2.19	5.7	5.7
avg[b]	−3.2	5.3	3.5	2.20 (1.39)	5.5	5.1

[a] Native structure was not detected as the lowest in energy. [b] Standard deviation in parentheses.

Table 3. Summary of Results for Detection of Structures Using the DFIRE-AA Potential Score on the Minimized Decoy Structures[a]

PDB ID	Z_{ener}	rmsd of top scorer	best rmsd of top 5 scoring structures	enrichment (15 × 15%)	av rmsd of top cluster	best av rmsd of top 5 clusters
1ail[b]	−1.4	9.7	7.7	0.64	9.2	6.6
1csp[b]	−3.1	4.3	3.9	2.78	6.0	6.0
1ctf	−3.3	3.3	3.3	1.87	4.8	4.8
1pgx	−3.6	5.9	2.4	2.52	4.2	4.1
1r69	−3.5	2.2	1.5	3.80	3.5	3.5
1tif	−3.8	7.8	3.8	1.08	5.1	5.0
1utg[b]	−0.5	5.4	5.4	0.30	6.2	6.2
1vif	−2.2	0.6	0.6	4.71	3.8	3.8
5icb[b]	−1.8	4.4	4.2	2.04	5.7	5.6
avg[c]	−2.6	4.8	3.6	2.19 (1.45)	5.4	5.1

[a] Structures were minimized using the protocol specified in the Methods section. [b] Native structure was not detected as the lowest in energy. [c] Standard deviation in parentheses.

The PARAM22/GB-SA potential is marginally better than DFIRE-AA at detecting a low-rmsd structure using score alone, as seen in Tables 2−4. Both potentials perform roughly the same in detection if the top five scoring conformations are considered. One can also see that the average Z-score for the PARAM22/GB-SA is slightly superior to the DFIRE-AA one. Furthermore, DFIRE-AA fails to detect the native X-ray crystal structure for four proteins (even with minimization), while PARAM22/GB-SA fails for only two proteins. The 15 × 15% enrichment scores for both potentials are on average roughly the same, while the standard deviation of these scores suggest DFIRE-AA can be either better or worse than PARAM22/GB-SA for a specific protein. For example, the DFIRE-AA potential fares worse than chance (i.e., enrichment scores less than 1) for three proteins, while the PARAM22/GB-SA enrichment values are above chance in each protein case. Using a clustering scheme to choose structures or sets of structures is somewhat worse than single conformation detection for DFIRE-AA (Tables 2 and 3) and significantly worse for PARAM22/GB-SA (Table 4). In principle, clustering should

Table 4. Summary of Results for the Detection of Structures Using the PARAM22/GB-SA Potential on the Minimized Decoy Structures[a]

PDB ID	Z_{ener}	rmsd of top scoring structure	best rmsd of top 5 scoring structures	enrichment (15 × 15%)	av rmsd of top cluster	best av rmsd of top 5 clusters
1ail	−3.3	10.7	4.0	2.58	6.6	6.6
1csp	−4.2	4.5	4.3	1.82	6.0	6.0
1ctf	−4.9	3.7	3.3	2.22	4.8	4.8
1pgx	−5.5	2.4	2.4	1.32	5.5	4.5
1r69	−5.8	2.4	1.7	2.59	3.5	3.5
1tif	−5.1	4.4	4.0	1.35	6.0	5.0
1utg[b]	−2.1	4.7	4.6	1.55	6.2	6.2
1vif	−3.2	0.5	0.4	4.22	3.8	3.8
5icb[b]	−1.8	4.1	4.0	1.66	8.4	5.6
avg[c]	−4.0	4.2	3.2	2.15 (0.92)	5.6	5.1

[a] Structures were optimized using the protocol specified in the Methods section. [b] Native structure was not detected as the lowest in energy. [c] Standard deviation in parentheses.

help smooth out noise in the scoring function. In practice, lingering clashes in specific structures are more penalized in the PARAM22/GB-SA results, likely leading to worse overall average cluster energies. Furthermore, cluster populations at this stage are unlikely to be fruitful, given that they are dependent on the "thermodynamic" sampling of the lower-resolution Rosetta united-residue function.

It is interesting to compare the results of the original paper associated with the decoys we used.[6] Tsai et al. reports an average Z-score of −4.5 and enrichment value of 1.86 for all 78 proteins using their single unified α/β scoring potential. This is not a fair comparison between our results and theirs as we are using a small manually selected subset of proteins from their large collection. However, it shows that our force-field results are in line with their analyses, which had used a different atomic resolution potential.

In Figure 2, we show two examples of the detection problem using the Rosetta decoy sets: an easy case and a difficult case. In the easy situation, 1vif (Figure 2a,b), there are several very near-native structures generated. Also, as the structures approach the native, there is a downward slope in energy. Structures below 1 Å in rmsd are detectable versus the rest of the set using the PARAM22/GB-SA potential. Furthermore, the lowest energy structure for this potential is nearly the lowest rmsd.[8] In contrast, in the difficult case, the 1ail (Figure 2c,d) decoy set has few structures that are better than 4 Å and only one structure better than 2 Å. Visually, one might consider the group of structures in Figure 2c at ∼4.5 Å have on average a better score than the other group, which suggests that by clustering a ∼4 Å conformation could be selected out. In reality, though, no such lower-scoring cluster was identified (Table 4). Figure 2c also illustrates how single structure detection can fail sometimes, as it picks out the low-scoring conformation at 10.7 Å. Figure 2d shows that DFIRE-AA cannot discern the native structure as lowest in energy. In addition, there are no visible trends in this scatter plot.

Figure 3 illustrates the point that even if many 2 and 3 Å structures are in the decoy set, they may have a lot of missing native contacts. This provides some evidence of why atomic resolution scoring functions may not detect these lower rmsd structures. Figure 3b shows very little funnel-like behavior, likely due to the large gap in native contact space between the best decoys and the native structure. In Figure 3c, the

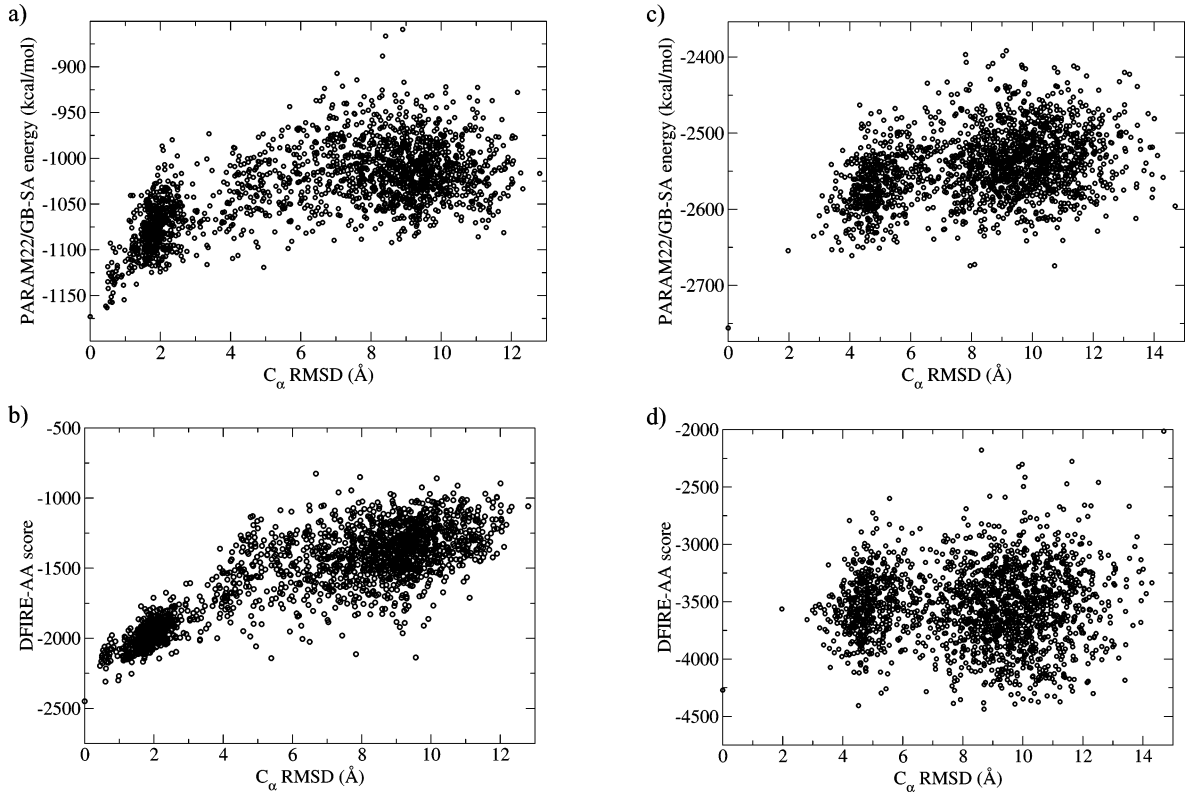


Figure 2. Scatter plots of the PARAM22/GB-SA and DFIRE-AA potentials vs C_α rmsd to native: (**a-b**) 1vif, an easy test case for detection and (**c-d**) 1ail, a difficult test case for detection.

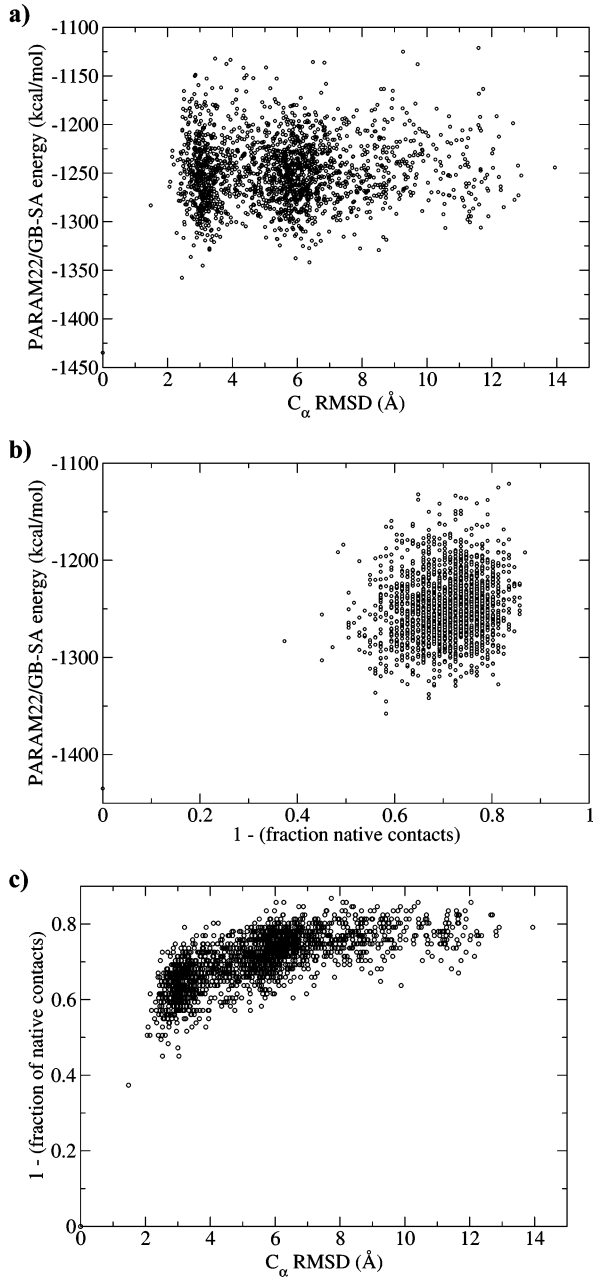


Figure 3. For the 1pgx decoy set, comparison of (**a**) PARAM22/GB-SA score with C_α rmsd, (**b**) PARAM22/GB-SA score with fraction of native contacts, and (**c**) fraction of native contacts with C_α rmsd.

structures between 2 and 4 Å begin to have some slope toward more native contacts than the continuum of structures in the set.

In Table 5, the funnel-like behavior of the two scoring functions is further quantified by looking at the correlation coefficient of the score to the rmsd of decoys which are close to the native.[7,50] In most proteins, small correlations do exist between score and rmsd. However, in some notable cases, such as 1tif and 1utg for DFIRE-AA and 1pgx for PARAM22/GB-SA, the correlations are nearly zero or negative, indicating no funnel-like behavior. Since the Rosetta-generated decoys do not completely span the conformation space of our test proteins, the correlations are probably, in general, underestimated. In fact, protein decoy sets obtained by perturbation of the native structure tend to show improved correlation between score and rmsd at various rmsd ranges (results not shown).[51]

Table 5. Correlation Coefficient of DFIRE-AA and PARAM22/GB-SA Scores vs RMSD as a Function of Different RMSD Ranges of Conformations

PDB ID	DFIRE-AA			PARAM22/GB-SA		
	<4 Å	<6 Å	all	<4 Å	<6 Å	all
1ail	−0.05	0.20	0.03	0.32	0.34	0.25
1csp	0.07	0.22	0.48	0.06	0.12	0.11
1ctf	0.06	0.22	0.43	0.17	0.30	0.23
1pgx	0.44	0.39	0.41	−0.03	−0.01	0.04
1r69	0.48	0.51	0.51	0.32	0.38	0.24
1tif	−0.09	0.08	0.39	−0.16	0.13	0.06
1utg	−0.09	−0.22	0.11	0.27	0.17	0.18
1vif	0.74	0.87	0.85	0.51	0.67	0.65
5icb	0.22	0.27	0.44	0.07	0.16	0.16
avg	0.20	0.28	0.41	0.17	0.25	0.21

Table 6. Summary of Results for Detection/Refinement of Structures Using Short-Time Molecular Dynamics (1 ps) with the PARAM22/GB-SA Potential

PDB ID	rmsd[a] of top scorer[b]	best rmsd[a] of top 5 scorers[b]	enrichment (15 × 15%)
1ail	10.8	4.0	2.53
1csp	7.7	4.5	1.87
1ctf	4.0	3.6	1.94
1pgx	11.8	2.5	2.21
1r69	1.7	1.7	3.92
1tif	4.0	3.4	1.44
1utg	11.0	4.5	1.41
1vif	0.9	0.9	4.29
5icb	4.1	4.2	1.90
avg[c]	6.2	3.3	2.39 (1.04)

[a] rmsd defined with respect to the final structures of the dynamics trajectories [b] Score defined as the average potential energy over the short-time dynamics simulation. [c] Standard deviation in parentheses.

3.2. Refinement. Short-run MD results on every decoy structure are presented in Table 6. The goal here was to obtain quick refinement of all of the decoy structures in the hopes that poor side-chain contacts in good rmsd structures might be rectified and detection would be improved. Unfortunately, single structure detection results were 2 Å worse on average than minimization alone. Side-by-side comparisons with the optimized structure results show that dynamics increased detection errors for a few of the difficult cases and 1pgx. Using the top five scoring conformations criteria, the dynamics results are on par with simple optimization. Finally, enrichment scores are overall enhanced somewhat by short-time dynamics. While these simulations lack equilibration at 298 K, which could be a source of error, there is a practical compromise with simulation runtime when thousands of structures must be simulated.[8]

Tables 7 and 8 summarize the results of ReX simulations on a diversity set of conformations ($N = 16$) for each protein. Each small set includes at least one structure of ∼3 Å rmsd quality. The sampling nature of ReX−MD simulations permits us to look at clusters and their respective populations

Table 7. Summary of Results for Detection and Refinement of Structures from 1-ns Replica Exchange Molecular Dynamics Simulations Using the DFIRE-MD Potential

PDB ID	best rmsd in diversity set	lowest rmsd (298 K)	lowest rmsd (all T)	lowest av energy cluster (rmsd)[a,b]	most populated cluster (rmsd)[a,b]
1ail	5.3	5.3	4.9	10.9	10.9
1csp	3.6	3.5	3.4	8.0	5.7
1ctf	3.6	3.4	3.4	10.4	5.2
1pgx	1.5	1.9	1.0	8.8	8.8
1r69	1.5	1.1	1.1	3.5	1.5
1tif	4.1	4.0	3.5	4.9	4.9
1utg	4.8	5.7	4.8	8.4	10.4
1vif	0.6	0.6	0.6	9.4	3.7[c]
5icb	4.3	4.1	3.3	4.8	4.4
avg	3.3	3.3	2.9	7.7	6.2

[a] Structures and energies obtained from the final step of the simulation block in the 298 K window. [b] rmsd was averaged over the structures in the specified cluster. [c] Averaged over two clusters tied for first, with RMSDs of 2.8 Å and 4.6 Å.

Table 8. Summary of Results for Detection and Refinement of Structures via 1-ns Replica Exchange Molecular Dynamics Simulations with the PARAM22/GB-SA Potential

PDB ID	best rmsd in diversity set	lowest rmsd (298 K)	lowest rmsd (all T)	lowest av energy cluster (rmsd)[a,b,c]	most populated cluster (rmsd)[a,b]
1ail	5.3	5.2	4.8	6.8	11.8
1csp	3.6	3.4	3.4	7.0	4.2
1ctf	3.6	3.1	3.1	11.3	10.9
1pgx	1.5	1.7	1.6	6.5	7.1
1r69	1.5	1.3	1.3	3.0	3.0
1tif	4.1	3.9	3.6	6.1	6.1
1utg	4.8	4.4	4.1	5.7	10.0
1vif	0.6	0.7	0.7	2.0	1.8
5icb	4.3	4.2	3.4	5.0	5.0
avg	3.3	3.1	2.9	5.9	6.7

[a] Structures and energies were obtained from the final step of each simulation block in the 298 K window. [b] rmsd was averaged over the structures in the specified cluster. [c] Clusters with less than 10 elements were filtered out.

at 298 K as analogous to free energies of these clusters. The data indicate the cluster population offered better detection than average energy for the DFIRE-MD potential but not the PARAM22/GB-SA potential. Given the limited number of proteins in this work, neither average energy nor free energy can be distinguished as better than the other. In only a few of the protein cases for both potentials did the lowest rmsd starting conformation contribute significantly to the lowest-energy cluster. This highlights the limitations of the potentials and the fact that no significant folding funnel could be discerned at such limited rmsd quality. Proteins 1ail and 1utg exemplify the latter constraint.

Interestingly, there are slight rmsd improvements, albeit undetectable via energy criteria, which take place for both potentials for most of the proteins.[7,12] At 298 K, the improvements average 0.2 Å for PARAM22/GB-SA. Over all the temperature windows, improvements average as much as 0.4 Å. In the case of DFIRE-MD, these rmsd improvements may not reflect refinement as much as compacting of the model structures. It is also noted that the best structures were not produced and preserved in the lowest temperature (298 K) window.

Some further details of a single replica exchange simulation (1pgx/PARAM22-GBSA) are presented in Figure 4. The progressions of the two lowest rmsd models, as seen in Figure 4a, are quite different. The 2.4 Å structure stabilizes and becomes lower in rmsd to about 2.0 Å, while the 1.4 Å structure gets significantly worse over time. This divergence can be explained in Figure 4b,c, where the 2.4 Å structure spent much more time in cooler temperature windows than the 1.4 Å model. Finally, as illustrated in Figure 4d, we note that in the first 600+ ps, a 7 Å model dominates the lowest temperature window. Consistent with this result, the data in Table 8 indicate that the lowest free-energy cluster had an average rmsd of ~7 Å. One might surmise that with further sampling the 2.5 Å model would dominate the 298 K window and be detected as the lowest in free energy.

Two more important points can be gleaned from Figure 4. First, the energy of a structure and not its rmsd to the native dictates how the models will percolate through the temperature windows. Hence, a poorly scoring low-rmsd structure inserted into a simulation may end up getting muddled by high temperatures. Second, the ReX simulations, as computationally intensive as they are, generally must be run for much longer simulation times than were done here (e.g., 10−100 ns) to get convergent population statistics.

Overall, the ReX results are not as remarkable as the simple detection schemes, despite the orders of magnitude more computational effort. Our PARAM22/GB-SA replica exchange on proteins of the size studied here required 2 days of computation per protein on 16 AMD Athlon 2200+ processors. In contrast, the PARAM22/GB-SA detection protocol on an entire decoy set required about 5 h on a single CPU.

Figure 5 illustrates that the DFIRE scores can be highly correlated with the compactness of the conformations. Although this trait may not be able to completely explain DFIRE-AA detection abilities, it does suggest that running on the DFIRE-MD energy surface could cause structures to become more compact. In fact, Figure 6 illustrates that DFIRE-MD tends to compress protein structures and make them more spherical in shape. This can be attributed to the fact that DFIRE-potential tends to maximize intraprotein contacts. An opposing protein contact breaker, such as a solvation term, is lacking.[36] Despite the distortions caused by DFIRE-MD, the potential is very expedient at forming contacts. In Figure 7, one can see that a partially extended conformation is quickly collapsed into a compact structure in a mere 5 ps of simulation time.

Given the quick collapsing propensities of DFIRE-MD, the Z-fold method was tested, and the results are summarized in Table 9. As one can see, the results are not much better than replica exchange on average. The lowest average energy and lowest free-energy clusters are on par with clustering results in the detection and replica exchange calculations. Most noticeably, the best rmsd structure is on average 0.3

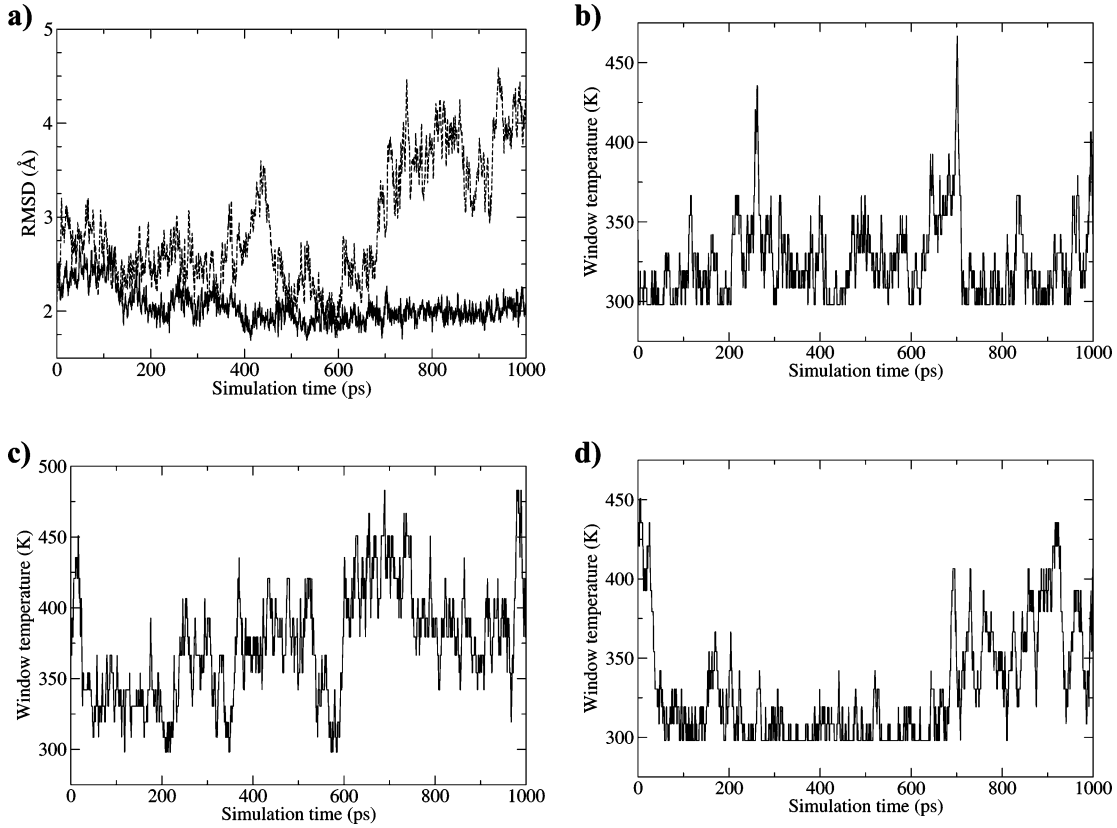


Figure 4. Replica exchange results for the 1pgx diversity set using the PARAM22/GB-SA potential. (**a**) Comparison of 2.4 Å (solid line) and 1.4 Å (dashed line) models. Temperature progressions of the (**b**) 2.4 Å, (**c**) 1.4 Å, and (**d**) 7 Å models.

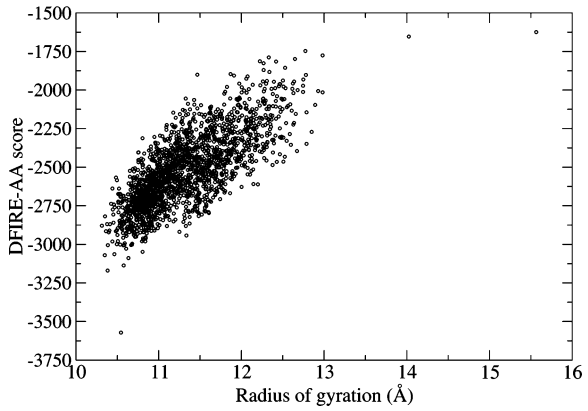


Figure 5. Comparison of the DFIRE-AA score with radius of gyration for the 1pgx decoy set.

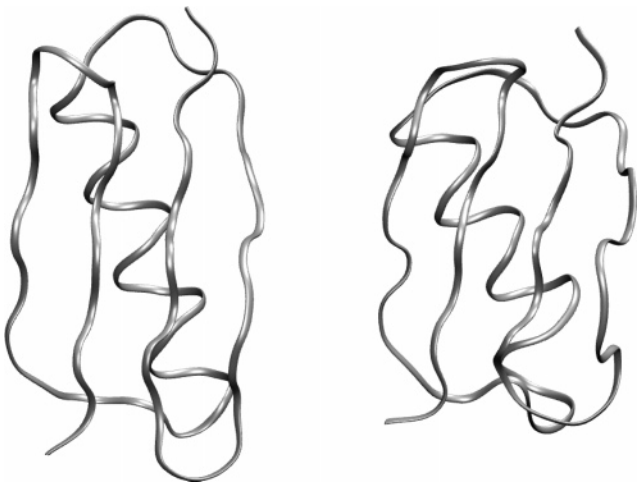


Figure 6. DFIRE-MD compresses native 1pgx protein structure over a simulation time period of 1 ns: (left) native structure and (right) after 1 ns of DFIRE-MD. Molecular graphics rendered with VMD software.[67]

Å better than the original best structure. Nevertheless, some of the rmsd improvement could be due to the compacting nature of the DFIRE-MD potential. Another issue is that neither the DFIRE-MD nor the PARAM22/GB-SA potential was able to detect the best rmsd structures. In Figure 8, it appears that improvements in rmsd were achieved for structures 2 Å and farther from the native. Sometimes rmsd improvements could be detected by DFIRE-MD as illustrated by the filled squares which lie above and below the zero line. Once again, some structure compacting may be occurring, and small rmsd improvements may not translate completely as refinements.

4. Discussion

4.1. Decoy Set Properties. The ability to detect near-native structures from a set of conformations is inevitably related to the quality of structures in the set. If enough low-rmsd structures are available, any good detection function should be able to pick up at least some of these structures as better in score than the rest. The extreme case of an easy decoy set would be one where model structures are developed from perturbations of the native. Small perturbation decoys would

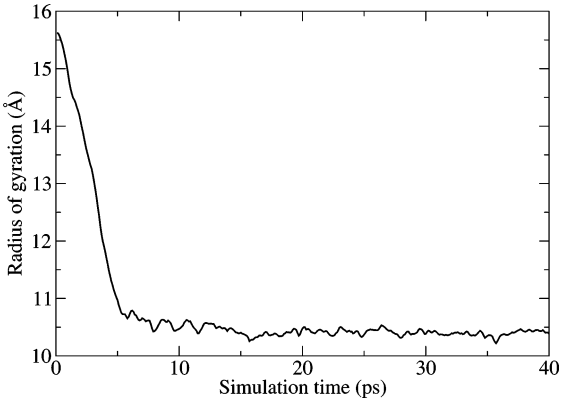


Figure 7. Radius gyration as a function of simulation time for the most extended model structure in the 1pgx decoy set using the DFIRE-MD potential at a temperature of 298 K.

be very near-native, and if the detection function labels the native as best, it would likely label near-natives better than misfolds.

Most of the low-rmsd structures in the Rosetta2 decoy sets are culled from Rosetta runs of nearly 200 000 structures per protein set. Generation of 200 000 structures for a single target is roughly an order magnitude larger than the standard automated server Robetta protocol. With today's computing, a single protein prediction would require effort on the order of CPU-weeks[52] to generate 200 000 models. Enrichment of the decoy sets with low-rmsd structures seems to increase the probability of detecting near-native structures, because the likelihood that at least one near-native structure will outscore all of the other structures increases. In addition, if clustering is performed, the near-native enrichment may provide a distinct cluster of structures from which to select. Furthermore, analyses such as the colony energy method,[14] which modifies the scores based on the presence of structural neighbors, would be biased by the enrichment protocol since as structures become closer to the native, they also become closer to each other, hence enhancing the pairwise rmsd weighting factors.

Bradley et al. shows that low-rmsd structures can often be found by using sequences homologous to the target in the Rosetta algorithm[8] where only a total of 10−20 thousand structures need to be built. Note that for the three proteins in common between their test set and ours (1tif, (1r69, and (1csp), their "Round 2" C_α rmsd results (4.1, 1.2, and 4.7 Å) are quite comparable to our simple PARAM22/GB-SA detection of their enriched decoy set (4.4, 2.4, and 4.5 Å). This favorable comparison is likely due to the fact that generating a large decoy set of 200 000 models increases the probability that there will be enough lower rmsd structures from which to detect. Moreover, it is unclear from the work of Bradley et al., whether their improvements were gained by increasing diversity or by using a computationally intensive refinement procedure (100−150 CPU-days per protein).

Despite the presence of near-native structures in every set in this work based on rmsd, other indicators such as fraction of native contacts suggest that the so-called near-natives are not near enough. With only an average of 60% native contacts for the best structures, it makes sense that the atomic resolution scoring functions may have some difficulty in detection. There are two reasons why the fraction of native contacts may be lacking. First, the side-chain prediction algorithm used for these decoy sets may not be optimal. Tests of rebuilding side chains with the SCAP method[53] led to better overall DFIRE-AA scores (results not shown). The other issue is that the fraction of native contacts may have, in analogy to a scoring function, a narrow funnel versus backbone rmsd. Most of the native contacts will collapse into place only when the protein is very close to the native in backbone rmsd space (see, for example, Figure 3c).

4.2. Scoring/Energy Functions. In this work, we looked at two diverse scoring/energy functions: one force field-based and the other statistically based. Force field-based functions are considered to be accurate but have many drawbacks. First, the standard van der Waals repulsion term is very sensitive to the positions of neighboring atoms such that structural minimization is required. Tsai et al. suggest the use of finite core repulsion terms to alleviate this issue.[6] A compromise, however, must be made to ensure that the core is repulsive enough to filter out incorrectly packed structures. Another problem with force field-based functions is that the folding funnel is trying to mimic the physical energy landscape of real proteins. As such, real proteins may have a subtle free energy gradient toward the native that requires long folding times (e.g., milliseconds to several seconds). Compared to the standard simulation times possibly using current computer resources, typically in the single-digit nanosecond range, there is a gap of several orders of magnitude.[13] A final problem with force field-based potentials is that they may be too inaccurate. Consequently, after exceptional computational effort of using them, simulations may still lead to unphysical structures.

One of the main problems with a pairwise-only statistical potential such as DFIRE-AA is the lack of a microenvironment or solvation term.[36] Many scoring functions already employ such additional terms.[6,20,36] This is needed because pairwise contacts are not statistically independent in known protein structures.[36] We believe such additions might increase the number of near-natives detected for some protein sets. The DFIRE-AA potential, like other statistical potentials, gleans information from native PDB structures. Consequently, unfolded state information is noticeably absent. This presumably leads to the large energy gradient in DFIRE-AA seen in the protein collapsing simulations (Figures 6 and 7). Atomic force fields, on the other hand, contain a relatively balanced description of unfolded and folded states. Thus, the energetic differences and subsequent propensities to drive folding are much more subtle and should be on the order of 5−15 kcal/mol, at least in terms of free energy.[54] Skolnick et al.[20] suggest parametrizing an energy function based on decoys/misfolds and near-natives. In this way, there is an enforced funnel or directionality between the two extremes which can be tuned to obtain a desired folding gradient.

The general issue regarding scoring functions is to what extent can they be optimized to achieve a significant folding funnel? Furthermore, the two key aspects of the funnel are its depth and width. Maximization of the Z-score by Tsai et

Table 9. Summary of Results for Detection and Refinement of Structures Using the *Z*-Fold Method

PDB ID	best rmsd diversity set	best rmsd	low av energy cluster	low free energy cluster	rmsd of lowest energy	best rmsd of top 5	top rescored[a]	best rmsd of top 5 rescored[a]
1ail	5.3	5.0	11.9	8.6	7.6	7.6	7.7	7.5
1csp	3.6	3.3	4.2	4.2	3.4	3.4	4.5	4.1
1ctf	3.6	2.9	8.3	4.4	3.8	3.5	3.7	3.0
1pgx	1.5	1.2	2.1	5.7	10.5	2.0	2.2	1.2
1r69	1.5	1.0	1.5	5.7	1.1	1.1	1.1	1.1
1tif	4.1	3.9	5.2	6.9	5.1	5.0	5.2	5.2
1utg	4.8	4.2	10.9	5.1	10.9	5.0	10.5	9.3
1vif	0.6	1.6	2.5	5.1	3.1	1.9	2.9	1.8
5icb	4.3	3.8	2.9	3.9	9.0	4.3	8.4	2.8
avg	3.3	3.0	5.5	5.5	6.1	3.8	5.1	4.0

[a] Rescoring potential is PARAM22/GB-SA after standard structure optimization (see text).

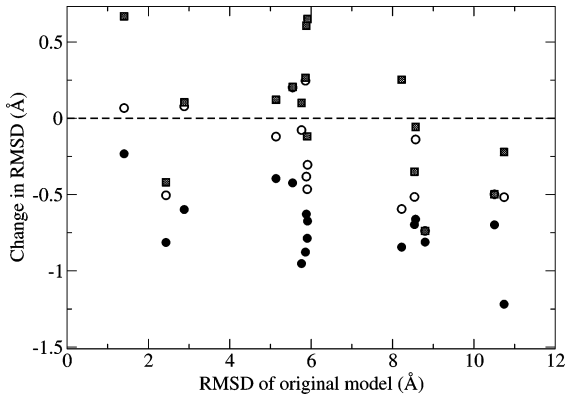


Figure 8. Refinement capability of Z-fold method as a function of rmsd of the original model for the 1pgx diversity set. Closed circles represent the lowest rmsd structures in the set, open circles denote the lowest rmsd structures out of the top five scoring conformations, and shaded squares represent the top scoring conformation.

al.[6] is an example of maximizing the depth of the scoring funnel such that the native is significantly lower in energy than any decoys. On the other hand, increasing the width of the funnel is also very important, since detection algorithms will work only if one or more model structures are within the funnel. It appears from our Z-score results, that the DFIRE-AA potential has a modestly smaller funnel depth than PARAM22/GB-SA. In addition, the overall increased enrichment scores suggest DFIRE-AA has a slightly larger funnel width. The problem with DFIRE-AA is that in some of the test sets, enrichment scores were 1 or less, suggesting that the funnel was nonexistent in the vicinity of the 15% lowest rmsd structures. The compromise to creating a wide and deep folding funnel is that, in general, the native structures of most, if not all, aqueous proteins will need to lie at or near the scoring function minimum.

4.3. Conformational Sampling. Many researchers have found that all-atom MD simulations are unable to explore diverse conformations at room-temperature despite simulation times on the order of several nanoseconds.[40] Fan and Mark[13] have suggested that even longer MD simulations on the order of hundreds of nanoseconds or even microseconds may be a viable technique for refinement. We agree that sufficiently long simulations probably would succeed some of the time. Invariably, though, simulation times of microseconds or longer are still outside most researchers' current computing capabilities.[55] Another difficulty is that the model structure to be refined may be, for all practical purposes, permanently trapped in a misfolded conformation. Misfolded proteins in vivo often require either intervention from chaperones or disposal by the cell machinery.[56,57]

In this work, we examined the ReX method which has been used successfully by Zhang et al. to sample conformational space with a sophisticated united-residue force field.[3] In fact, Misura et al.[7] commented that the addition of temperature might enhance sampling. Regrettably, the combination of an all-atom force field and ReX may not be useful without restraints, because high-temperature unfolding leads to destruction of the informational content of the original model. Furthermore, low-temperature refolding of a partially denatured structure can take an inordinately long simulation time when force-field potentials are used. In contrast, ReX simulations can be successful in loop modeling,[44] because the number of degrees of freedom are small enough to be sampled well within a feasible simulation time. In addition, the restraints of the two loop stems limit the extent of possible unfolded conformations.

Perhaps other sampling schemes such as Monte Carlo might fare better. Misura et al. performed multiple zero temperature Monte Carlo runs on small sets of decoys. They employed backbone and side-chain rotamer move sets which were able to find lower rmsd structures than the original models. One drawback was their inability to sometimes detect the lowest rmsd structures via an energy function alone. Furthermore, there was a compromise between the size of the move sets and the ability to sample rare side-chain conformations that might be crucial to achieve correct packing.[7]

The Z-fold method which entails a slight unfolding and refolding of a model conformation is a compelling alternative. It stands in contrast to simply simulating the rearrangement of a protein that is trapped in a misfolded compact state. Also, the Z-fold approach benefits from a statistical potential with a fast refolding process because the energy gradient from the partially unfolded state to a compact state is large. Nonetheless, there are several problems with using a statistical potential. First, compacting will occur at local levels causing distortion in secondary structures. This can be ameliorated somewhat through the use of secondary

structure restraints. In addition, the folds produced will be limited by the accuracy of the statistical potential. The most visible effect of this occurrence is that proteins will tend to form compact spherical structures as the competition with solvent interactions is neglected (Figure 6). Furthermore, lacking hydrogen atoms, detailed steric volume exclusions and explicit hydrogen bonding are neglected. Perhaps, a careful reweighing of energy terms and the introduction of solvation-like terms may offer the best of both worlds—a relatively fast compacting potential with diminished unphysical artifacts.

4.4. Future Directions. The fact that decoy sets with more near-native rmsd structures fared better in the detection results suggests that one should use lower-resolution models to their fullest extent before constructing and scoring all-atom models. Furthermore, all-atom model potentials are replete with local minima which hindered our dynamics-based optimization approaches. A good example of pushing the limits of united residue models is the work of Zhang et al.[45] which describes a new generation of lattice-based united residue models which can refine homology models to some degree. Furthermore, Misura et al. have shown that searches within the united-residue-based Rosetta protocol are capable of building homology models better than those created by simply constructing from a template.[58]

Given that the rmsd/score correlation values in Table 5 were suboptimal in critical rmsd ranges for most proteins, another improvement we suggest is optimizing scoring functions such as DFIRE-AA and PARAM22/GB-SA for the protein structure detection and refinement problem. For instance, the scoring funnel can be both deepened[6] and optimized to expedite folding.[59] In addition, the energy function can be smoothed[60] or transformed to enhance sampling.[61,62] Finally, hybrid strategies for conformational sampling that combine both knowledge-based and physical-based energy functions may prove to be particularly effective in refinement.[12,26]

Finally, all-atom molecular dynamics and standard replica-exchange protocols may not be the optimal methods for refinement as seen in our results. Large scale conformational changes induced by molecular dynamics are likely to be slow compared to large-scale moves possible in a Monte Carlo approach. Alternatively, enumerative sampling methods have been shown useful in small search problems such as modeling of loop regions.[15] Perhaps, local enumerative optimization could be performed on structural regions deemed to be unfavorable in energy. Regarding replica exchange, recent work of Zuckerman, et al. suggests limitations in this approach for the sole purposes of canonical sampling at 298 K.[63] Alternative sampling approaches should be considered such as genetic algorithms[64] and resolution exchange.[63]

5. Conclusion

Statistical potentials are a fast alternative to force-field-based potentials. Unfortunately, without reference to unfolded and misfolded states in their parametrization they may not be well suited to temperature-based sampling schemes.[3] Ironically, this feature makes them useful in a framework where fast refolding of structures is desired. The Z-fold method, which can produce random rearrangements of model conformations, benefits greatly from the fast refolding capabilities of the DFIRE-AA potential. Undesirably, the DFIRE-AA potential, in particular, lacks certain multibody solvent effects which will tend to cause a protein to minimize its surface area and "sphericalize" regardless of the protein's actual fold type.

The force field potential we used here includes a state-of-the-art implicit solvent model.[65] As a tool for detecting near-native structures, we believe this potential is on par with other force field potentials currently available.[66] However, there are many deficiencies in the physics of most implicit solvent force fields that still need to be addressed (e.g., charge polarization, treatment of structural waters, etc.). Deficiencies aside, the noisy nature of the energy landscape will require creative new methods in exploring conformations adjacent to the models generated by a lower-resolution potential. Temperature-based sampling schemes, such as replica exchange using different temperature windows, may not be helpful for the refinement problem on the atomic scale without additional enhancements.

Acknowledgment. We thank Drs. M. Feig and L. Carlacci for helpful discussions. M.S.L. acknowledges financial support from the Department of Defense High Performance Computing Modernization Program Office (HPCMO), the Biotechnology High Performance Computing Software Applications Institute (HSAI), and the U.S. Army Medical Research and Material Command (Project No. RIID 02-4-1R-069). We thank the Army Research Laboratory Major Shared Resource Center for computer time. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the U.S. Army.

References

(1) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389.

(2) Miller, R. T.; Jones, D. T.; Thornton, J. M. *FASEB J.* **1996**, *10*, 171.

(3) Zhang, Y.; Kolinski, A.; Skolnick, J. *Biophys. J.* **2003**, *85*, 1145.

(4) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. *J. Mol. Biol.* **1997**, *268*, 209.

(5) Vincent, J. J.; Tai, C. H.; Sathyanarayana, B. K.; Lee, B. *Proteins* **2005**.

(6) Tsai, J.; Bonneau, R.; Morozov, A. V.; Kuhlman, B.; Rohl, C. A.; Baker, D. *Proteins* **2003**, *53*, 76.

(7) Misura, K. M.; Baker, D. *Proteins* **2005**, *59*, 15.

(8) Bradley, P.; Misura, K. M.; Baker, D. *Science* **2005**, *309*, 1868.

(9) Dill, K. A.; Phillips, A. T.; Rosen, J. B. *J. Comput. Biol.* **1997**, *4*, 227.

(10) Vorobjev, Y. N.; Almagro, J. C.; Hermans, J. *Proteins* **1998**, *32*, 399.

(11) Kryshtafovych, A.; Venclovas, C.; Fidelis, K.; Moult, J. *Proteins* **2005**, *61* Suppl 7, 225.

(12) Lu, H.; Skolnick, J. *Biopolymers* **2003**, *70*, 575.

(13) Fan, H.; Mark, A. E. *Protein Sci.* **2004**, *13*, 211.

(14) Xiang, Z.; Soto, C. S.; Honig, B. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7432.

(15) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A. *Proteins* **2004**, *55*, 351.

(16) Rohl, C. A.; Strauss, C. E.; Chivian, D.; Baker, D. *Proteins* **2004**, *55*, 656.

(17) Feig, M.; Brooks, C. L., III *Proteins* **2002**, *49*, 232.

(18) Hsieh, M. J.; Luo, R. *Proteins* **2004**, *56*, 475.

(19) Zhou, H.; Zhou, Y. *Protein Sci.* **2002**, *11*, 2714.

(20) Skolnick, J.; Zhang, Y.; Arakaki, A. K.; Kolinski, A.; Boniecki, M.; Szilagyi, A.; Kihara, D. *Proteins* **2003**, *53* Suppl 6, 469.

(21) Mackerell, A. D., Jr.; Bashford, D.; Bellott, D. M.; Dunbrack Jr., 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, I., 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.

(22) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999.

(23) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(24) Gilson, M. K.; Honig, B. *Proteins: Struct., Funct., Genet.* **1988**, *4*, 7.

(25) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III *J Comput. Chem.* **2003**, *24*, 1348.

(26) Zhu, J.; Xie, L.; Honig, B. *Proteins* **2006**, *65*, 463.

(27) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. *J. Med. Chem.* **2005**, *48*, 2325.

(28) Muegge, I.; Martin, Y. C. *J. Med. Chem.* **1999**, *42*, 791.

(29) Wang, K.; Fain, B.; Levitt, M.; Samudrala, R. *BMC Struct. Biol.* **2004**, *4*, 8.

(30) Dominy, B. N.; Brooks, C. L. *J. Comput. Chem.* **2002**, *23*, 147.

(31) Sippl, M. J.; Weitckus, S. *Proteins* **1992**, *13*, 258.

(32) Melo, F.; Feytmans, E. *J. Mol. Biol.* **1997**, *267*, 207.

(33) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 899.

(34) Zhang, C.; Liu, S.; Zhou, Y. *Protein Sci.* **2004**, *13*, 391.

(35) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins* **2004**, *56*, 310.

(36) Summa, C. M.; Levitt, M.; Degrado, W. F. *J. Mol. Biol.* **2005**, *352*, 986.

(37) Zhang, C.; Liu, S.; Zhou, H.; Zhou, Y. *Protein Sci.* **2004**, *13*, 400.

(38) Wang, G.; Dunbrack, R. L., Jr. *Nucleic Acids Res.* **2005**, *33*, W94.

(39) Skeel, R. D.; Tezcan, I.; Hardy, D. J. *J. Comput. Chem.* **2002**, *23*, 673.

(40) Lee, M. R.; Tsai, J.; Baker, D.; Kollman, P. A. *J. Mol. Biol.* **2001**, *313*, 417.

(41) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminatham, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.

(42) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96.

(43) Zhou, R.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12777.

(44) Olson, M. A.; Feig, M.; Brooks, C. L., III *J. Comput. Chem.*, submitted for publication.

(45) Zhang, Y.; Arakaki, A. K.; Skolnick, J. *Proteins* 2005.

(46) Ishikawa, Y.; Sugita, Y.; Nishikawa, T.; Okamoto, Y. *Chem. Phys. Lett.* **2001**, *333*, 199.

(47) Feig, M.; Karanicolas, J.; Brooks, C. L., III *J. Mol. Graph. Model* **2004**, *22*, 377.

(48) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577.

(49) Sheinerman, F. B.; Brooks, C. L., III *J. Mol. Biol.* **1998**, *278*, 439.

(50) Stumpff-Kane, A. W.; Feig, M. *Proteins* **2006**, *63*, 155.

(51) Samudrala, R.; Levitt, M. *Protein Sci.* **2000**, *9*, 1399.

(52) Chivian, D.; Kim, D. E.; Malmstrom, L.; Bradley, P.; Robertson, T.; Murphy, P.; Strauss, C. E.; Bonneau, R.; Rohl, C. A.; Baker, D. *Proteins* **2003**, *53* Suppl 6, 524.

(53) Xiang, Z.; Honig, B. *J. Mol. Biol.* **2001**, *311*, 421.

(54) Bursulaya, B.; Brooks, C. L., III *J. Am. Chem. Soc.* **1999**, *121*, 9947.

(55) Pande, V. S.; Baker, I.; Chapman, J.; Elmer, S. P.; Khaliq, S.; Larson, S. M.; Rhee, Y. M.; Shirts, M. R.; Snow, C. D.; Sorin, E. J.; Zagrovic, B. *Biopolymers* **2003**, *68*, 91.

(56) Dalton, W. S. *Semin. Oncol.* **2004**, *31*, 3.

(57) Dobson, C. M. *Semin. Cell Dev. Biol.* **2004**, *15*, 3.

(58) Misura, K. M.; Chivian, D.; Rohl, C. A.; Kim, D. E.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 5361.

(59) Xia, Y.; Levitt, M. *Proteins* **2004**, *55*, 107.

(60) Tappura, K.; Lahtela-Kakkonen, M.; Teleman, O. *J. Comput. Chem.* **2000**, *21*, 388.

(61) Berne, B. J.; Straub, J. E. *Curr. Opin. Struct. Biol.* **1997**, *7*, 181.

(62) Fujitsuka, Y.; Takada, S.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Proteins* **2004**, *54*, 88.

(63) Zuckerman, D. M.; Lyman, E. *J. Chem. Theor. Comput.* **2006**, *2*, 1200.

(64) Yang, Y.; Liu, H. *J. Comput. Chem.* **2006**, *27*, 1593.

(65) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L., III *J. Comput. Chem.* **2004**, *25*, 265.

(66) Zhu, J.; Alexov, E.; Honig, B. *J. Phys. Chem. B* **2005**, *109*, 3008.

(67) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33.

CT600195F