

JCTC

Journal of Chemical Theory and Computation

Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change

David L. Mobley,[†] John D. Chodera,[‡] and Ken A. Dill^{*,†}

Department of Pharmaceutical Chemistry, and Graduate Group in Biophysics, University of California at San Francisco, San Francisco, California 94158-2517

Received February 5, 2007

Abstract: Free energy calculations are increasingly being used to estimate absolute and relative binding free energies of ligands to proteins. However, computed free energies often appear to depend on the initial protein conformation, indicating incomplete sampling. This is especially true when proteins can change conformation on ligand binding, as free energies associated with these conformational changes are either ignored or assumed to be included by virtue of the sampling performed in the calculation. Here, we show that, in a model protein system (a designed binding site in T4 lysozyme), conformational changes can make a difference of several kcal/mol in computed binding free energies and that they are neglected in computed binding free energies if the system remains kinetically trapped in a particular metastable state on simulation timescales. We introduce a general “confine-and-release” framework for free energy calculations that accounts for these free energies of conformational change. We illustrate its use in this model system by demonstrating that an umbrella sampling protocol can obtain converged binding free energies that are independent of the starting protein structure and include these conformational change free energies.

INTRODUCTION

Computational tools are becoming increasingly important in drug discovery.¹ A major goal is to use these methods to

predict (absolute or relative) protein–ligand binding free energies. A great deal of effort^{2–4} has been focused on identifying which protein structures (i.e. *apo*, *holo*, or optimized in some manner) work best for estimating binding affinities. This emphasis on a single bound structure or conformation begs the question, “Can protein–ligand binding free energies be accurately predicted only a single protein conformation, or only some of the relevant protein conformations are considered?”. We demonstrate here that the answer is a decisive *no* in at least the model system considered here. There can be significant strain energies and free energy costs associated with trapping a protein into *any* metastable state, and, as we show here, the neglect of these costs can lead to substantial errors that depend on the metastable state chosen. (Here, we will use the term “structure” to refer to a single static structure and the term “metastable state” to refer to a favorable region of configuration space (set of structures) that is kinetically distinct from other such regions).

Computed binding free energies are often sensitive to the starting protein structure, even with alchemical free energy methods,^{5–10} which should not be the case if these simulations are converged. We believe this is for a similar reason: Even if full protein flexibility is allowed, the full range of relevant protein states may not be accessible on simulation timescales. This means that the protein is kinetically trapped in a particular metastable state, and the free energy cost of this trapping is neglected. Here, the problem is fundamentally a kinetic one: Large energy barriers can separate metastable protein states and trap the protein in a single metastable state on simulation timescales. Unfortunately, this trapping is inevitable whenever energy barriers are sufficiently large,¹¹ yet inadequate sampling even at the level of a single side chain rotameric state can lead to a difference in several kcal/mol in computed binding free energies.⁸ The problem is that it is necessary to adequately sample multiple relevant protein metastable states, including at least the metastable states containing both the *apo* and *holo* structures.

Here, we describe a framework we call “confine-and-release” for computing absolute binding free energies that correctly accounts for multiple relevant metastable states, such as protein conformational changes on ligand binding. The framework is general, in that it may be implemented in a number of different ways. We demonstrate the framework in a model binding site using one particular approach based on umbrella sampling, below.

In this work, we will refer to the problem of kinetic trapping or confinement as “[virtual] confinement”, to distinguish it from real confinement, where an external biasing potential is used. The confine-and-release approach

* Corresponding author e-mail: dill@maxwell.compbio.ucsf.edu.

[†] Department of Pharmaceutical Chemistry.

[‡] Graduate Group in Biophysics.

discussed here can deal with both cases, but we illustrate it here with virtual confinement.

The basic theory underlying absolute binding free energy calculations has previously been described in detail (for example, in refs 12 and 13). The absolute binding free energy is given as

$$\Delta G_{bind} = -k_B T \ln \frac{C^\circ}{8\pi^2} \frac{\sigma_P \sigma_L}{\sigma_{PL}} \frac{Z_{PL}}{Z_P Z_L} + P^\circ \Delta V_{PL}$$

where the protein-complex partition function is given by

$$Z_{PL} = \int_{\text{complex}} e^{-\beta U(\vec{r})} d\vec{r}$$

which is an integral over all of the protein–ligand conformations defining the bound state, and Z_P and Z_L are the corresponding partition functions consisting of integrals for the protein and ligand alone in solvent, respectively. C° denotes the standard concentration (1 M), and the σ factors are the symmetry numbers for the protein, for the ligand, and for the complex. These terms as well as the $P^\circ \Delta V_{PL}$ pressure-volume work term relate to the standard state and are explained in detail elsewhere.¹⁴

The essential point here is that evaluating the binding free energy necessarily involves integrating over all of the relevant (low potential energy) conformations of the protein and ligand, including all metastable states. If that integration is incomplete, as in the case of inadequate sampling, the quantity calculated will not be a true binding free energy. In such cases of kinetic trapping, the free energy that is calculated can be called a “confined” binding free energy—it measures the binding free energy of the system [virtually] confined to a metastable state (for example, the region corresponding to the *holo* structure) and hence neglects certain components of the true binding free energy such as strain energies. This observation is related to that made earlier by Straatsma and McCammon in the context of solvation for molecules with multiple relevant rotameric states: Unless all relevant metastable states are sampled in some manner, computed free energies are “unreasonable” and incorrect.¹⁵

We illustrate the problem with the example of *p*-xylene binding in a simple apolar cavity (an engineered cavity in T4 lysozyme) studied computationally by Deng and Roux.⁸ Here, a single valine side chain reorients upon ligand binding (as seen by comparing the *apo* and *holo* structures¹⁹).

We use simulation protocols employed previously¹⁴ with minor modifications described in the Supporting Information. These modifications involve improved parameters for the Particle mesh Ewald¹⁶ treatment of long-range electrostatics, addition of a separate vacuum calculation in order to finish the cycle for computing binding free energies, and addition of a long-range correction term to account for attractive dispersion interactions between the ligand and protein that are neglected when simulations are run with a short cutoff. Very briefly, the overall procedure involves first restraining the ligand in complex, then annihilating the ligand’s electrostatic interactions, followed by decoupling its Lennard-Jones interactions. The restraints are then analytically removed, and this is equivalent to having a protein with no ligand, plus a noninteracting, neutral ligand in solvent. The ligand electrostatic interactions are then restored in solvent, completing the thermodynamic cycle. The free energy of

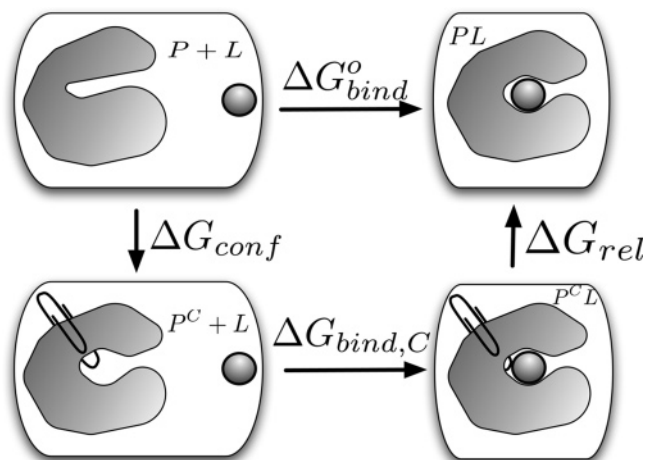


Figure 1. Thermodynamic cycle for the confine-and-release framework. The quantity we want to calculate is ΔG_{bind}^o (top), the free energy difference for the process $P + L \rightarrow PL$. Kinetic trapping (virtual confinement) or deliberate confinement can keep conformations changes from being sampled (shown graphically by a paperclip). When this happens, computed free energies are actually confined binding free energies, $\Delta G_{bind,C}^o$ (bottom arrow). To relate these to true binding free energies, it is necessary to compute the free energy of confining the protein in the absence of the ligand (left arrow) and releasing the protein in the presence of the ligand (right arrow).

making each of these transformations is computed using free energy methods with a series of separate simulations at different alchemical intermediate states (λ values).

We start from the *apo* structure. We observe that the system remains trapped in the metastable state containing that structure over the course of all equilibration and production trajectories involved in the free energy calculation (1.11 ns at each λ value). The resulting computed binding free energy (at standard pressure and 300 K) is -2.96 ± 0.06 kcal/mol (where the uncertainty represents 1 SD over a set of block bootstrap trials as described in the Supporting Information and previously¹⁴). If, instead, we start from the *holo* structure, we compute a binding free energy of -7.27 ± 0.09 kcal/mol. If we examine the valine side chain χ_1 dihedral angle as a function of time for every simulation in these free energy calculations, we find that, in each case, it remains in its initial rotameric state. The valine does not cross its torsional energy barrier on simulation timescales. This causes significant errors: One computed binding free energy indicates *p*-xylene is a millimolar binder; the other indicates it is a micromolar binder.

We solve this problem using the “confine-and-release” framework, depicted in the thermodynamic cycle in Figure 1; there, confinement (in this case virtual confinement) is illustrated by a paper clip. We begin by recognizing that our calculated free energies are “confined” binding free energies, that is, free energies for binding of the ligand to a protein that is restricted to a particular metastable state. Then, to compute the true binding free energy, we must add the free energy of confining the protein to that metastable state when no ligand is bound and the free energy of releasing the protein from its confinement when the ligand is bound. Hence $\Delta G_{bind}^o = \Delta G_{conf} + \Delta G_{bind,C}^o + \Delta G_{rel}$. In this expression, ΔG_{bind}^o is the true (standard) binding free energy; $\Delta G_{bind,C}^o$ is the standard binding free energy of the ligand to the confined protein; ΔG_{conf} is the free energy of confining the protein to

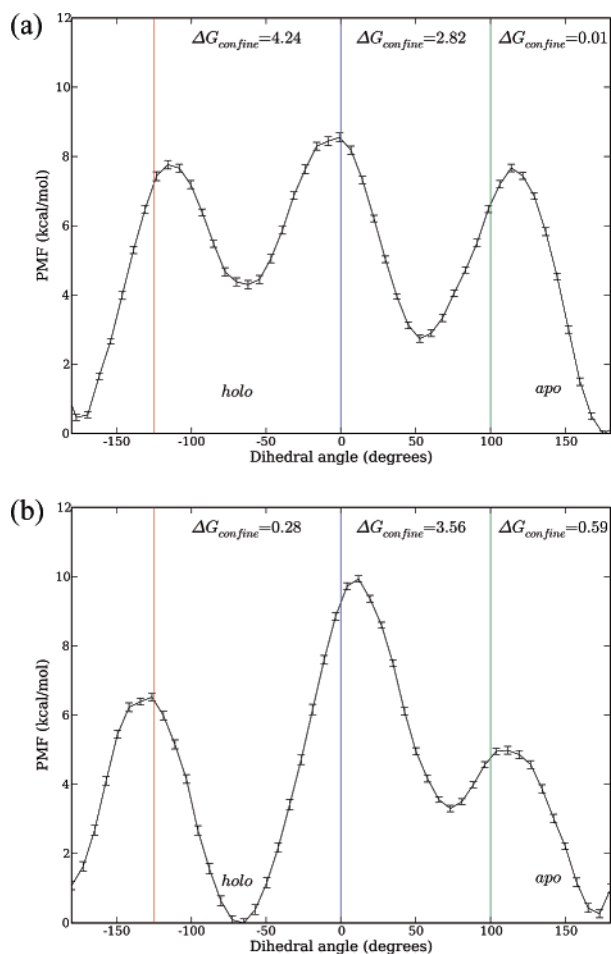


Figure 2. Potential of mean force for rotating the valine 111 side chain, with (b) and without (a) the ligand. Above each of the three regions is shown the free energy of confining Val111 to that metastable state. The *apo* metastable state corresponds to the first region on the left and the far right region (since the dihedral angle is periodic). Error bars represent statistical uncertainties corresponding to 1 SD. Uncertainties for confinement to each well are given in the text.

this smaller region of configuration space in the unbound state; and ΔG_{rel} is the free energy of releasing the protein from conformational confinement in the bound state. This can be thought of as a generalization of conformational biasing potentials.⁸

Free energies of confinement and release can be computed using a variety of different algorithms. Here, since there is a single relevant degree of freedom that needs to be sampled, we employed umbrella sampling.¹⁷ We computed the potential of mean force (PMF) for rotating the side chain of Val111 throughout its range of motion in both the bound and unbound states (Figure 2) (details available in the Supporting Information). From the PMF, we computed the free energy of confining the side chain to each rotameric state (as described in the Supporting Information). To test reproducibility of the corrected true free energy, the entire confine-and-release procedure was performed twice: Once using the *apo* structure and the associated metastable state (beginning from the *apo* crystal structure) for the binding calculation, and once using the *holo* structure (and metastable state) for the binding calculation. The same framework applies in either case.

Using the *apo* metastable state, we compute a confinement free energy in the unbound state of 0.01 ± 0.04 and a release free energy of -0.6 ± 0.1 kcal/mol in the bound state. Combining this with the computed confined binding free energy of -2.96 ± 0.06 kcal/mol yields a total binding free energy of -3.5 ± 0.2 kcal/mol. Alternatively, using the *holo* metastable state, the confinement free energy is 4.2 ± 0.2 kcal/mol, and the release free energy is 0.28 ± 0.08 kcal/mol, which, when added to the computed confined binding free energy of -7.27 ± 0.09 kcal/mol, yields a total binding free energy of -3.3 ± 0.2 kcal/mol. The difference between the total binding free energies computed from the different crystal structures is now only 0.2 ± 0.3 kcal/mol—statistically indistinguishable from zero. Hence, we believe these values now represent the overall binding free energy, corrected for inadequate sampling of Val111. In this case, the experimental binding free energy is -4.67 ± 0.06 kcal/mol, so our approach substantially improves agreement with experiment, especially when beginning from the *holo* structure.

Figure 2 shows that, for *p*-xylene, a single rotameric state dominates when the ligand is absent (Figure 2a), and a different rotameric state dominates when the ligand is present (Figure 2b), although in (Figure 2b), both rotameric states are relevant—that is, both states contribute significant fractions to the free energy. In general, the relevant rotameric state may differ in the presence and absence of the ligand, or there may be multiple relevant states in either case.

Previous work on this binding site, beginning from the *holo* structure for each ligand, produced binding free energies that were 2.05 to 4.40 kcal/mol too negative relative to experiment⁸ for those compounds where Val111 reorients on ligand binding (*p*-xylene, *o*-xylene, and *n*-butylbenzene, isobutylbenzene).¹⁹ Indeed, these compounds were essentially the worst outliers in that study. Here, due to kinetic trapping, we had to apply a positive correction of 3.9 kcal/mol for *p*-xylene beginning from the *holo* structure. Though the previous work used a different force field and parameters, it seems likely that kinetic trapping of Val111 can explain a significant portion of the observed errors there. For example, if we applied our correction to their calculated value for *p*-xylene (-9.06 kcal/mol), the resulting binding free energy would be -5.06 kcal/mol (calculated) versus -4.67 kcal/mol (experiment).

Here, the confine-and-release technique was applied to a single degree of freedom. As other situations will undoubtedly require careful sampling of more than a single (known) degree of freedom, the calculation of free energies of confinement and release from potentials of mean force is not necessarily a general strategy for applying this framework. Rather, the key points here are as follows: First, correct binding free energies can only be obtained when protein conformational change is correctly accounted for. Second, protein conformational change contributes substantially to the overall binding free energy, even for changes as small as the reorientation of single side chains. Thus, protein conformational changes should not simply be ignored in binding free energy calculations.

To compute confine-and-release free energies with the umbrella sampling approach discussed here, there are several requirements. First, one must know (i.e., crystallographically) or be able to predict (i.e., from side chain Monte Carlo sampling⁴) all of the relevant slow degrees of freedom.

Second, there must be relatively few of these degrees of freedom so that deliberate sampling of them is tractable. In this particular binding site, crystallographic evidence suggests that the only side chain reorientation on ligand binding is that of Val111,¹⁹ thus it is straightforward to apply this umbrella sampling approach. In general, however, umbrella sampling may prove impractical. But the confine-and-release approach itself (Figure 1) only requires a method of computing the confine and release free energies; this need not be done with umbrella sampling.

The confine-and-release cycle used here, involving confining and releasing the protein to compute true binding free energies, can easily be extended to a variety of other applications. In the example above, [virtual] confinement is due to kinetic trapping. But deliberate confinement by external restraints may also be useful. This could help, for example, for proteins that undergo relatively large conformational changes on ligand binding, such flap closure in HIV protease. Without this confinement, the protein could begin to deform back to its *apo* structure as the ligand is alchemically removed, leading to sampling problems. These sampling problems can be severe: At some alchemical intermediate states, *both* metastable states could be relevant, and the protein would need to sample both several times during the simulation. In HIV protease, for example, these conformational changes may take place on the microsecond to millisecond time scale and are difficult to sample even with long molecular dynamics trajectories.¹⁸ Thus, this confinement approach can also potentially aid convergence at intermediate alchemical states.

We conclude that computing binding free energies requires more than just computing the binding free energy of the ligand to a particular conformational state of the protein; it also requires a calculation of the free energy associated with confining the protein to that particular conformational state with and without the ligand present. These confinement free energies can be substantial, even for the relatively rigid binding site considered here. Elsewhere, we have noted that similar problems can arise when sampling ligand orientations.¹⁴ Unless free energy calculations include sufficient sampling to adequately include these conformational changes at all stages of the transformation, computed “binding free energies” are not true binding free energies. In short, a dependence of free energy estimates on initial protein or ligand structure can indicate that simulations are not adequately sampling the relevant regions of configuration space. The confine-and-release framework we introduce here can be used to design approaches that isolate and solve these sampling problems in a systematic and controlled manner for free energy calculations.

The importance of conformational change in binding free energies has ramifications that extend beyond just alchemical free energy calculations. Virtual screening methods that rely on docking and scoring using a single structure need to reconsider the assumption that binding free energies can be estimated given an appropriate bound structure. Free energy costs associated with trapping the protein to the *holo* structure, or to any structure chosen, may be significant and probably need to be correctly accounted for to accurately predict binding free energies. This problem cannot be avoided simply by comparing relative binding free energies of different ligands, either. In this binding site, for example, it

is known that some ligands bind *without* reorientation of the Val111 side chain, while others require the reorientation seen here in the case of *p*-xylene.¹⁹ This means that free energy costs required to bind different ligands can be substantially different—up to several kcal/mol, based on the data presented here. Thus, when estimating relative binding free energies using the same protein structure, errors will be different for different ligands rather than canceling out.

In summary, the confine-and-release framework presented here provides a rigorous way to correct for inadequate or restricted computational sampling of protein degrees of freedom in ligand binding free energy calculations. This approach can give binding free energies that are independent of the starting protein structure (i.e., *apo* or *holo*) and therefore yield true binding free energies for the given force field. Here we have demonstrated this approach using an umbrella sampling technique for computing the confine-and-release free energies; sampling requirements will probably limit this particular technique to accounting for inadequate sampling of a limited number of degrees of freedom. But the framework is more general.

ACKNOWLEDGMENT

We thank Benoît Roux (University of Chicago) for a critical reading of the manuscript. J.D.C. was supported in part by HHMI and IBM predoctoral fellowships. D.L.M. and K.A.D. acknowledge NIH grant GM63592, Anteon Corporation grant USAF-5408-04-SC-0008, and a UCSF Sandler Award.

Supporting Information Available: Details of umbrella sampling calculations, simulation protocols, and computation of confinement and release free energies. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES

- (1) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813–1818.
- (2) McGovern, S. L.; Shoichet, B. K. Information Decay in Molecular Docking Screens Against Holo, Apo, and Modeled Conformations of Enzymes. *J. Med. Chem.* **2003**, *46*, 2895–2907.
- (3) Huang, N.; Kalyanaraman, C.; Irwin, J. J.; Jacobson, M. P. Molecular Mechanics Methods for Predicting Protein-Ligand Binding. *J. Chem. Inf. Model.* **2006**, *46*, 243–253.
- (4) Meiler, J.; Baker, D. ROSETTALIGAND: Protein-Small Molecule Docking with Full Side-Chain Flexibility. *Proteins: Struct., Funct., Bioinf.* **2006**, *66*, 538–548.
- (5) Wang, J.; Deng, Y.; Roux, B. Absolute Binding Free Energy Calculations Using Molecular Dynamics Simulations with Restraining Potentials. *Biophys. J.* **2006**, *91*, 2798–2814.
- (6) Shirts, M. R. Calculating Precise and Accurate Free Energies in Biomolecular Simulations. Ph.D. Dissertation, Stanford University, 2004. ProQuest location: <http://www.lib.umi.com/dissertations/fullcit/3153076> (accessed Feb 1, 2007).
- (7) Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C. D.; Shirts, M. R.; Sorin, E. J.; Pande, V. S. Direct Calculation of the Binding Free Energies of FKBP Ligands. *J. Chem. Phys.* **2005**, *123*, 084108.
- (8) Deng, Y.; Roux, B. Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *J. Chem. Theor. Comput.* **2006**, *2*, 1255–1273.
- (9) van den Bosch, M.; Swart, M.; Snijderst, J. G.; Berendsen, H. J. C.; Mark, A. E.; Oostenbrink, C.; van Gunsteren, W. F.; Canters, G. W. Calculation of the Redox Potential of the Protein Azurin and Some Mutants. *Chem. Bio. Chem.* **2005**, *6*, 738–746.
- (10) Zhou, Y.; Oostenbrink, C.; Jongejan, A.; Hagen, W. R.; de Leeuw, S. W.; Jongejan, J. A. Computational Study of Ground-State Chiral Induction in Small Peptides: Comparison of the Relative Stability of

- Selected Amino Acid Dimers and Oligomers in Homochiral and Heterochiral Combinations. *J. Comput. Chem.* **2006**, *27*, 857–867.
- (11) Leitgeb, M.; Schroder, C.; Boresch, S. Alchemical Free Energy Calculations and Multiple Conformational Substates. *J. Chem. Phys.* **2005**, *122*, 084109.
- (12) Gilson, M. K.; Given, J. A.; Bush, G. L.; McCammon, J. A. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys. J.* **1997**, *72*, 1047–1069.
- (13) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *J. Phys. Chem. B* **2003**, *107*, 9535–9551.
- (14) Mobley, D. L.; Chodera, J. D.; Dill, K. A. On the Use of Orientational Restraints and Symmetry Number Corrections in Alchemical Free Energy Calculations. *J. Chem. Phys.* **2006**, *125*, 084902.
- (15) Straatsma, T. P.; McCammon, J. A. Treatment of Rotational Isomers in Free Energy Evaluations. Analysis of the Evaluation of Free Energy Differences by Molecular Dynamics Simulations of Systems with Rotational Isomeric States. *J. Chem. Phys.* **1989**, *90*, 3300–3304.
- (16) Essmann, U.; Perera, L.; Berkowitz, M.; Darden, T.; Lee, H.; Pedersen, Lee, G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (17) Kumar, S. J.; Rosenberg, M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
- (18) Hornak, V.; Okur, A.; Rizzo, R. C.; Simmerling, C. HIV-1 Protease Flaps Spontaneously Open and Reclose in Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 915–920.
- (19) Morton, A.; Matthews, B. W. Specificity of Ligand Binding in a Buried Nonpolar Cavity of T4 Lysozyme: Linkage of Dynamics and Structural Plasticity. *Biochemistry* **1995**, *34*, 8576–8588.

CT700032N

Dynamic Formation and Breaking of Disulfide Bonds in Molecular Dynamics Simulations with the UNRES Force Field

M. Chinchio,[†] C. Czaplewski,^{†,‡} A. Liwo,^{†,‡} S. Ołdziej,^{†,‡} and H. A. Scheraga^{*,†}

Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301, and Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland

Received April 6, 2007

Abstract: Many proteins contain disulfide bonds that are usually essential for maintaining function and a stable structure. Several algorithms attempt to predict the arrangement of disulfide bonds in the context of protein structure prediction, but none can simulate the entire process of oxidative folding, including dynamic formation and breaking of disulfide bonds. In this work, a potential function developed to model disulfide bonds is coupled with the united-residue (UNRES) force field, and used in both canonical and replica exchange molecular dynamics simulations to produce complete oxidative folding pathways. The potential function is obtained by introducing a transition barrier that separates the bonded and nonbonded states of the half-cystine residues. Tests on several helical proteins show that improved predictions are obtained when dynamic disulfide-bond formation and breaking are considered. The effect of the disulfide bonds on the folding kinetics is also investigated, particularly their role in stabilizing folding intermediates, resulting in slower folding.

1. Introduction

Disulfide bonds are often present in the native conformations of proteins and contribute to their stability and function. In some cases the disulfide bonds are essential for maintaining the structure of the protein,¹ while for other proteins some bonds can be broken without causing drastic conformational changes.^{2–4} Oxidative folding is a very complex process, by which a fully reduced and unfolded protein reaches its native structure (conformational folding) with all native disulfide bonds. Over the years, experimentally determined folding pathways for different proteins have shown a high degree of diversity in the number and type of intermediate states that appear during folding.^{1,5–8} Two limiting-case mechanisms have been described:⁹ the folded-precursor mechanism, in which conformational folding precedes the formation of the native disulfide bonds,

and the quasi-stochastic mechanism, in which disulfide bonds (native and non-native) form in unfolded conformations. Evidence has been found for both,^{9,10} and for many proteins the correct mechanism probably lies somewhere in between.

Early theoretical work identified the decreased entropy of the unfolded state as the main source of the increased stability observed for disulfide-bonded proteins.^{11–14} In this chain-entropy model, the effect becomes stronger as the sequence separation of the pair of linked residues grows. However, individual proteins can deviate significantly from this ideal behavior, as evidenced by the difficulties encountered in engineering disulfide-stabilized proteins.^{15–19} In many cases, one cannot ignore the effect of the introduction of a disulfide bond on the folded state, particularly when the protein structure is perturbed and strained by this bond.

Various aspects of this problem have been attacked computationally using both Monte Carlo and molecular dynamics techniques, and the methods employed have ranged from very simplified lattice Gō-like models to the most

* Corresponding author phone: (607)255-4034; fax: (607)254-4700; e-mail: has5@cornell.edu.

[†] Cornell University.

[‡] University of Gdańsk.

detailed all-atom representations.^{17,20–23} However, in most cases the disulfide bonds are not allowed to form and break during the simulations, but rather they are set at the beginning and remain fixed throughout. A few approaches have attempted to model this process dynamically, either by employing a highly simplified 2-D lattice representation²⁴ or by restricting the dynamics to the packing of preassigned secondary structure elements.²⁵ While several disulfide-bond prediction algorithms exist,^{26–30} the first general approach to protein structure prediction including dynamic disulfide-bond formation and breaking was recently proposed by some of the present authors.³¹ In that approach, no assumptions were made as to the positions of native disulfides, and predictions were produced by an unbiased global search for an energy minimum, based on local energy minimization. The energy function and united-residue (UNRES) model employed were previous versions of those presented here, which will be described in more detail in later sections. However, due to its use of local energy minimization in the global search procedure, thermodynamic or kinetic information could not be obtained directly. The work presented here addresses this limitation by using the recently developed molecular dynamics algorithm applied to the UNRES model³² and introducing dynamic formation and breaking of disulfide bonds. The method is tested on four α -helical proteins, covering a range of fold types and various arrangements of disulfide bonds. The results show that the addition of disulfide bonds can significantly improve the quality of blind structure predictions while, at the same time, providing insight into the role of disulfide bonds in the stabilization of proteins and their effect on the kinetics of folding.

2. Methods

2.1. The UNRES Force Field. In the UNRES model,^{33–47} a polypeptide chain is represented by a sequence of α -carbon (C^α) atoms linked by virtual bonds with attached united side chains (SC) and united peptide groups (p). Each united peptide group is located in the middle between two consecutive α -carbons. Only these united peptide groups and the united side chains serve as interaction sites, the α -carbons serving only to define the chain geometry, as shown in Figure 1.

The UNRES force field has been derived as a Restricted Free Energy (RFE) function of an all-atom polypeptide chain plus the surrounding solvent, where the all-atom energy function is averaged over the degrees of freedom that are lost when passing from the all-atom to the simplified system (i.e., the degrees of freedom of the solvent, the dihedral angles χ for rotation about the bonds in the side chains, and the torsional angles λ for rotation of the peptide groups about the $C^\alpha \cdots C^\alpha$ virtual bonds).^{37,38,48} The RFE is further decomposed into factors arising from interactions within and between a given number of united interaction sites.³⁸ Expansion of the factors into generalized Kubo cumulants⁴⁹ facilitated the derivation of approximate analytical expressions for the respective terms,^{37,38} including the *multibody* or *correlation* terms. The theoretical basis of the force field is described in detail elsewhere.³⁸

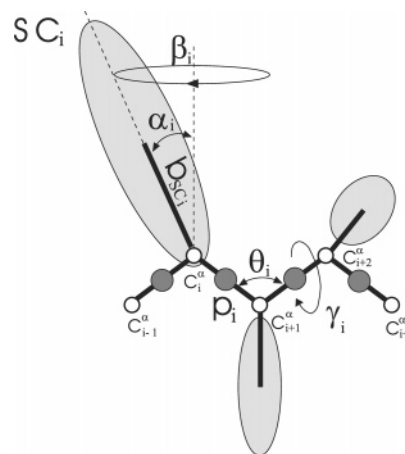


Figure 1. The UNRES model of polypeptide chains. The interaction sites are side-chain centroids of different sizes (SC) and peptide-bond centers (p) indicated by shaded circles, whereas the α -carbon atoms (small empty circles) are introduced only to assist in defining the geometry. The virtual $C^\alpha \cdots C^\alpha$ bonds have a variable length centered around 3.8 Å, corresponding to a planar trans peptide group; the virtual-bond (θ) and dihedral (γ) angles are variable. Each side chain is attached to the corresponding α -carbon with a variable “bond length”, b_{SC_i} , variable “bond angle”, α_{SC_i} , formed by SC_i and the bisector of the angle defined by C^α_{i-1} , C^α_i , and C^α_{i+1} , and with a variable “dihedral angle” β_{SC_i} , of counterclockwise rotation about the bisector, starting from the right side of the C^α_{i-1} , C^α_i , C^α_{i+1} frame.

The energy of the virtual-bond chain is expressed by eq 1:

$$\begin{aligned}
 U = & w_{SC} \sum_{i < j} U_{SC_i SC_j} + w_{SCp} \sum_{i \neq j} U_{SC_i p_j} + w_{pp}^{VDW} \sum_{i < j-1} U_{p_i p_j}^{VDW} + \\
 & w_{pp}^{el} f_2(T) \sum_{i < j-1} U_{p_i p_j}^{el} + w_{tor} f_2(T) \sum_i U_{tor}(\gamma_i) + \\
 & w_{tord} f_3(T) \sum_i U_{tord}(\gamma_i, \gamma_{i+1}) + w_b \sum_i U_b(\theta_i) + \\
 & w_{rot} \sum_i U_{rot}(\alpha_{SC_i}, \beta_{SC_i}) + \sum_{m=3,4} w_{corr}^{(m)} f_m(T) U_{corr}^{(m)} + \\
 & \sum_{m=3,4} w_{turn}^{(m)} f_m(T) U_{turn}^{(m)} + w_{bond} \sum_{i=1}^{nbond} U_{bond}(d_i) \quad (1)
 \end{aligned}$$

The term $U_{SC_i SC_j}$ represents the mean free energy of the hydrophobic (hydrophilic) interactions between the side chains, which implicitly contains the contributions from the interactions of the side chain with the solvent. The term $U_{SC_i p_j}$ denotes the excluded-volume potential of the side-chain-peptide-group interactions. The peptide-group interaction potential is split into two parts: the Lennard-Jones interaction energy between peptide group centers ($U_{p_i p_j}^{VDW}$) and the average electrostatic energy between peptide group dipoles ($U_{p_i p_j}^{el}$); the second of these terms accounts for the tendency to form backbone hydrogen bonds between peptide groups p_i and p_j . U_{tor} , U_{tord} , U_b , and U_{rot} are the virtual-bond torsional terms, virtual-bond double-torsional terms, virtual-bond angle bending terms, and side-chain rotamer terms; these terms account for the local propensities of the polypeptide chain. The terms $U_{corr}^{(m)}$ represent the *correlation* or *multibody*

contributions from the coupling between backbone-local and backbone-electrostatic interactions, and the terms $U_{\text{turn}}^{(m)}$ are correlation contributions involving m consecutive peptide groups; they are, therefore, termed turn contributions. The correlation contributions were derived^{37,38} from a generalized-cumulant expansion⁴⁹ of the restricted free energy (RFE) of the system consisting of the polypeptide chain and the surrounding solvent. The multibody terms are indispensable for reproduction of regular α -helical and β -sheet structures. The terms $U_{\text{bond}}(d_i)$, where d_i is the length of the i th virtual bond, are simple harmonic potentials of virtual bond distortions, introduced for the molecular dynamics implementation, and nbond is the number of virtual bonds. The $f_m(T)$ terms express the temperature dependence of the restricted free energy function. Their main effect is to reduce the weight of multibody terms at high temperatures, thus preventing the premature formation of secondary structure, characteristic of older versions of this force field.

The internal parameters of $U_{p_i p_j}$, U_{tor} , U_{tord} , $U_{\text{corr}}^{(m)}$, and $U_{\text{turn}}^{(m)}$ were derived by fitting the analytical expressions to the RFE surfaces of model systems computed by quantum mechanics at the MP2/6-31G** ab initio level,^{42,43} while the parameters of $U_{\text{SC}_i \text{SC}_j}$, $U_{\text{SC}_i p_j}$, U_b , and U_{rot} were derived by fitting the calculated distribution functions to those determined from the PDB;³⁶ work is currently in progress to obtain these parameters from quantum mechanical ab initio calculations of the potentials of mean force of appropriate model systems. The w 's (the weights of the energy terms), the internal parameters within each cumulant term, and the mean free energies of side-chain interactions of the $U_{\text{SC}_i \text{SC}_j}$ energy term were obtained by a hierarchical optimization⁵⁰ of the energy function based on protein 1GAB, a three-helix bundle. This force field is the first version of UNRES parametrized specifically for canonical simulations. It was designed so that folding occurs at physiological temperatures (≈ 300 K).

In this version of the force field, the side-chain pairwise interaction potential is represented by the orientation dependent Gay-Berne⁵¹ functional form, given by

$$U_{\text{SC}_i \text{SC}_j}^{(\text{GB})} = 4(|\epsilon_{ij}^{(\text{GB})}|x_{ij}^{12} - \epsilon_{ij}^{(\text{GB})}x_{ij}^6) \quad (2)$$

$$\epsilon_{ij}^{(\text{GB})} = \epsilon_{ij} \epsilon_{0ij}^{(\text{GB})}$$

$$x_{ij} = \frac{\sigma_{0ij}}{r_{ij} - (\sigma_{ij} - \sigma_{0ij})}$$

where $\epsilon_{0ij}^{(\text{GB})}$ and σ_{0ij} are constant internal parameters ($\epsilon_{0ij}^{(\text{GB})} > 0$ only for hydrophobic-hydrophobic interactions), ϵ_{ij} and σ_{ij} depend on the relative orientation of the two side chains (more details are given in an earlier publication³⁵), and r_{ij} is the distance between the side-chain centroids, as shown in Figure 2. For any given orientation, the van der Waals well depth is given by $\epsilon_{ij}^{(\text{GB})}$. However, ϵ_{ij} remains in the range 1.0–2.0 for different orientations; therefore, the well depth is always at least $\epsilon_{0ij}^{(\text{GB})}$.

2.2. Disulfide Bonds in UNRES. Disulfide-bond potentials were first introduced into the UNRES force field as simple harmonic potentials which depended only on the

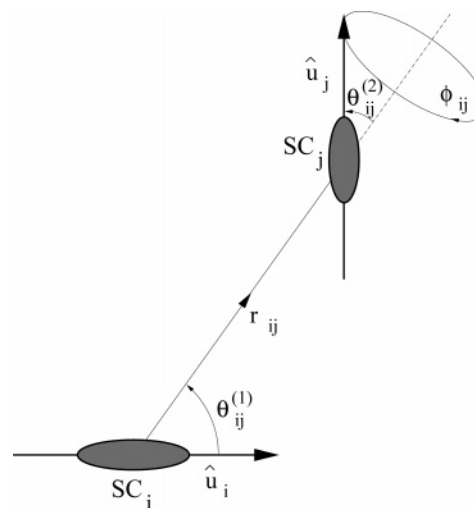


Figure 2. Definition of the orientation of two anisotropic side chains, SC_i and SC_j , represented by ellipsoids of revolution.³⁵ The relative position of the centers of the side chains are given by the vector \mathbf{r}_{ij} (of length r_{ij}). The principal axes of the ellipsoids are assumed to be colinear with the C^α – SC lines; their directions are given by the unit vectors $\hat{\mathbf{u}}_i$ and $\hat{\mathbf{u}}_j$. The variables defining the relative orientations of the ellipsoids are the angles $\theta_{ij}^{(1)}$ (the planar angle between $\hat{\mathbf{u}}_i$ and \mathbf{r}_{ij}), $\theta_{ij}^{(2)}$ (the planar angle between $\hat{\mathbf{u}}_j$ and \mathbf{r}_{ij}), and ϕ_{ij} (the angle of counterclockwise rotation of the vector $\hat{\mathbf{u}}_j$ about the vector \mathbf{r}_{ij} from the plane defined by $\hat{\mathbf{u}}_i$ and \mathbf{r}_{ij} when looking from the center of SC_j toward the center of SC_i).

distance between the side-chain centroids of the relevant half-cystines.³¹ A more sophisticated model was later developed which takes into account the relative orientation of the two side chains as well as the distance between them.⁵² In this model, the disulfide-bond energy is expressed by

$$U_{\text{SC}_i \text{SC}_j}^{(\text{SS})} = \epsilon_0^{(\text{SS})} + \sum_{n=1}^3 v_n \zeta_{ij}^n + k_0((\eta_{ij}^{(1)})^2 + (\eta_{ij}^{(2)})^2) + k_1(\eta_{ij}^{(1)} + \eta_{ij}^{(2)})(r_{ij} - d_0) + k_2(r_{ij} - d_0)^2$$

$$\eta_{ij}^{(1)} = 1 - \hat{\mathbf{u}}_i \cdot \hat{\mathbf{r}}_{ij}$$

$$\eta_{ij}^{(2)} = 1 + \hat{\mathbf{u}}_j \cdot \hat{\mathbf{r}}_{ij}$$

$$\zeta_{ij} = \hat{\mathbf{u}}_i \cdot \hat{\mathbf{u}}_j - (\hat{\mathbf{u}}_i \cdot \hat{\mathbf{r}}_{ij})(\hat{\mathbf{u}}_j \cdot \hat{\mathbf{r}}_{ij}) \quad (3)$$

where the vectors $\hat{\mathbf{u}}$ and \mathbf{r} are defined in Figure 2, $\epsilon_0^{(\text{SS})}$ is an adjustable parameter which defines the well depth, and the internal parameters d_0 , v_n , and k_n were derived based on ab initio calculations of diethyl disulfide at the RHF/6-31G** level.⁵² The numerical values of these parameters are summarized in Table 1.

Until now,³¹ $U_{\text{SC}_i \text{SC}_j}^{(\text{GB})}$ had to be replaced with $U_{\text{SC}_i \text{SC}_j}^{(\text{SS})}$ to calculate the interaction energy between half-cystines forming a disulfide bond. Formation (or breaking) of a disulfide bond was effectively achieved by modifying the relevant term in the UNRES force field. This approach works very well for minimization-based search methods, where the goal is simply to identify the lowest energy conformations. It was successfully applied³¹ to the CSA search procedure,^{53–55} a

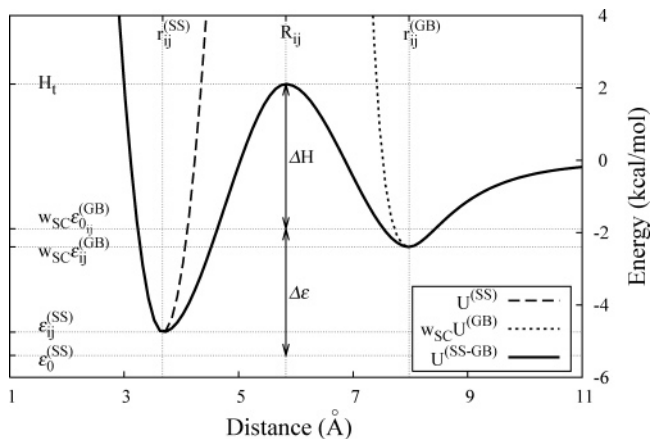


Figure 3. Distance dependence of the energy term $U_{SC,SC_j}^{(SS-GB)}$, for a representative arrangement of side chains ($\theta_{ij}^{(1)} = 30^\circ$, $\theta_{ij}^{(2)} = 150^\circ$, $\phi_{ij} = 120^\circ$). The dashed and dotted curves represent $U_{SC,SC_j}^{(SS)}$ and $U_{SC,SC_j}^{(GB)}$. The most crucial distances (the two minima $r_{ij}^{(SS)}$ and $r_{ij}^{(GB)}$ and the maximum R_{ij}) are shown. The portions of these potentials between $r_{ij}^{(SS)}$ and $r_{ij}^{(GB)}$ are replaced by an artificial transition barrier. The key energy levels are also illustrated, viz., the top of the transition barrier (H_t); the depth ($w_{SC}\epsilon_{ij}^{(GB)}$) of the $U_{SC,SC_j}^{(GB)}$ potential well and its defining parameter ($\epsilon_{ij}^{(GB)}$); and the depth of the $U_{SC,SC_j}^{(SS)}$ potential well ($\epsilon_{ij}^{(SS)}$) and its defining parameter ($\epsilon_0^{(SS)}$). Also shown are the two adjustable parameters ΔH and $\Delta\epsilon$.

Table 1. Internal Parameters of the Disulfide-Bond Potential

parameter	value
d_0	3.78 Å
v_1	-1.08 kcal/mol
v_2	7.61 kcal/mol
v_3	13.70 kcal/mol
k_0	11.0 kcal/mol
k_1	12.0 kcal/mol/Å
k_2	15.1 kcal/mol/Å ²

genetic algorithm which has been used with UNRES for the past several years. CSA was extended to allow for dynamic rearrangement of disulfide bonds during the simulation by introducing new genetic operators to treat the formation and breaking of the bonds.³¹ However, in the case of molecular dynamics simulations, this approach cannot be applied easily without destroying the thermodynamic properties of the algorithm. Instead, a new approach is presented here that combines the terms $U_{SC,SC_j}^{(GB)}$ and $U_{SC,SC_j}^{(SS)}$ by introducing an artificial transition barrier between the two minima. The resulting potential function, given by eq 4, is shown in Figure 3.

$$U_{SC,SC_j}^{(SS-GB)} = \begin{cases} U_{SC,SC_j}^{(SS)} & \text{if } r_{ij} < r_{ij}^{(SS)} \\ \epsilon_{ij}^{(SS)} g(r_{ij}; R_{ij}, r_{ij}^{(SS)}) + H_t g(r_{ij}; r_{ij}^{(SS)}, R_{ij}) & \text{if } r_{ij}^{(SS)} \leq r_{ij} < R_{ij} \\ H_t g(r_{ij}; r_{ij}^{(GB)}, R_{ij}) + w_{SC}\epsilon_{ij}^{(GB)} g(r_{ij}; R_{ij}, r_{ij}^{(GB)}) & \text{if } R_{ij} \leq r_{ij} < r_{ij}^{(GB)} \\ w_{SC}U_{SC,SC_j}^{(GB)} & \text{if } r_{ij} \geq r_{ij}^{(GB)} \end{cases} \quad (4)$$

In the version of UNRES used for the calculations presented here, the values⁵⁰ of w_{SC} and $\epsilon_0^{(GB)}$ are 1.35279 and -1.4013, respectively. The side-chain distances corresponding to the minima of the $U_{SC,SC_j}^{(SS)}$ and $U_{SC,SC_j}^{(GB)}$ potentials ($r_{ij}^{(SS)}$ and $r_{ij}^{(GB)}$, respectively), and the values of these energy terms at the minima ($\epsilon_{ij}^{(SS)}$ and $w_{SC}\epsilon_{ij}^{(GB)}$) are easily obtained by differentiation of the respective defining functions, eqs 2 and 3. The distance R_{ij} corresponding to the maximum of the transition barrier is chosen arbitrarily to be the midpoint between the two minima.

The potential function $U_{SC,SC_j}^{(SS-GB)}$ contains only two adjustable parameters, $\Delta\epsilon$ and ΔH , which control the depth of the disulfide-bond potential well and the height of the barrier, respectively:

$$\epsilon_0^{(SS)} = w_{SC}\epsilon_{ij}^{(GB)} - \Delta\epsilon \quad (5)$$

$$H_t = w_{SC}\epsilon_{ij}^{(GB)} + \Delta H \quad (6)$$

The function $g(x;a,b)$ was chosen as a sigmoid function between the limits a and b and must have the following properties to define a smooth barrier:

$$g(x = a; a, b) = 0 \quad (7a)$$

$$g(x = b; a, b) = 1 \quad (7b)$$

$$\left. \frac{\partial g}{\partial x} \right|_{x=a} = 0 \quad (7c)$$

$$\left. \frac{\partial g}{\partial x} \right|_{x=b} = 0 \quad (7d)$$

The exact form of the function $g(x; a, b)$ was shown to have no significant effect on the performance of the potential function (data not shown); therefore, the following simple sigmoid form was chosen:

$$g(x; a, b) = \frac{(x-a)^2}{(b-a)^2} \left(3 - 2\frac{x-a}{b-a} \right) \quad (8)$$

To dynamically simulate the formation and breaking of disulfide bonds, the new term $U_{SC,SC_j}^{(SS-GB)}$ is inserted into eq 1 to replace the corresponding $w_{SC}U_{SC,SC_j}$, for all possible pairs of cysteines. Formation or breaking of a disulfide bond simply corresponds to crossing the barrier at $r_{ij} = R_{ij}$, and occurs naturally during the simulation. The thermodynamic properties of the equilibrium state are expected to be influenced mainly by the parameter $\Delta\epsilon$, which determines the strength of a disulfide bond and will therefore also affect the rate of bond breaking. From our previous work³¹ and an analysis of Doig and Williams,⁵⁶ we estimate $\Delta\epsilon \approx 3.5$ kcal/mol. The other adjustable parameter (ΔH) determines the rate of disulfide-bond formation and breaking and will therefore strongly influence the kinetics of folding. Both in vivo and in vitro, oxidative folding is controlled by various agents that promote disulfide formation and reshuffling. However, while the folding kinetics can be strongly affected by varying experimental conditions, there is evidence that the main features of the folding mechanism are conserved.^{57,58} Since the parameters $\Delta\epsilon$ and ΔH can be thought of as mimicking experimental conditions, this suggests that varying

ΔH within a range that makes the simulations feasible (if it is too large, transitions will never occur) should not change the character of the folding pathways. These effects will be investigated in section 3.

2.3. Simulation Methodology. The kinetics of folding are studied by carrying out molecular dynamics (MD) runs at 300 K. The initial state is obtained by performing a short (100 000 steps) run at 600 K to produce an unfolded state. The simplest method to achieve canonical simulations is to employ the Berendsen thermostat.⁵⁹ This approach is not as physically accurate as Langevin dynamics with explicit friction and stochastic forces, but it was shown to produce much faster simulated folding, while still producing qualitatively accurate folding pathways, albeit on a shorter time scale.⁶⁰ These features make it ideal for the purpose of testing the new methodology for disulfide-bond formation. As in earlier work,⁶⁰ the coupling parameter (τ_T in eq 18 of ref 60) was set to 1 mtu (where 1 mtu = 48.9 fs), and the extension⁴⁷ of the velocity Verlet algorithm⁶¹ to include a variable time step was used to integrate the equations of motion, with a basic time step of 0.1 mtu.

In order to enhance sampling and investigate the temperature dependence of various properties of the equilibrium state ensemble (such as free energy and native content), multiplexing replica-exchange molecular dynamics (MREMD^{62,63}) runs are also performed. The MREMD method is based on replica-exchange molecular dynamics (REMD^{64–68}). In REMD, M canonical MD simulations are performed at different temperatures, covering the range 200–440 K. After every 20 000 steps, an exchange of temperatures is attempted between neighboring trajectories, based on the Metropolis criterion defined by eq 9

$$\Delta = [\beta_{i+1}U(\mathbf{X}_{i+1}, \beta_{i+1}) - \beta_i U(\mathbf{X}_{i+1}, \beta_i)] - [\beta_{i+1}U(\mathbf{X}_i, \beta_{i+1}) - \beta_i U(\mathbf{X}_i, \beta_i)] \quad (9)$$

where $\beta_i = 1/RT_i$, T_i is the temperature of the i th trajectory, \mathbf{X}_i represents all the variables defining the conformation of the i th trajectory (at the time of the exchange), and $U(\mathbf{X}_i, \beta_i)$ is the corresponding UNRES energy. If $\Delta \leq 0$, T_i and T_{i+1} are exchanged; otherwise, the exchange is performed with probability $\exp(-\Delta)$. MREMD builds on REMD by running several trajectories at any given temperature and attempting exchanges between all trajectories at neighboring temperatures. Multiplexing was recently shown to improve the convergence of the simulations significantly.⁶³

2.4. Data Analysis. To compute thermodynamic quantities and averages from the results of MREMD simulations, we employ the weighted histogram analysis method (WHAM),⁶⁹ a procedure which was recently adapted for use with MREMD simulations and the UNRES potential.⁵⁰ Given M simulations at different temperatures producing an ensemble of N conformations, we solve the following set of self-consistent equations for the probabilities (P_i) of all conformations and the dimensionless free energies (f_k) of all simulations

$$\omega_i = -\ln \sum_{k=1}^M \exp[-f_k + \beta_k U(\mathbf{X}_i, \beta_k)]$$

$$P_i(\beta_j) = \frac{\exp[-\beta_j U(\mathbf{X}_i, \beta_j)]}{\sum_{k=1}^M \exp[-f_k + \beta_k U(\mathbf{X}_i, \beta_k)]} = \exp[\omega_i - \beta_j U(\mathbf{X}_i, \beta_j)]$$

$$f_k = -\ln \sum_{i=1}^N P_i(\beta_k) \quad (10)$$

where ω_i can be considered as the entropy of the i th conformation. With these quantities we can compute the average of any quantity A at any temperature T (or $\beta = 1/RT$) over any subset of conformations $\{\mathbf{X}\}$

$$\langle A \rangle_{\beta, \{\mathbf{X}\}} = \frac{1}{Z(\beta, \{\mathbf{X}\})} \sum_{i \in \{\mathbf{X}\}} A_i \exp[\omega_i - \beta U(\mathbf{X}_i, \beta)] \quad (11)$$

where $Z(\beta, \{\mathbf{X}\})$ is the partition function defined by

$$Z(\beta, \{\mathbf{X}\}) = \sum_{i \in \{\mathbf{X}\}} \exp[\omega_i - \beta U(\mathbf{X}_i, \beta)] \quad (12)$$

Based on eq 12, we define the free energy

$$F(\beta, \{\mathbf{X}\}) = -\frac{1}{\beta} \ln Z(\beta, \{\mathbf{X}\}) \quad (13)$$

and the heat capacity

$$C_V(T, \{\mathbf{X}\}) = \frac{\partial}{\partial T} E(T, \{\mathbf{X}\})$$

$$E(T, \{\mathbf{X}\}) = -RT^2 \frac{\partial}{\partial T} \ln Z(T, \{\mathbf{X}\}) \quad (14)$$

The heat capacity is particularly useful in identifying the temperature at which the folding transition occurs, which corresponds to a peak in the heat capacity curve when plotted against temperature.

To identify the dominant conformations in the equilibrium state ensemble at any particular temperature, the conformations are clustered with the minimum-variance clustering algorithm,^{70,71} using the rms deviation over C^α coordinates (rmsd) as the distance measure. The rmsd cutoff in clustering is adjusted to achieve a balance between the total number of clusters and their compactness. The clusters are then ranked by their probability, which is calculated by adding up the probabilities for the individual conformations making up each cluster, obtained by solving eqs 10. For each cluster, the average rmsd is calculated using eq 11, where A_i is taken as the rmsd of the i th conformation from the native structure.

3. Results

The potential function and methodology described in the previous sections are applied to several proteins, identified by their PDB codes and described in detail in the following sections. The proteins chosen for this work are relatively simple and small but cover different disulfide-bond arrangements and fold types. They were selected primarily to test the various features of the new method, without adding unnecessary complexity. The version of UNRES used here was parametrized on a single small α -helical protein⁵⁰

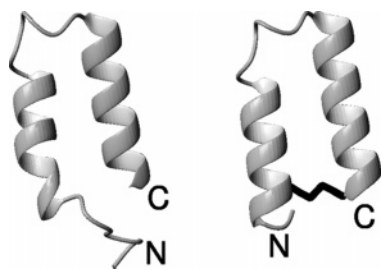


Figure 4. Native conformations⁴ of 1ZDB (left) and 1ZDD (right). The disulfide bond in the structure of 1ZDD is highlighted in black. The N- and C-termini are labeled for clarity.

(1GAB) and therefore does not in general perform well on large proteins or proteins containing β structure. As improved versions of the potential are produced (work under way), it will become possible to study more complex systems.

For each protein, several canonical MD and MREMD runs are carried out with different values for the parameters $\Delta\epsilon$ and ΔH . To assess the effect of introducing disulfide bonds, one run (labeled NOSS) is performed with the $U_{SC,SC_j}^{(GB)}$ potential instead of $U_{SC,SC_j}^{(SS-GB)}$, i.e., without allowing for disulfide-bond formation. Each MREMD simulation consists of 100–200 trajectories. The ensemble of conformations used to calculate thermodynamics properties from MREMD simulations is obtained by sampling over the last 8 000 000 steps of each trajectory. The total length of each trajectory is 24 000 000 steps. For each canonical MD simulation, statistics are obtained from 160 trajectories, each 20 000 000 steps long (just under 100 ns).

3.1. Proteins 1ZDB and 1ZDD. Protein 1ZDB is a 38-residue fragment of the B-domain of protein A, comprising its first two helices. Several residues have been mutated to enhance its binding affinity to IgG,⁷² and the fragment was shown to fold independently to the same structure as in the full protein.⁴ Removal of the first five N-terminal residues further improves binding; we refer to this shorter fragment as 1ZDB*. According to ref 4, 1ZDD is a one-disulfide variant of 1ZDB, showing greatly enhanced thermal stability. The three sequences are

1ZDB : AVAQSFNMQQRRFYALHDPNLNNEEQRNAKIKSIRDD

1ZDB* : FNMQQRRFYALHDPNLNNEEQRNAKIKSIRDD

1ZDD : FNMQCRRFYALHDPNLNNEEQRNAKIKSIRDDC

The native conformations⁴ of proteins 1ZDB and 1ZDD are shown in Figure 4, with the disulfide bond highlighted. The unstructured N-terminal region of 1ZDB, clearly visible in the Figure, is removed to produce 1ZDB*, for which no experimental structure was determined.⁴

The heat capacity curves for 1ZDB, 1ZDB*, and 1ZDD are shown in Figure 5. Despite some differences, all curves point to a well-defined folding transition around 310 K, which justifies the choice of 300 K to study the folded state. The results shown in Table 2, obtained from the MREMD runs, confirm that natively like conformations are more prevalent for the shorter fragment, consistent with experimental

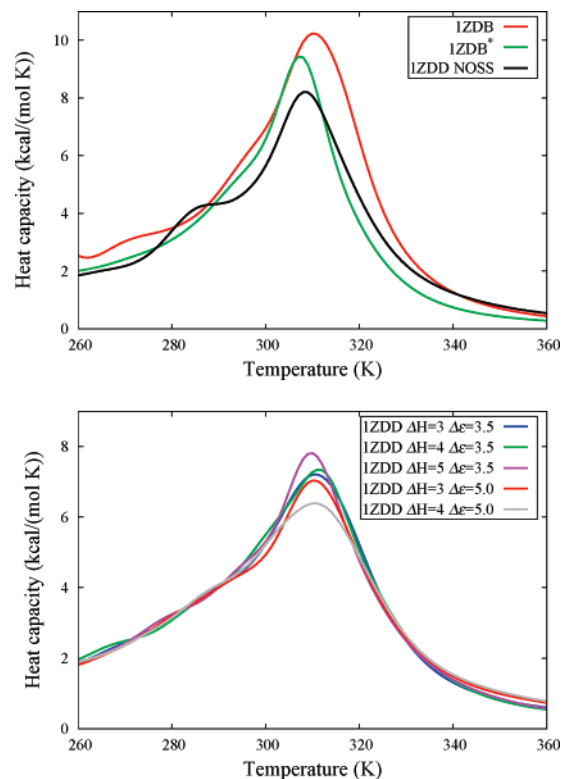


Figure 5. Temperature dependence of the heat capacity for proteins 1ZDB, 1ZDB*, and 1ZDD, with data obtained from MREMD simulations.

Table 2. Percentage of Conformations within a Given RMSD Cutoff from the Corresponding Native Structure in the Equilibrium Ensemble at 300 K^a

test system	rmsd < 4 Å	rmsd < 5 Å	rmsd < 6 Å
1ZDB	2	16	38
1ZDB*	9	39	63
1ZDD			
NOSS	25	65	89
$\Delta\epsilon = 3.5$			
$\Delta H = 3$	47	78	95
$\Delta H = 4$	49	78	95
$\Delta H = 5$	47	79	95
$\Delta\epsilon = 5.0$			
$\Delta H = 3$	57	84	97
$\Delta H = 4$	52	81	97

^a Data obtained from MREMD runs.

observations.⁷² It is also clear from the results of the NOSS run that greater prevalence of native structures is obtained with the mutation of one residue to cysteine and the addition of another cysteine at the C-terminus, even without considering the formation of the disulfide bond. However, allowing the disulfide bond to form enhances the quality of the prediction significantly, with approximately twice as many conformations found within 4 Å from the native (from 25% in the NOSS run to $\approx 50\%$ when the disulfide bond is allowed to form).

As discussed in section 2.2, broad thermodynamic properties are affected mainly by the parameter $\Delta\epsilon$ and not by ΔH . This can be seen clearly in Figure 6, which shows the temperature dependence of the average rmsd from the native structure and the average disulfide-bond content. For the

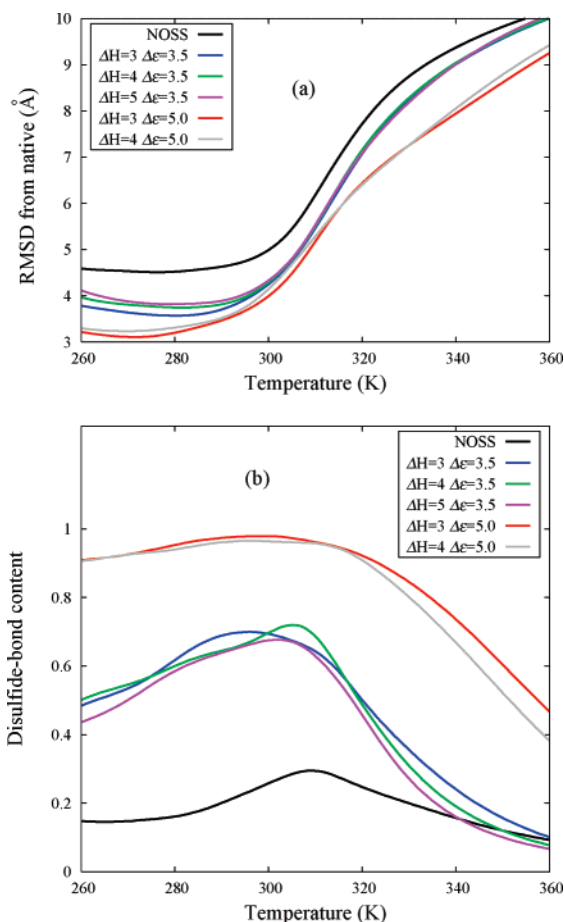


Figure 6. Temperature dependence of (a) the average rmsd from the native structure and (b) the average disulfide-bond content for protein 1ZDD, with data obtained from MREMD simulations. In the NOSS run, a disulfide bond is considered formed when the distance between Cys–Cys side-chain centroids is less than 8 Å.

purpose of calculating the disulfide-bond content in the NOSS run, a bond is considered formed when the distance between Cys–Cys side-chain centroids is less than 8 Å. The rmsd improves significantly as more disulfide bonds are present in the folded state (Figure 6), which is consistent with the data in Table 2. These results, particularly the low disulfide-bond content (less than 30%) for the NOSS run, imply that, without adding a real disulfide bond, the three-dimensional structure cannot accommodate even a half-cystine arrangement to readily form a disulfide bond, i.e., in order to form a disulfide bond, some structural modifications (with a possible introduction of strain) are necessary. For this reason, only with the higher value of $\Delta\epsilon$ is a folded state obtained, in which the bond is almost always formed.

The canonical MD simulations at 300 K show that this protein has no significant intermediates along the folding pathway. As seen in Figure 7(a), the time evolution of the fraction of folded structures (defined as those having an rmsd from the native structure below 6 Å) is almost unaffected by the introduction of the disulfide bond. In fact, the protein is already folded when the disulfide bond first appears in 70% of the cases for $\Delta H = 3$ and 80% for $\Delta H = 4$, regardless of $\Delta\epsilon$ (not shown here). The absence of an effect of ΔH is seen more clearly in Figure 7(b), which shows the

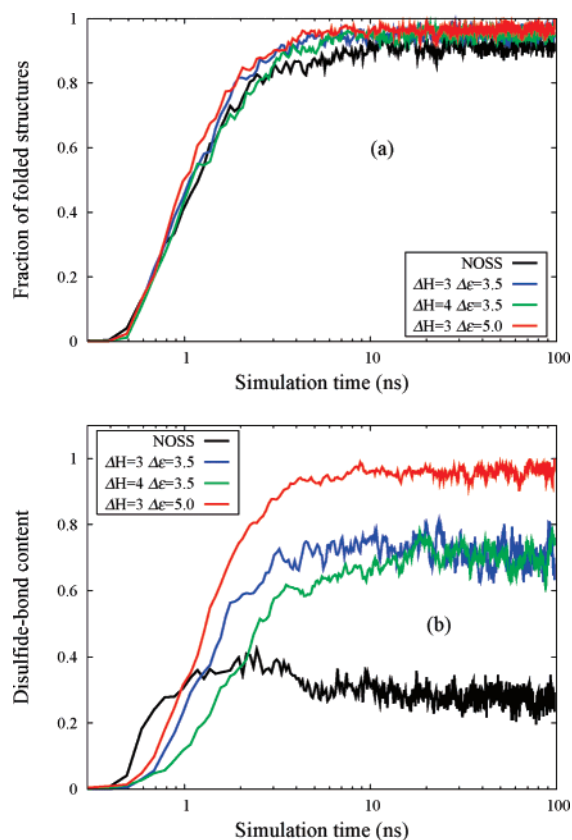


Figure 7. Time evolution of (a) the fraction of folded structures (rmsd from the native structure below 6 Å) and (b) the average disulfide-bond content in canonical MD simulations at 300 K for protein 1ZDD.

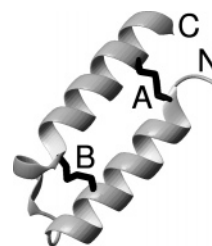


Figure 8. Native conformation⁷³ of 1E10. The two disulfide bonds are highlighted in black and labeled A and B, and the N- and C-termini are labeled for clarity.

time evolution of the average disulfide-bond content. As expected, the two curves corresponding to $\Delta\epsilon = 3.5$ converge to the same value, but disulfide-bond formation is faster for lower ΔH (i.e., lower barrier), even though the final equilibrium value is independent of ΔH . It should be noted that these final values are also consistent with the results obtained from the MREMD simulations shown in Figure 6(b).

3.2. Protein 1E10. 1E10, shown in Figure 8, is a 38-residue protein with a fold very similar to that of 1ZDD. However, its structure⁷³ is stabilized by two disulfide bonds, between residues 3-34 (A) and 13-24 (B). The presence of four cysteine residues introduces the possibility of forming non-native disulfide bonds during folding. Experimental information^{73,74} suggests that correct positioning of the helices precedes the formation of the disulfide bonds, which are known to form spontaneously. It was also shown that the

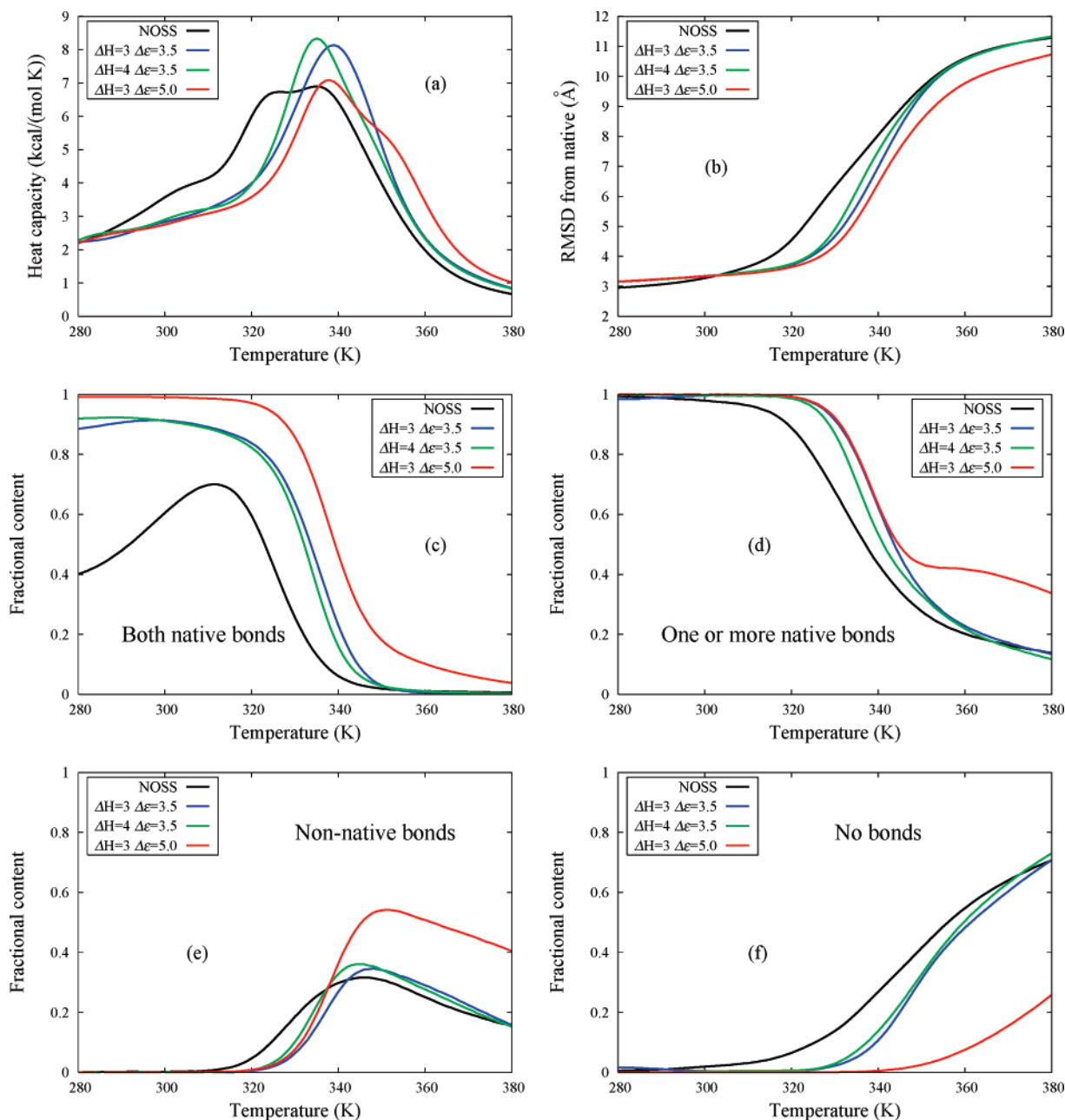


Figure 9. Temperature dependence of (a) the heat capacity and (b) the average rmsd from the native structure. Temperature dependence of the fraction of conformations with (c) both native disulfide bonds, (d) at least one native disulfide bond, (e) non-native disulfide bonds, and (f) no disulfide bonds. All plots refer to protein 1E10, with data obtained from MREMD simulations.

corresponding peptide devoid of disulfide bonds is less stable but still has a high propensity to adopt a helical conformation.⁷³

Figure 9 shows the results obtained with MREMD simulations. The folding transition occurs around 330–340 K, which is somewhat higher than for 1ZDD, and is sharper when disulfide-bond formation is allowed, as seen in Figure 9(a). Figure 9(b) shows that the results are in very good agreement with the experimental structure for all simulations (even NOSS), with average rmsd from the native structure just over 3 Å at 300 K. This protein has been studied previously with the UNRES force field, coupled with the CSA search method and dynamic disulfide-bond formation.³¹ It was found in that study that the lowest-energy structure had only disulfide bond A of Figure 8 formed, even when

using force field parameters designed especially for this protein. Figure 9(c),(d), on the other hand, shows that both native disulfide bonds are present in the majority of the conformations at 300 K, and virtually all conformations have at least one native bond [generally bond A (not shown here), as in the previous study³¹], even for the NOSS simulation. By contrast to 1ZDD, the introduction of disulfide bonds does not alter the conformations in the folded ensemble appreciably; the disulfide bonds can be formed with little strain and therefore lead to the formation of a high population of species with native disulfide bonds. While native disulfide bonds dominate the folded state (below the transition temperature), a significant number of structures containing non-native disulfide bonds (or no disulfide bonds at all) is found at higher temperatures [Figure 9(e),(f)]. It should be

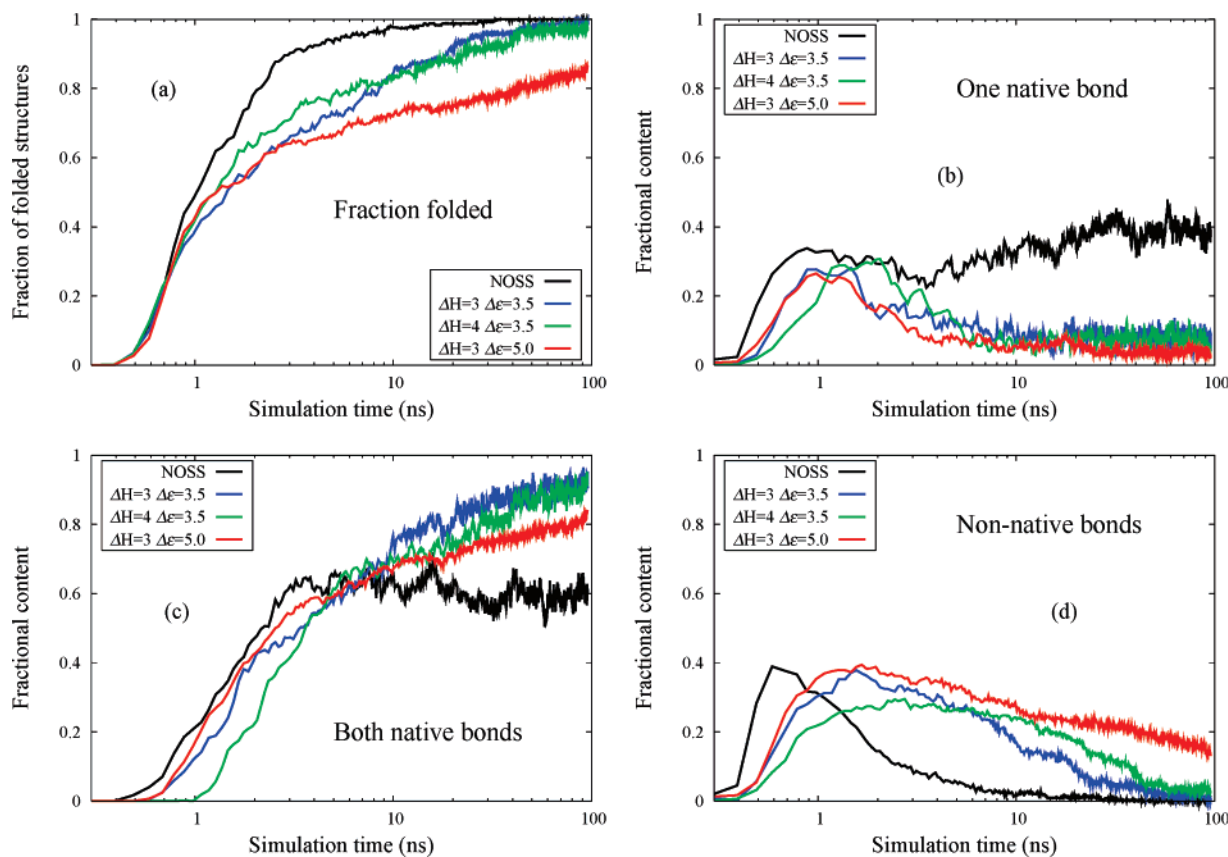


Figure 10. (a) Time evolution of the fraction of folded structures (rmsd from the native structure below 6 Å). Time evolution of the fraction of conformations with (b) one native disulfide bond, (c) both native disulfide bonds, and (d) non-native disulfide bonds. All plots refer to protein 1E10, with data obtained from canonical MD simulations at 300 K.

noted that these conformations do not form any families with similar structure but rather a diverse collection of unfolded structures containing all possible disulfide-bond arrangements. As noted previously, the parameter ΔH has little effect on thermodynamic averages. On the other hand, the higher value of $\Delta\epsilon$ results in greater prevalence of disulfide bonds (native or otherwise) at higher temperatures.

Figure 10 shows the results obtained with canonical MD simulations at 300 K. By contrast to the results obtained for 1ZDD and shown in Figure 7(a), it is clear from Figure 10(a) that folding is significantly slowed by the introduction of disulfide bonds. In fact, 100 ns were not sufficient to reach convergence for the run with $\Delta\epsilon = 5.0$. Disulfide bond B of Figure 8 generally forms before disulfide bond A (60–70% of the time, not shown here), but conformations with one native disulfide bond do not accumulate [Figure 10(b)] because the second disulfide bond forms readily [Figure 10(c)]. Conformational folding in large part precedes the formation of fully-bonded structures, as seen from a comparison of Figure 10(a),(c) and from Table 3, which shows the percentage of structures that are already folded when various disulfide-bond arrangements first appear. The reason for the slower folding is the accumulation of structures containing non-native disulfide bonds [Figure 10(d)]. The NOSS run also visits these regions of conformational space, but only the presence of disulfide bonds turns them into kinetic traps, highlighting the downside of a large $\Delta\epsilon$. As mentioned in the previous paragraph, these structures do not represent a well-defined intermediate but rather a collection

Table 3. Percentage of Structures That Are Already Folded (rmsd from the Native Structure below 6 Å) When the Given Disulfide-Bond Arrangement (A, B, or A+B) First Forms in Canonical MD Runs at 300 K for Protein 1E10

test system	A	B	A+B
NOSS	36	4	75
$\Delta\epsilon = 3.5, \Delta H = 3$	65	36	97
$\Delta\epsilon = 5.0, \Delta H = 3$	55	32	95
$\Delta\epsilon = 3.5, \Delta H = 4$	85	58	97

of misfolded conformations containing all possible single-bond arrangements (conformations with two non-native disulfide bonds never appeared).

3.3. One-Disulfide Variant of 1GAB. The current force field can produce nativelike conformations for both 1ZDD and 1E10, even without the introduction of disulfide bonds. While this is often the case for disulfide-stabilized proteins, there are examples of proteins for which the disulfide bonds are necessary for folding.¹ However, these proteins often contain β structure and are too complex to treat at present. Instead, two substitutions are made in the sequence of protein 1GAB, which was used to parametrize the force field used for this work.⁵⁰ 1GAB is a 47-residue, three-helix bundle.⁷⁵ In the new sequence, labeled 1GAB*, residues 9 and 26 (both Ala) were replaced by Cys. These sites were chosen because their side chains are not close in the native conformation, but they are close enough to form a disulfide bond in its mirror image; in our simulations here of 1GAB without

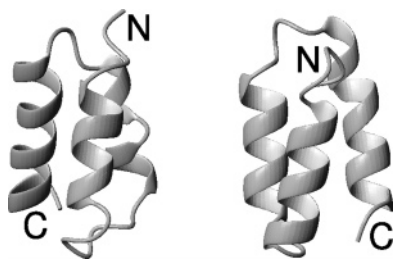


Figure 11. Native⁷⁵ (left) and mirror image (right) conformations of 1GAB. The N- and C-termini are labeled for clarity.

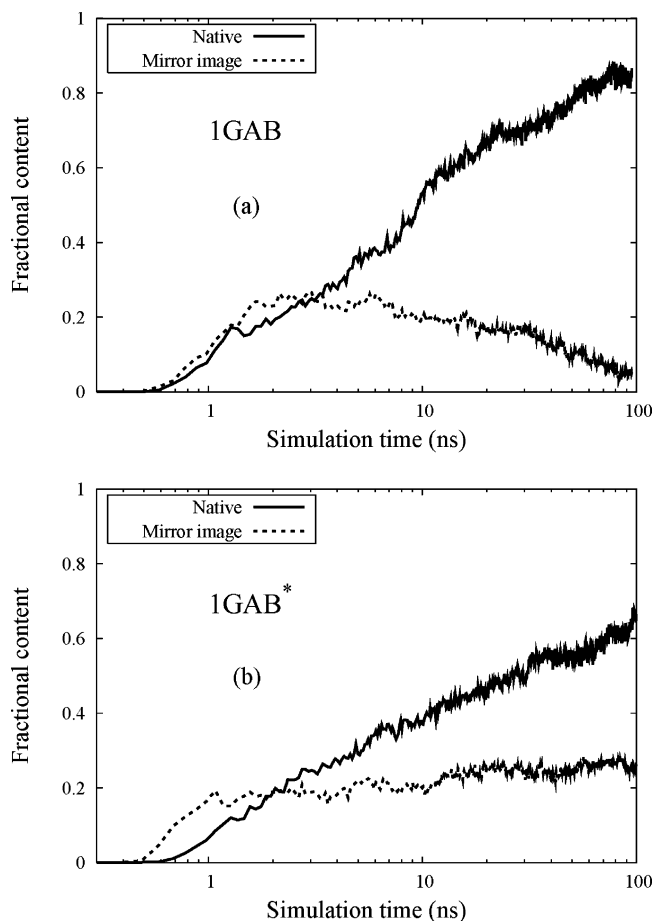


Figure 12. Time evolution of the native and mirror image populations in canonical MD simulations at 300 K for (a) 1GAB and (b) 1GAB* (with $\Delta H = 4$ and $\Delta \epsilon = 3.5$).

disulfide bonds, the mirror image appears as an intermediate during folding but is present only in less than 5% of the conformations in the equilibrium ensemble at 300 K, which is mainly nativelylike. NOSS simulations of 1GAB* produce the same results, i.e., the effect of the introduction of Cys residues is negligible unless disulfide bonds are allowed to form. The native and mirror-image conformations are shown in Figure 11, while Figure 12(a),(b) shows the time evolution of the fraction of nativelylike and mirror-image conformations (defined, as before, as those with rmsd within 6 Å from the corresponding native or mirror-image structure, respectively). In 1GAB, the mirror-image population grows as fast as the native population for the first 2 ns and subsequently decays slowly and almost disappears. In 1GAB*, on the other hand, the mirror-image population initially grows even faster than the nativelylike population, but it never decays and in fact

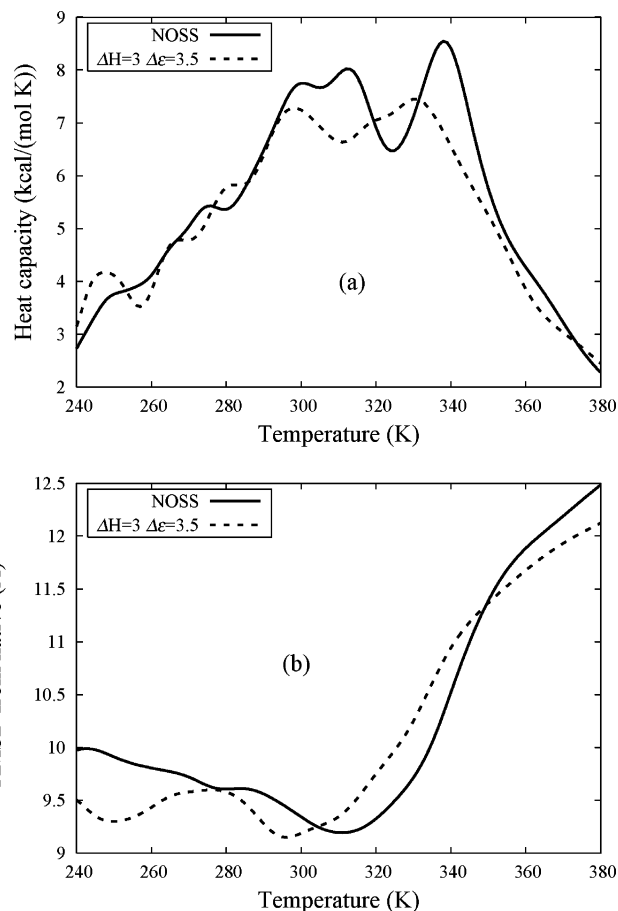


Figure 13. Temperature dependence of (a) the heat capacity and (b) the average rmsd from the native structure for protein 1NKL, with data obtained from MREMD simulations.

represents approximately 25% of the final state. Similar results for the final populations were obtained from MREMD simulations. Although we did not completely reverse the relative stabilities of the two main forms (native and mirror image), this result shows that it is possible to significantly alter the composition of the equilibrium state and not just increase its thermal stability as for 1ZDD and 1E10.

3.4. Protein 1NKL. As a final test case, a larger and more complex protein was chosen, for which the current force field produces poorer predictions. 1NKL is a 78-residue four-helix bundle containing three disulfide bonds,⁷⁶ two of which staple the N-terminus to the C-terminus (4-76 and 7-70), while the third is in the middle of the sequence (35-45). The results of MREMD simulations are shown in Figure 13. In this case, the folding transition is not well defined, and the average rmsd of the predictions is very poor. The simulations were carried out for 200 ns (almost twice as long as for the other proteins) to ensure that lack of convergence was not causing the poor results. Although most predictions are far from native, the equilibrium state at 280 K (below the irregular transition region) does include a small fraction of nativelylike conformations, as shown in Figure 13. These are present in both the NOSS run and in the run with dynamic disulfide-bond formation with $\Delta H = 3.0$ kcal/mol and $\Delta \epsilon = 3.5$ kcal/mol (which we label DYNSS) but are clearly more likely for the DYNSS run. In fact, the NOSS run does not produce any conformations with rmsd below 4 Å. We conclude that

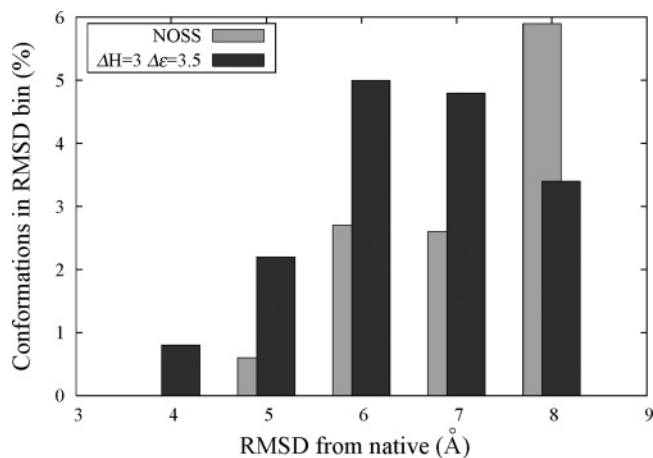


Figure 14. Percentage of conformations vs rmsd from the native structure for protein 1NKL, with data obtained from MREMD simulations.

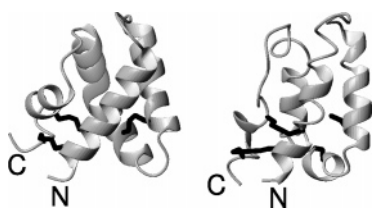


Figure 15. Native conformation⁷⁶ (left) and representative of the best family of predictions (right) for protein 1NKL. The disulfide bonds are highlighted in black, and the N- and C-termini are labeled for clarity.

the disulfide bonds again improve the quality of the predictions and the relative abundance of nativelike conformations but are not sufficient to turn them into the most probable family. When clustering is applied to the results of the MREMD simulations, a fragmented ensemble is revealed, reflecting the poorly defined transition. Approximately 20 families are found for each run, and, for both the NOSS and the DYNSS simulations, the third-ranking family is nativelike and represents 7–8% of the ensemble. However, in the NOSS run, this family has an average rmsd from the native structure of 6.2 Å, while disulfide bonds improve the rmsd of the corresponding DYNSS family to 5.0 Å. The native conformation and a representative from the nativelike DYNSS family (4.4 Å rmsd from the native structure) are shown in Figure 15. The predicted conformation contains the two native disulfide bonds near the termini, which appear to form much more easily than disulfide bond 35-45.

4. Conclusions

A new approach presented in this paper enables us to model dynamic formation and breaking of disulfide bonds in molecular dynamics simulations of protein folding. By using a reduced representation (UNRES), the time scale over which oxidative folding takes place becomes accessible. Disulfide-bond formation and breaking are simulated by introducing a transition barrier between the energy minima describing the disulfide bond on one side and the interaction between free half-cystine side chains on the other. To the best of our knowledge, this is the first algorithm based on physical

principles to produce entire oxidative folding pathways for proteins, starting from the sequence alone and without using prior knowledge of the disulfide-bond arrangement.

The approach is tested on several helical proteins. Two of them (1ZDD and 1EIO) have a simple two-helix fold, with either one or two disulfide bonds. The force field produces nativelike conformations in both cases, even when dynamic formation of disulfides is not considered. However, when it is considered, the disulfide bonds are correctly predicted and the stability of the structure is improved, consistent with experimental findings. For protein 1ZDD, the quality of the prediction is also significantly improved. In both cases, many conformations are already nativelike by the time the native disulfide-bond arrangement is formed, a result which supports the folded-precursor mechanism of oxidative folding. However, this is not observed in all trajectories, suggesting that the quasi-stochastic mechanism also plays a role, albeit a smaller one for these proteins. Tests on a larger, more complex protein (1NKL) containing three disulfide bonds also support the conclusion that allowing the disulfide bonds to form results in improved predictions, with greater frequency and improved quality of nativelike conformations in the equilibrium ensemble. However, the nativelike family is not the dominant one in this case, and the results indicate that improvements are still necessary in the underlying UNRES force field (work currently under way). In a different kind of test on protein 1GAB, we show that mutation of the appropriate residues to cysteine can lead to overstabilized intermediate states, in which a disulfide bond is present that cannot form in the native conformation. Similar results have been observed experimentally.⁸

One possible problem with the potential function employed here is the formation of structures with more than two cysteines within bonding distance. Nothing in the model explicitly prevents this, but the results show that it is very rare, and therefore insignificant, because it is energetically unfavorable. However, an explicit term could be added to the potential to remove this possibility if it became a problem in future applications. Much more important for future work is the development of a more transferable force field that is able to treat more complex proteins, particularly those containing β structure. Several examples of such proteins exist for which oxidative pathways are known experimentally and would provide ideal test systems. The good results obtained here also suggest that it may be possible to use a similar approach to simulate other rare events in protein folding. One example is the cis–trans isomerization of proline residues.^{77,78}

Acknowledgment. We thank Daniel Ripoll, Lovy Pradeep, and Robert Gahl for valuable suggestions and help with the selection of test proteins. This research was carried out by using the resources of our 820-processor Beowulf cluster at the Baker Laboratory of Chemistry and Chemical Biology, Cornell University, of the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center, of the National Center for Supercomputing Applications System at the University of Illinois at Urbana–Champaign, and of the Center for Computation and

Technology at Louisiana State University. All figures were prepared with the programs MOLMOL,⁷⁹ GNUPLOT,⁸⁰ and XFIG.⁸¹

References

- (1) Wedemeyer, W. J.; Welker, E.; Narayan, M.; Scheraga, H. A. *Biochemistry* **2000**, *39*, 4207–4216.
- (2) Shimotakahara, S.; Rios, C. B.; Laity, J. H.; Zimmerman, D. E.; Scheraga, H. A.; Montelione, G. T. *Biochemistry* **1997**, *36*, 6915–6929.
- (3) Laity, J. H.; Lester, C. C.; Shimotakahara, S.; Zimmerman, D. E.; Montelione, G. T.; Scheraga, H. A. *Biochemistry* **1997**, *36*, 12683–12699.
- (4) Starovasnik, M. A.; Braisted, A. C.; Wells, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10081–10085.
- (5) Creighton, T. E. *Science* **1992**, *256*, 111–114.
- (6) Rothwarf, D. M.; Li, Y.-J.; Scheraga, H. A. *Biochemistry* **1998**, *37*, 3760–3766.
- (7) Rothwarf, D. M.; Li, Y.-J.; Scheraga, H. A. *Biochemistry* **1998**, *37*, 3767–3776.
- (8) Mason, J. M.; Cliff, M. J.; Sessions, R. B.; Clarke, A. R. *J. Biol. Chem.* **2005**, *280*, 40494–40499.
- (9) Welker, E.; Wedemeyer, W. J.; Narayan, M.; Scheraga, H. A. *Biochemistry* **2001**, *40*, 9059–9064.
- (10) Boudko, S. P.; Engel, J. J. *Mol. Biol.* **2004**, *335*, 1289–1297.
- (11) Flory, P. J. *J. Am. Chem. Soc.* **1956**, *28*, 5222–5235.
- (12) Poland, D. C.; Scheraga, H. A. *Biopolymers* **1965**, *3*, 379–399.
- (13) Anfinsen, C. B.; Scheraga, H. A. *Adv. Prot. Chem.* **1975**, *29*, 205–300.
- (14) Pace, C. N.; Grimsley, G. R.; Thomson, J. A.; Barnett, B. J. *J. Biol. Chem.* **1988**, *263*, 11820–11825.
- (15) Zhou, N. E.; Kay, C. M.; Hodges, R. S. *Biochemistry* **1993**, *32*, 3178–3187.
- (16) Betz, S. F. *Protein Sci.* **1993**, *2*, 1551–1558.
- (17) Abkevich, V. I.; Shakhnovich, E. I. *J. Mol. Biol.* **2000**, *300*, 975–985.
- (18) Zavodszky, M.; Chen, C.-W.; Huang, J.-K.; Zolkiewski, M.; Wen, L.; Krishnamoorthi, R. *Protein Sci.* **2001**, *10*, 149–160.
- (19) Siadat, O. R.; Lougarre, A.; Lamouroux, L.; Ladurantie, C.; Fournier, D. *BMC Biochem.* **2006**, *7*, 12.
- (20) Regan, L.; Rockwell, A.; Wasserman, Z.; DeGrado, W. *Protein Sci.* **1994**, *3*, 2419–2427.
- (21) Rey, A.; Skolnick, J. *J. Chem. Phys.* **1994**, *100*, 2267–2276.
- (22) Wang, Y.; Goh, S. Y.; Kuczera, K. *J. Pept. Res.* **1999**, *53*, 188–200.
- (23) Qin, M.; Zhang, J.; Wang, W. *Biophys. J.* **2006**, *90*, 272–286.
- (24) Camacho, C. J.; Thirumalai, D. *Proteins* **1995**, *22*, 27–40.
- (25) Kobayashi, Y.; Sasabe, H.; Akutsu, T.; Saito, N. *Biophys. Chem.* **1992**, *44*, 113–127.
- (26) Martelli, P. L.; Fariselli, P.; Malaguti, L.; Casadio, R. *Protein Eng.* **2002**, *15*, 951–953.
- (27) O'Connor, B. D.; Yeates, T. O. *Nucleic Acids Res.* **2004**, *32*, W360–W364.
- (28) Ferrè, F.; Clote, P. *Nucleic Acids Res.* **2005**, *33*, W230–W232.
- (29) Ceroni, A.; Passerini, A.; Vullo, A.; Frascioni, P. *Nucleic Acids Res.* **2006**, *34*, W177–W181.
- (30) Cheng, J.; Saigo, H.; Baldi, P. *Proteins: Struct., Funct., Bioinformatics* **2006**, *62*, 617–629.
- (31) Czaplewski, C.; Oldziej, S.; Liwo, A.; Scheraga, H. A. *Protein Eng. Des. Sel.* **2004**, *17*, 29–36.
- (32) Liwo, A.; Khalili, M.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2362–2367.
- (33) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci.* **1993**, *2*, 1697–1714.
- (34) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci.* **1993**, *2*, 1715–1731.
- (35) Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849–873.
- (36) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Oldziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874–887.
- (37) Liwo, A.; Kaźmierkiewicz, R.; Czaplewski, C.; Groth, M.; Oldziej, S.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A. *J. Comput. Chem.* **1998**, *19*, 259–276.
- (38) Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. *J. Chem. Phys.* **2001**, *115*, 2323–2347.
- (39) Lee, J.; Ripoll, D. R.; Czaplewski, C.; Pillardy, J.; Wedemeyer, W. J.; Scheraga, H. A. *J. Phys. Chem. B* **2001**, *105*, 7291–7298.
- (40) Pillardy, J.; Czaplewski, C.; Liwo, A.; Wedemeyer, W. J.; Lee, J.; Ripoll, D. R.; Arłukowicz, P.; Oldziej, S.; Arnautova, Y. A.; Scheraga, H. A. *J. Phys. Chem. B* **2001**, *105*, 7299–7311.
- (41) Liwo, A.; Arłukowicz, P.; Czaplewski, C.; Oldziej, S.; Pillardy, J.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1937–1942.
- (42) Oldziej, S.; Kozłowska, U.; Liwo, A.; Scheraga, H. A. *J. Phys. Chem. A* **2003**, *107*, 8035–8046.
- (43) Liwo, A.; Oldziej, S.; Czaplewski, C.; Kozłowska, U.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 9421–9438.
- (44) Oldziej, S.; Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 16934–16949.
- (45) Oldziej, S.; Łagiewka, J.; Liwo, A.; Czaplewski, C.; Chinchio, M.; Nianias, M.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 16950–16959.
- (46) Oldziej, S. et al. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7547–7552.
- (47) Khalili, M.; Liwo, A.; Rakowski, F.; Grochowski, P.; Scheraga, H. A. *J. Phys. Chem. B* **2005**, *109*, 13785–13797.
- (48) Nishikawa, K.; Momany, F. A.; Scheraga, H. A. *Macromolecules* **1974**, *7*, 797–806.
- (49) Kubo, R. *J. Phys. Soc. Jpn.* **1962**, *17*, 1100–1120.
- (50) Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Oldziej, S.; Wachucik, K.; Scheraga, H. A. *J. Phys. Chem. B* **2007**, *111*, 260–285.

- (51) Gay, J. G.; Berne, B. J. *J. Chem. Phys.* **1981**, *74*, 3316–3319.
- (52) Kozłowska, U.; Liwo, A. To be submitted for publication.
- (53) Lee, J.; Scheraga, H. A.; Rackovsky, S. *J. Comput. Chem.* **1997**, *18*, 1222–1232.
- (54) Lee, J.; Liwo, A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2025–2030.
- (55) Czaplewski, C.; Liwo, A.; Pillardy, J.; Ołdziej, S.; Scheraga, H. A. *Polymer* **2004**, *45*, 677–686.
- (56) Doig, A. J.; Williams, D. H. *J. Mol. Biol.* **1991**, *217*, 389–398.
- (57) Rothwarf, D. M.; Scheraga, H. A. *Biochemistry* **1993**, *32*, 2671–2679.
- (58) Rothwarf, D. M.; Scheraga, H. A. *Biochemistry* **1993**, *32*, 2680–2689.
- (59) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (60) Khalili, M.; Liwo, A.; Jagielska, A.; Scheraga, H. A. *J. Phys. Chem. B* **2005**, *109*, 13798–13810.
- (61) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637–649.
- (62) Rhee, Y. M.; Pande, V. S. *Biophys. J.* **2003**, *84*, 775–786.
- (63) Czaplewski, C.; Kalinowski, S.; Scheraga, H. A. To be submitted for publication.
- (64) Hansmann, U. H. E.; Okamoto, Y. *Physica A (Amsterdam)* **1994**, *212*, 415–437.
- (65) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **2000**, *329*, 261–270.
- (66) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *J. Chem. Phys.* **2003**, *118*, 6664–6675.
- (67) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *J. Chem. Phys.* **2003**, *118*, 6676–6688.
- (68) Naniias, M.; Czaplewski, C.; Scheraga, H. A. *J. Chem. Theory Comput.* **2006**, *2*, 513–528.
- (69) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
- (70) Murtagh, F. *Multidimensional clustering algorithms*; Physika Verlag: Vienna, Austria, 1985.
- (71) Murtagh, F.; Heck, A. *Multivariate data analysis*; Kluwer Academic: Dordrecht, Holland, 1987.
- (72) Braisted, A. C.; Wells, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 5688–5692.
- (73) Barthe, P.; Rochette, S.; Vita, C.; Roumestand, C. *Protein Sci.* **2000**, *9*, 942–955.
- (74) Barthe, P.; Chiche, L.; Strub, M. P.; Roumestand, C. *J. Mol. Biol.* **1997**, *274*, 801–815.
- (75) Johansson, M. U.; de Chateau, M.; Wikstrom, M.; Forsen, S.; Drakenberg, T.; Bjorck, L. *J. Mol. Biol.* **1997**, *266*, 859–865.
- (76) Liepinsh, E.; Andersson, M.; Ruyschaert, J. M.; Otting, G. *Nat. Struct. Biol.* **1997**, *4*, 793–795.
- (77) Boulegue, C.; Milbradt, A. G.; Renner, C.; Moroder, L. *J. Mol. Biol.* **2006**, *358*, 846–856.
- (78) Pradeep, L.; Shin, H. C.; Scheraga, H. A. *FEBS Lett.* **2006**, *580*, 5029–5032.
- (79) Koradi, R.; Billeter, M.; Wüthrich, K. *J. Mol. Graphics* **1996**, *14*, 51–55.
- (80) <http://www.gnuplot.info>.
- (81) <http://www.xfig.org>.

CT7000842

JCTC

Journal of Chemical Theory and Computation

Theoretical Investigation of Tautomeric Equilibria for Isonicotinic Acid, 4-Pyridone, and Acetylacetone in Vacuo and in Solution

Peter I. Nagy,^{*,†} Giuliano Alagona,[‡] and Caterina Ghio^{*,‡}

Department of Medicinal and Biological Chemistry and Center for Drug Design and Development, The University of Toledo, Toledo, Ohio 43606-3390, and CNR-IPCF, Institute for Physico-Chemical Processes, Molecular Modeling Lab, Via Moruzzi 1, I-56124 Pisa, Italy

Received July 7, 2006

Abstract: Tautomeric equilibria have been theoretically calculated for isonicotinic acid (neutral and zwitterionic forms), the 4-pyridone/4-hydroxypyridine system, and the keto–enol transformation for acetylacetone in vacuo and in tetrahydrofuran, methanol, and water solvents. Solvent, basis set, and cavity model effects have been studied in the integral equation formalism for the polarizable continuum model (IEF-PCM)/B3LYP framework, as well as the effect of the procedure, CHELPG or RESP, applied in fitting atomic charges to the in-solution molecular electrostatic potential (ELPO). The in-solution optimized geometries obtained at the IEF-PCM/B3LYP/6-31G* and 6-311++G** levels differ moderately but deviate from their gas-phase counterparts. Atomic charges fitted to the in-solution ELPO show small variations in the considered solvents, as well as when the united-atom cavity model, or a model with explicit consideration of polar hydrogens and scaled Bondi radii, has been applied. In contrast, the fitting procedure considerably affects the derived charges producing more separated atomic charges when the CHELPG rather than the RESP procedure is utilized. The fitted charges increase up to 20% in absolute value when the basis set is enlarged from 6-31G* to 6-311++G** in the IEF-PCM/B3LYP calculations. The relative free energy, calculated as $\Delta G_{\text{tot}} = \Delta E_{\text{int}} + \Delta G(\text{solv}) + \Delta G_{\text{thermal}} + (\text{symmetry correction})$, in an ab initio/density functional theory (DFT) + free energy perturbation (FEP)/Monte Carlo (MC) approximation strongly depends on the accepted value for the relative internal energy, ΔE_{int} , of the tautomers. ΔE_{int} is to be calculated at the IEF-PCM/QCISD(T)/cc-pVTZ//IEF-PCM/B3LYP/6-31G* level for the isonicotinic acid tautomers for producing relative free energies in aqueous solution close to experimental values. In other solvents, for this system and for the other two tautomeric equilibria, calculation of ΔE_{int} at the IEF-PCM/B3LYP/6-31G* level produces ΔG_{tot} in agreement up to 1 kcal/mol with the experimental values. FEP/MC $\Delta G(\text{solv})$ calculations provide robust results with RESP charges derived by a fit to the in-solution ELPO generated at the IEF-PCM/B3LYP/6-31G* level. Molecular dynamics simulations pointed out that isonicotinic acid forms a dimeric zwitterion in tetrahydrofuran, in contrast to what happens in aqueous solution, and this structural peculiarity was interpreted as the reason for the failure of the ab initio/DFT + FEP/MC method in this particular solution.

I. Introduction

Explicit solvent models of in-solution physicochemical processes use intermolecular potentials for calculating a large number of atomic interaction energies in most cases nowadays. Pure ab initio or density functional theory (DFT) calculations are not feasible (mainly at a high theoretical level

and/or with large basis sets) for systems with medium-size solutes and several hundreds of solvent molecules. Such calculations are even less feasible for statistical averaging of millions of geometric arrangements for the members of the system. Although Car–Parrinello molecular dynamics simulations¹ have been performed for systems where all molecular interactions are considered at the DFT level,² the method has technical limitations at present for large systems.

A computationally more affordable method for modeling in-solution processes is based on the quantum mechanical/molecular mechanical (QM/MM) approach, where the system is divided into two regions characterized at different levels

* Corresponding author phone (419) 530-1945 (P.I.N.), +39-050-3152449 (C.G.); fax: (419) 530-1909 (P.I.N.), +39-050-3152442 (C.G.); e-mail: pnagy@utnet.utoledo.edu (P.I.N.), C.Ghio@ipcf.cnr.it (C.G.).

† The University of Toledo.

‡ Institute for Physico-Chemical Processes.

of theory.³ The region considered to be most responsible for the physicochemical process is described at a QM level. For the remainder of the system, and for considering the interactions between quantum and classical regions, MM approximations are employed. The MM methods make use of different approaches for calculating the intermolecular interaction energy. In force fields, one of the most important contributions is the electrostatic term: when atom–atom interactions are to be taken into account, atomic charges for the individual atoms or groups of atoms are needed in the different molecules. Accordingly, the simulation reliability may significantly depend on the selected set for the atomic (group) charge parameters.

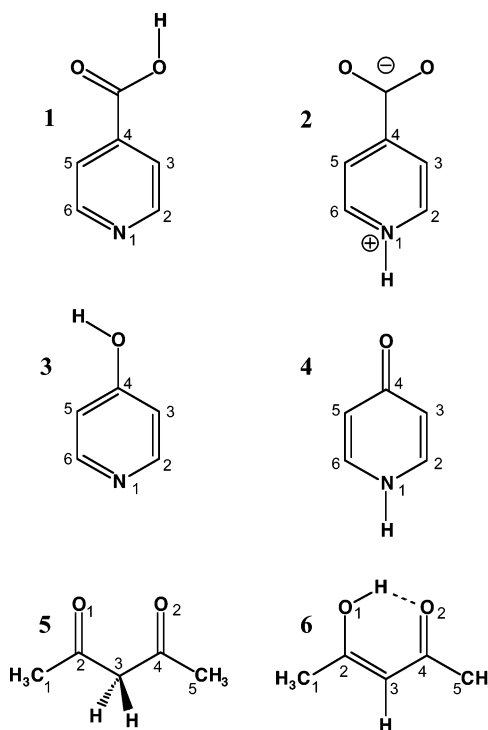
QM/MM methods are used in cases where quantum effects are decisive in some region of the system. An obvious example is the study of chemical reactions when the breaking and forming of chemical bonds occurs. A “softer” problem is investigating conformational and/or tautomeric equilibria for a solute molecule. In these processes, only the relative free energies of some particular structures are to be calculated, and further simplifications are possible throughout the modeling.

In tautomeric equilibria, in particular, new chemical bonds are formed while old ones vanish. However, we are not interested here in the reaction mechanism involved in the tautomeric effect, but rather in the relative stability of the two forms in vacuo and in solution. Quantum chemical considerations are therefore requested, although only for some specific solute structures. On the other hand, the selected quantum-chemical level has to account also for the solvent effects in these cases.

The present authors have been investigating conformational/tautomeric changes in solution for a long time.⁴ Net relative free energies were calculated as a sum of the relative internal and solvation free energies. It has become clear, however, that in order to reach accord with available experimental equilibrium constants obtained in different solvent systems, a sophisticated modeling approach is required. The mutual solute–solvent polarization modifies the electron distribution of the partners, and as a result, both the geometries and the relative internal energies change for the stable in-solution solute species as compared to their gas-phase counterparts.

When a polarizable continuum dielectric approximation is applied for a solution model, explicit structural changes can be reached for the solute. When this information is used, the net relative free energy for the conformational/tautomeric isomers of the solute can be obtained as a sum of the relative internal free energies calculated quantum chemically and solvation free energy changes obtained with some explicit solvent model. To get good values for the latter, determination of reliable solute atomic charges to be used in calculations of the solute–solvent interaction energies via pair potentials is needed. In the present investigation, the effects of different simulation parameters on the calculated free-energy changes for tautomeric transformations of isonicotinic acid, 4-OH-pyridine, and acetylacetone in tetrahydrofuran, methanol, and aqueous solution are studied. Net atomic charges derived for a solute molecule immersed in a

Chart 1. Numbering for the Isonicotinic Acid Pair, 4-OH-Pyridine/4-Pyridone, and Acetylacetone



polarizable continuum dielectric are compared. As variable modeling parameters, the cavity formation procedure, the quantum mechanical level as well as the basis set, and the methods employed for fitting net atomic charges to the in-solution molecular electrostatic potential were also considered. In Monte Carlo (MC) simulations with explicit solvent molecules, the effect of the applied charge sets and the approximations used for calculating the long-range electrostatic interactions have been studied as a byproduct. The relative total free-energy differences calculated for the tautomeric pairs have been compared with those available from experimental equilibrium studies. In a number of cases, molecular dynamics (MD) simulations have been carried out in different solvents as well.

II. Methods and Calculations

Investigated structures are shown in Chart 1. Geometries of the neutral (1) and zwitterionic (2) tautomers for isonicotinic acid, 4-hydroxypyridine (3), 4-pyridone (4), and the diketo (5) and keto–enol (6) forms of acetylacetone were optimized in the gas phase as well as in tetrahydrofuran (THF), water, and methanol solvents at the DFT/B3LYP⁵ (the Becke gradient-corrected three-parameter hybrid exchange and Lee–Yang–Parr correlation functionals) and integral equation formalism for the polarizable continuum model (IEF-PCM)⁶/B3LYP levels, respectively, using the 6-31G* and the 6-311++G** basis sets.⁷ Calculations were performed with the aid of the Gaussian 03 software⁸ and applying default dielectric constants with the IEF-PCM.

Two sets of atomic radii (reported in Table 1) were applied in forming cavities in the solvents. When the united–atom standard set, UA0 (with a scaling factor of 1.0) was used,

Table 1. Atomic Cavity Radii

	UAO ($\alpha = 1.0$)	Bondi ($\alpha = 1.2$)	
	R	R	$R \times \alpha$
CH	2.125	1.90	2.280
CH ₂	2.325	2.00	2.400
CH ₃	2.525	2.00	2.400
C	1.925	1.70	2.040
=O	1.750	1.52	1.824
-OH	1.850		
N	1.830	1.55	1.860
NH	1.930		
O ⁻	1.750	1.52	1.824
-O		1.52	1.824
H		1.20	1.440

the CH, CH₂, and CH₃ as well as the OH and NH groups acted as single-sphere centers. The final cavity was formed by the union of the overlapping spheres around the atomic centers in the molecules. When Bondi radii⁹ were used, a united-atom model was still maintained for the CH_n groups; for polar hydrogen atoms (i.e., those linked to O and N), separate centers in the cavity formation were considered, and a scaling factor of 1.2 was employed throughout.

Geometries corresponding to local energy minima were identified by all positive vibrational frequencies. Single-point energy calculations for these structures were performed at the IEF-PCM/B3LYP/6-311++G** level as well as at the IEF-PCM/QCISD(T)¹⁰ (quadratic configuration interaction with single and double excitations and perturbative triple excitations) and CCSD(T)¹¹ (coupled-cluster with excitations as above) levels with the 6-31G*, 6-311++G**, and cc-pVTZ sets.¹² Relative internal energies with corrections up to the QCISD(T) level were obtained as

$$\Delta E_{\text{int}}(\text{corr}) = \Delta \langle \psi | H | \psi \rangle + \{ \Delta E[\text{IEF-PCM/QCISD(T)}] - \Delta \langle \psi | H + 1/2V | \psi \rangle \} \quad (1)$$

where H and V are the Hamiltonian and the reaction field operators, respectively, and ψ is the converged wavefunction in solution (where the solvent is treated at the HF/SCF level). $\Delta E[\text{IEF-PCM/QCISD(T)}]$ includes the relative electrostatic solute-solvent interaction energy of $\Delta E_{\text{elst}} = \langle \psi | 1/2V | \psi \rangle$ as well; thus, the $\Delta E[\text{IEF-PCM/QCISD(T)}] - \Delta \langle \psi | H + 1/2V | \psi \rangle$ term accounts for the high-level internal energy correction. CCSD(T) corrections were obtained by using eq 1 and applying the corresponding IEF-PCM/CCSD(T) values.

Thermal corrections at the IEF-PCM/B3LYP/6-31G* level for obtaining relative internal free energies for the species were calculated in the rigid rotator, harmonic oscillator approximation¹³ as

$$\Delta G_{\text{therm}}(T) = \Delta \text{ZPE} + \Delta [H_{\text{vibr}}(T) - \text{ZPE}] - T\Delta S(T) \quad (2)$$

where ZPE, $H_{\text{vibr}}(T)$, and $S(T)$ stand for the zero-point energy, the vibrational enthalpy, and the total entropy, respectively, at $T = 298$ K and $p = 1$ atm. The total relative free energy of the tautomers was calculated at the PCM level as

$$\Delta G_{\text{tot}} = \Delta E_{\text{int}} + \Delta E_{\text{elst}} + \Delta G_{\text{drc}} + \Delta G_{\text{therm}}(T) \quad (3)$$

Table 2. 12-6-1 Potential Parameters for the Pure Liquid THF in the Twist Conformation

	σ^a (Å)	ϵ^b (kcal)	q (atomic unit)	density (g/cm ³)	heat of vaporization (kcal/mol)
O	2.900	0.140	-0.4194		
C _α	3.450	0.070	0.0879		
C _β	3.450	0.070	-0.0212		
H _{αeq}	2.450	0.036	0.0400		
H _{βax}	2.450	0.036	0.0464		
H _{βeq}	2.450	0.036	0.0482		
H _{βax}	2.450	0.036	0.0084		
calcd				0.886	7.70
exp ^c				0.889	7.65

^a The σ parameter corresponds to the atom separation when the 12-6 Lennard-Jones (LJ) potential has the zero value. ^b ϵ provides the negative of the minimum energy of the 12-6 LJ potential. ^c From ref 20.

where ΔG_{drc} is the relative nonelectrostatic free-energy term accounting for solute-solvent dispersion and repulsion energies, and for the cavity formation.

Relative solvation free energies were obtained by using the free-energy perturbation method (FEP)¹⁴ as implemented in Monte Carlo simulations.¹⁵ Calculations were carried out by the use of the BOSS 4.7 software.¹⁶

MC simulations were performed in NpT (isobaric-isothermal) ensembles at $T = 298$ K and $p = 1$ atm.¹⁷ Water boxes including 503-505 TIP4P water molecules¹⁸ and a single solute were considered for the aqueous solution model. The solution models with THF and methanol solvents were comprised of 262-264 solvent molecules, using three- and five-point models for methanol^{19a} and THF,^{19b} respectively, and a single solute. For THF, an all-atom model was also developed. Three different conformations for a THF solute were optimized at the IEF-PCM/B3LYP/6-31G* level in a THF environment, and a C₂ twisted structure was found of lowest energy. By determining the atomic charges upon the RESP fit (see below) and slightly modifying the OPLS 12-6 steric parameters, the calculated density for the solvent box comprised of 267 all-atom THF molecules as well as the heat of vaporization were in good agreement with the experimental values (Table 2). The heat of vaporization (HV) was calculated as

$$\text{HV} = (E_{\text{gas}}^{\text{int}} + RT) - [E_{\text{liq}}^{\text{int}} + E(\text{MC})/267 + pV_{\text{liq}}] \quad (4)$$

where $E_{\text{gas}}^{\text{int}} - E_{\text{liq}}^{\text{int}}$ is the change of the molar internal energy as calculated in the gas phase and in a THF environment (using the IEF-PCM) at the QCISD(T)/cc-pVTZ//B3LYP/6-31G* level. $E(\text{MC})$ is the MC energy for the liquid model comprised of 267 THF molecules, and V_{liq} is the THF molar volume.

Nonphysical reaction paths were selected in the FEP calculations when the tautomeric species 1-2, 4-3, and 5-6 were transformed. A nonphysical path means a gradual and contemporaneous annihilation and development of the proton involved in the tautomeric transformation at the proper sites. The reaction-coordinate parameter, λ , has values of 0 and 1 referring to the chemically correct starting and final structures

of the tautomers. When double-wide sampling¹⁵ was used in FEP calculations, $\Delta\lambda$ was selected in the 0.0125–0.05 range in order to keep the free-energy increments at about 1 kcal/mol or less. Geometric and interaction potential parameters along the transformation path were calculated by linear interpolation between the corresponding starting and end values with a λ coupling parameter.¹⁵ The optimized geometries for the tautomers were obtained from IEF-PCM/B3LYP/6-31G* calculations.

Interaction energies of the solution elements were calculated by using the 12–6–1-type OPLS-AA pair potential.²¹ For the 12–6 steric OPLS parameters, the all-atom literature values were accepted.²¹ The solvent–solvent cutoff (RCUT) and the solute–solvent cutoff (SCUT) were set to 9.75–12.0 Å and 12.0 Å, respectively. Random translation and rotation for the solute were limited to 0.1 Å and 10°, respectively. Solute trial moves were attempted every 50 steps, while volume alteration (with a maximum of 250 Å³) was attempted every 1000 steps. Periodic boundary conditions and preferential sampling were applied with $c = 120$ in the sampling factor, $1/(R^2 + c)$, where R is the distance between the solute's reference atom and the central atom of the selected solvent molecule. With these simulation parameters, 40–60% of the newly generated configurations were accepted out of 3500 K and 5000 K configurations considered in the equilibrium and averaging phases, respectively.

The charge parameters were determined in the present study. A total of 36 ($3 \times 2 \times 2 \times 3$) charge sets were derived for each form of the three tautomeric pairs, considering three solvents (water, methanol, and THF), two cavity models (UA0 and Bondi), two fitting procedures (CHELPG²² and RESP²³) of the atomic charges to the in-solution molecular electrostatic potential (ELPO), and three in-solution wave functions for calculating the ELPO. Further charge sets were derived for structures **1** and **2** in vacuo for comparison. The wavefunctions were obtained either for the species optimized in the given solvent at the IEF-PCM/B3LYP/6-31G* and the IEF-PCM/B3LYP/6-311++G** levels or from IEF-PCM/B3LYP/6-311++G**//IEF-PCM/B3LYP/6-31G* single-point calculations (optimized geometries in vacuo and in solution are reported in Tables S1–S3 of the Supporting Information).

Long-range electrostatic effects (LRE) were obtained at the IEF-PCM/B3LYP level using the corresponding basis set in cases when the FEP calculations were performed with RCUT = 9.75 Å. Upon the ICUT = 2 option in the BOSS program, every solvent molecule is seen by the solute if the central atom of the solvent is within a sphere of $R = \text{SCUT}$ around any solute atoms. Accordingly, the IEF-PCM energy providing the solute–solvent interaction energy out of the SCUT-defined volume was calculated for the tautomers with a cavity formed by interlocking spheres around the solute atoms with $R = \text{SCUT} = 12$ Å. In several simulations, both RCUT and SCUT were set to 12.0 Å, and the LRE was considered throughout the Ewald summation^{24a} or by applying a reaction field (RF).²⁵ For the RF calculations, dielectric constants of 7.43, 32.63, and 78.39, relevant at $T = 298$ K, were applied in THF, methanol, and water, respectively.

The total relative free energy was calculated as

$$\Delta G_{\text{tot}} = \Delta E_{\text{int}} + \Delta G(\text{solv}) + \Delta G_{\text{therm}} \quad (5)$$

where ΔE_{int} and ΔG_{therm} are from IEF-PCM calculations and $\Delta G(\text{solv})$ is the relative solvation free energy for the tautomers, as calculated by the FEP/MC method.

Constant pressure ($p = 1$ atm) and temperature ($T = 298$ K) MD simulations for 2 ns have been carried out on dimers of isonicotinic acid and its zwitterion in THF using AMBER9.²⁶ The THF box from the pure liquid MC simulation (see above) has been used to solvate both dimers applying periodic boundary conditions, a 12 Å cutoff, and the particle-mesh Ewald method^{24b–e} to treat LRE. Both for the THF solvent and each individual solute molecule, RESP charges obtained for the IEF-PCM/B3LYP/6-31G* optimized geometries in THF using Bondi radii have been employed. A preliminary minimization (80–90 K steps) has been carried out, followed by six constant volume MD runs (using SHAKE^{24f} for bonds involving H, 2 fs time-step) of 20 ps each, raising the temperature in 50 K increments until a temperature of 298 K has been reached. Then, a constant pressure equilibration (1 atm, for further 200 ps) has been carried out to approach the experimental density of 0.889 g/cm³ for the pure THF. MD simulations in water (TIP4P) have been carried out following analogous equilibration and production phases.

III. Results and Discussion

A. Continuum Solvent Calculations. The equilibrium composition is a sensitive function of the chemical environment for many tautomeric systems. The relative standard free energy of the involved species, related to the K equilibrium constant as $-RT \ln K = \Delta G^\circ$, changes not only through the solvation from the gas phase, but it generally varies in different solvents too.^{27,28} Although continuum methods are capable of accounting for the energetic aspects of the tautomeric processes, they are inherently unable to predict the solvent structure around the solute. Both goals, thermodynamic and solution-structure characterizations of the tautomeric system, can be achieved, however, by a combined application of the ab initio/DFT continuum method and MC simulations considering explicit solvent molecules. The final relative free energies can be calculated from eq 5.

In the present MC simulations, the atom–atom interactions are calculated by pair potentials with predefined parameters. The atomic charges are the least transferable parameters, which change with the composition of the molecule including the specific atom and are affected by the chemical environment.

Coupling continuum and MC calculations, it is straightforward to resort to solvent-dependent charges. When the gas-phase HF/6-31G* charges are used, the calculated gas-phase dipole moment should be systematically overestimated, but it is likely that this value would not correctly account for the polarization both in low dielectric constant solvents and water at a time. Since the optimized geometry also changes going from the gas-phase into different solvents (torsional angles may be especially sensitive to the environment), the electron distribution also would follow these

changes. Thus, the gas-phase charges would not apply for all different solvents in principle. Neither is the use of IEF-PCM with a specific ϵ particularly appealing: this method could be advisable for obtaining charges to be used in simulations inside proteins or nucleic acids,²⁹ but not for dilute solutions with a wide variety of solvents. There are several other distinct methods in the literature that basically determine the atomic charges for the atom in a specific functional group or chemical environment. It is, however, outside the aim of this study to review them. Nonetheless, it is worth mentioning at least the CMx methods,³⁰ because their semiempirical rapid procedures are incorporated directly in BOSS³¹ and have been subjected to extensive validations in solution, including the tautomeric equilibrium for 2-hydroxypyridine and 2-pyridone,³² and for computations of absolute free energy of hydration with the TIP4P water model.³³ In our investigations, however, besides the reason put forward below, we prefer to make use of ab initio or DFT-derived ELPO charges because those fittings (only a very small fraction of the total computational time) originated from our earlier (either atom-centered or not) partial charge models in vacuo.^{3a–b,34}

By the application of the PCM method, however, chemical system and solvent-dependent solute charges can be derived. In mutual modifications of the solute structure and the polarization state of the dielectric medium in PCM studies, a solute structure corresponding to a local energy minimum can be obtained if the electrostatic solute–solvent interaction is considered throughout the optimization. Because of the different solute–solvent interactions in various solvents, the optimized solute structure, and consequently the relative internal energy for the elements of a tautomeric pair, also show solvent dependence in general.^{27d}

The derivation process for the atomic charges is far from being unique: even employing a single method, obtained atom-centered charges may significantly differ. In many cases, moreover, possible intramolecular hydrogen bond(s) or changes in the molecular conformation also have a non-negligible effect on the atomic charges. In addition, as mentioned above, these charges change more or less in different solvents.

If the relative internal energy has been calculated on the basis of molecular structures as obtained from a continuum approximation, a procedure for calculating the solute–solvent interaction energy/free energy in MC is consistent when the charge distribution of the solute is represented as in the continuum solvent method. These net atomic values have to mimic the overall charge distribution of the solute molecule in the given solvent. For calculating tautomeric equilibria by the FEP method, consideration of only two involved tautomeric forms suffices in general, but for determination of a rotational potential, charge derivations for a large number of conformers are needed.³⁵ Our goal in the present paper is to study three tautomeric systems, taking into account the effects of the applied internal parameters of the IEF-PCM method, of the basis set, and of the charge fitting procedure used to derive the in-solution relevant atomic charges on the calculated relative free energies. Results will be compared with available experimental values.

The effect of the chemical environment on the optimized structure may be assessed comparing geometric parameters from gas-phase and in-solution optimizations (Tables S1–S3, Supporting Information). Changes in the bond lengths and angles are generally monotonic in the gas phase, THF, methanol, and water-solvent series. Bond lengths may differ by up to 0.02 Å in the gas phase and in solution. For example, we found a remarkable solvent effect for the C–O distance in **1** and N–H and C₄–C_{carb} variation for **2** at the IEF-PCM/B3LYP/6-31G* level when the UA0 cavity was formed (Table S1, Supporting Information). The solvent effect is 4–5° for the OCO angle in **2** with any cavity method and with both basis sets.

The C₄–O bond of 4-pyridone changes generally by up to 0.02 Å upon solvation (Table S2, Supporting Information), but the difference is even larger in water and with the 6-311++G** basis set. Table S3 (Supporting Information) indicates a very large solvent effect on the C₂C₃C₄O₂ torsion angle in the diketo form with any basis set and cavity model. The solvent effect on the O₁C₂C₃C₄ torsion angle is 10–17° with the two basis sets.

Tables 3–5 compare the relative energies/free energies for the tautomeric pairs in the gas phase and in solution at different levels of the theory. The cavity models applied in the IEF-PCM calculations are indicated. The dominating relative energy term for the isonicotinic acid tautomers is the relative internal energy (Table 3). $\Delta G_{\text{tot}} = \Delta E_{\text{int}} + \Delta E_{\text{elst}} + \Delta G_{\text{drc}}$ is always positive and would become slightly more positive if the ΔG_{therm} were added. The positive ΔG_{tot} means that the neutral form was predicted by the IEF-PCM method as the prevailing form in solution. The prevalence of this form, however, is solvent-dependent. ΔG_{tot} decreases from THF to methanol and further decreases in water. Thus, the IEF-PCM method correctly reproduces the trend of the stability of the tautomers, indicating the increasing population of the zwitterionic form in the more polar solvent. The basis set effect on the calculated values is also important. For example, ΔG_{tot} is 5.55, 0.56, and 0.55 kcal/mol in water when the UA0 cavity is used. The decrease of 5.55 to 0.56 indicates a pure basis set effect because the two values were calculated at identical, IEF-PCM/B3LYP/6-31G*-optimized geometries with 6-31G* and 6-311++G** basis sets, respectively. The geometry effect is small, as revealed by the change of ΔG_{tot} from 0.56 to 0.55 kcal/mol, using the 6-311++G** basis sets in both cases, but on the 6-31G*- and 6-311++G** optimized geometries.

The cavity model also affects the calculated energy terms. Using the Bondi instead of the UA0 cavity, the absolute values of both the ΔE_{int} and ΔE_{elst} terms decrease by, however, different amounts. Furthermore, the ΔG_{drc} term, although small in absolute value, is of different sign with the two cavity models. Overall, the Bondi ΔG_{tot} values obtained at a given level are about 2 kcal/mol more positive for the zwitterionic isonicotinic acid than the corresponding value calculated with the UA0 cavity. The largest problem with the IEF-PCM results for this tautomeric pair is that the method does not reproduce the switch of the tautomeric preference in water compared to THF and methanol. Experiments found^{28a} about 1% and 2–4% zwitterionic isonicotinic

Table 3. Energy and Free-Energy Terms (kcal/mol) for the Zwitterionic Isonicotinic Acid Relative to the Neutral Form in the Gas Phase and from IEF-PCM/B3LYP Calculations with Different Basis Sets

B3LYP		THF		methanol		water	
6-31G*	gas phase	UA0	Bondi	UA0	Bondi	UA0	Bondi
ΔE_{int}	32.45	39.36	38.05	42.30	40.38	42.96	41.03
ΔE_{elst}		-28.60	-25.78	-35.75	-31.99	-37.29	-33.59
ΔG_{drc}		-0.11	0.08	-0.09	0.08	-0.12	0.09
ΔG_{tot}^a		10.65	12.35	6.46	8.47	5.55	7.53
ΔG_{therm}		0.39	0.38	0.69	0.45	0.47	0.46
6-311++G**//6-31G*		UA0	Bondi	UA0	Bondi	UA0	Bondi
ΔE_{int}		37.03	35.82	40.27	38.46	40.99	39.18
ΔE_{elst}		-30.84	-28.20	-38.65	-35.11	-40.31	-36.86
ΔG_{tot}^b		6.08	7.70	1.53	3.43	0.56	2.41
6-311++G**		UA0	Bondi	UA0	Bondi	UA0	Bondi
ΔE_{int}	29.37	37.58	36.21	40.96	38.97	41.74	39.68
ΔE_{elst}		-31.39	-28.57	-39.36	-35.62	-41.08	-37.37
ΔG_{drc}		-0.10	0.08	-0.08	0.08	-0.11	0.10
ΔG_{tot}		6.09	7.72	1.52	3.43	0.55	2.41
ΔG_{exp}^c		2.7		2.3		-2.6	

^a $\Delta G_{\text{tot}} = \Delta E_{\text{int}} + \Delta E_{\text{elst}} + \Delta G_{\text{drc}}$. ^b ΔG_{drc} from the 6-31G* calculations. ^c Derived from experimental compositions determined in ref 28a. The estimated uncertainty in ΔG_{exp} is up to a few tenths of a kilocalorie per mole.

Table 4. Energy and Free-Energy Terms (kcal/mol) for 4-Pyridone Relative to 4-OH-Pyridine in the Gas Phase and from IEF-PCM/B3LYP Calculations with Different Basis Sets

B3LYP		THF		methanol		water	
6-31G*	gas phase	UA0	Bondi	UA0	Bondi	UA0	Bondi
ΔE_{int}	1.51	3.85	3.37	4.98	4.30	5.24	4.51
ΔE_{elst}		-6.69	-6.17	-8.62	-8.08	-9.04	-8.45
ΔG_{drc}		-0.14	0.14	-0.13	0.12	-0.18	0.15
ΔG_{tot}^a		-2.98	-2.66	-3.77	-3.66	-3.98	-3.79
ΔG_{therm}		0.45	0.39	0.51	0.45	0.53	0.45
6-311++G**//6-31G*		UA0	Bondi	UA0	Bondi	UA0	Bondi
ΔE_{int}		3.94	3.44	5.46	4.65	5.69	4.93
ΔE_{elst}		-8.12	-7.52	-10.69	-9.94	-11.13	-10.42
ΔG_{tot}^b		-4.32	-3.94	-5.36	-5.17	-5.62	-5.34
6-311++G**		UA0	Bondi	UA0	Bondi	UA0	Bondi
ΔE_{int}	1.12	4.38	3.71	6.03	5.09	6.43	5.39
ΔE_{elst}		-8.55	-7.78	-11.28	-10.37	-11.90	-10.89
ΔG_{drc}		-0.15	0.13	-0.13	0.11	-0.18	0.14
ΔG_{tot}		-4.32	-3.94	-5.38	-5.17	-5.65	-5.36
ΔG_{exp}^c	> 1.36					-4.5	

^a $\Delta G_{\text{tot}} = \Delta E_{\text{int}} + \Delta E_{\text{elst}} + \Delta G_{\text{drc}}$. ^b ΔG_{drc} from the 6-31G* calculations. ^c Ref 28b.

acid in THF and methanol, respectively, but this tautomer population is about 95% in aqueous solution. Our best theoretical estimate is $\Delta G_{\text{tot}} = 0.56$ kcal/mol compared to -2.59 kcal/mol, derived from the experiment (Table 3).

The relative internal energies for the tautomers in the gas phase compared to the solution phase show significant changes, ranging from 6 (7) to 11 (12) kcal/mol, depending on solvent permittivity, cavity radii, and the basis set (B3LYP/6-311++G** values in parentheses). The relative stability of the zwitterionic form decreases upon interaction with the solvent, because ΔE_{int} increases. The polarity of the solvent has some effect on the relative internal energies. Whereas the ΔE_{int} values are slightly closer in the gas phase and in THF, the strong interaction of the zwitterion with polar

solvents (methanol and water) leads to a distortion of the geometry (Table S1, Supporting Information) in order to make the electrostatic solute-solvent interaction energy optimally negative. The basis set effect can also be remarkable in some cases, such as for the COO group arrangement in vacuo: at the B3LYP/6-311++G** level, the carboxylate is nearly perpendicular to the ring ($\text{OCCC} = 67^\circ$), whereas at the B3LYP/6-31G* level, it is located in the ring plane, as occurs in solution as well, without noticeable differences between applied levels.

Energy results for the 4-pyridone/4-OH-pyridine tautomeric pair are summarized in Table 4. Acceptance of the UA0 versus Bondi cavity has a relatively small effect on the calculated energies, although the ΔG_{drc} values are of

Table 5. Energy and Free-Energy Terms (kcal/mol) for the Diketo Relative to the Keto–Enol Form of Acetylacetone in the Gas Phase and from IEF-PCM/B3LYP Calculations with Different Basis Sets

B3LYP	gas phase	THF		methanol		water	
		UA0	Bondi	UA0	Bondi	UA0	Bondi
6-31G*							
ΔE_{int}	3.24	4.54	4.51	5.04	5.06	5.27	5.28
ΔE_{elst}		-2.26	-2.16	-3.17	-3.14	-3.51	-3.46
ΔG_{drc}		0.24	0.50	0.26	0.48	0.33	0.64
ΔG_{tot}^a		2.52	2.85	2.12	2.41	2.09	2.46
ΔG_{therm}		-1.42	-1.51	-1.70	-1.98	-1.69	-1.73
6-311++ G**//6-31G*		UA0	Bondi	UA0	Bondi	UA0	Bondi
ΔE_{int}		6.82	6.79	7.39	7.42	7.66	7.67
ΔE_{elst}		-2.79	-2.70	-3.95	-3.99	-4.39	-4.41
ΔG_{tot}^b		4.27	4.59	3.69	3.92	3.60	3.90
6-311++ G**		UA0	Bondi	UA0	Bondi	UA0	Bondi
ΔE_{int}	5.35	7.21	7.23	8.18	8.12	8.39	8.46
ΔE_{elst}		-3.25	-3.24	-4.85	-4.78	-5.22	-5.32
ΔG_{drc}		0.32	0.56	0.34	0.55	0.43	0.74
ΔG_{tot}		4.28	4.55	3.67	3.90	3.60	3.88
ΔG_{exp}^c		0.8-1.1		0.5-0.7		-0.64	

^a $\Delta G_{\text{tot}} = \Delta E_{\text{int}} + \Delta E_{\text{elst}} + \Delta G_{\text{drc}}$. ^b ΔG_{drc} from the 6-31G* calculations. ^c Ref 28c.

different sign in the two approximations again (see Table 3). The absolute values of the ΔE_{int} and ΔE_{elst} terms are smaller with the Bondi than with the UA0 cavity by 0.5–0.7 kcal/mol, but ΔG_{tot} differs only by 0.3–0.4 kcal/mol. Experimental results are available both in the gas phase and in solution. The gas-phase experiment indicates more than 90% 4-OH-pyridine in the tautomeric mixture, whereas 4-pyridone is almost exclusively present in aqueous solution.^{28b} Our calculations predict 4-OH-pyridine as the prevailing form in the gas phase, whereas the 4-pyridone tautomer is the overwhelming fraction in all considered solutions. The calculated relative free energies show remarkable basis set dependence: $\Delta G_{\text{tot}} = -2.7$ to -3.0 and -3.9 to -4.3 kcal/mol in THF with the 6-31G* and 6-311++G** basis sets, respectively. In water, the calculated corresponding ΔG_{tot} values are -3.8 to -4.0 and -5.3 to -5.7 kcal/mol. Differences in the optimized geometries obtained with the 6-31G* and 6-311++G** basis sets have, however, negligible effect on the relative energy terms also for this tautomeric pair, as revealed by the comparison of the B3LYP/6-311++G**//B3LYP/6-31G* and B3LYP/6-311++G** results (Table 4). The consideration of the relative ΔG_{therm} values would make ΔG_{tot} less negative by 0.4–0.5 kcal/mol, still maintaining, however, the negative sign as well as a strong preference for 4-pyridone in both solvents. The geometry relaxation of the gas-phase structure is modest, especially with the Bondi cavity, in any solvent.

Concerning acetylacetone, the lowest-energy structures of its tautomers in vacuo are shown in Figure 1. While the most stable keto–enol form (6) is satisfactorily represented in Chart 1, with heavy atoms approximately in the same plane; the lowest-energy diketo form (5) significantly differs from the structure sketched in the chart both in vacuo and in solution (Figure 2). The only minimum energy structures in vacuo within 10 kcal/mol of the lowest minimum are shown in Figure 3. The second minimum (Figure 3a) is 3.5 kcal/mol less favorable than 5, while the structure (Figure 3b) of

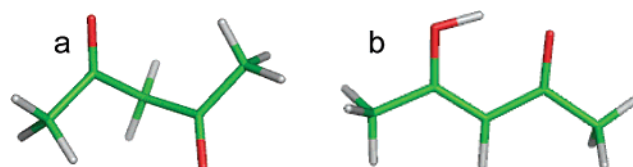


Figure 1. B3LYP/6-31G* lowest-energy structures in vacuo for the acetylacetone tautomers: (a) diketo and (b) keto–enol.

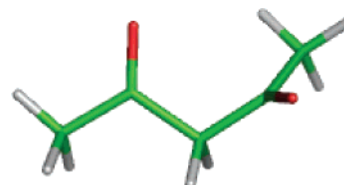


Figure 2. IEF-PCM/B3LYP/6-31G* lowest-energy structure in water for the diketo form of acetylacetone.

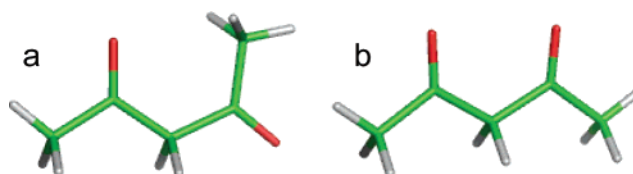


Figure 3. Additional minimum-energy structures at the B3LYP/6-31G* level for the diketo form of acetylacetone in vacuo: (a) $\Delta E = 3.48$ kcal/mol and (b) $\Delta E = 6.78$ kcal/mol.

Chart 1, a minimum indeed, is about twice as much less stable. Interestingly enough, the IEF-PCM/B3LYP/6-31G* optimization starting from either one of the aforementioned structures produces the arrangement displayed in Figure 2 which, in aqueous solution, is the most stable one, while the solvated diketo form of Figure 1a is less favorable by 0.39 kcal/mol.

IEF-PCM calculations predict that the keto–enol form of acetylacetone is more stable than the diketo tautomer both in THF and in water (Table 5). The results hardly depend

Table 6. IEF-PCM/B3LYP/6-31G* Unscaled Vibrational Frequencies (cm^{-1}) for Selected Modes of the Acetylacetone Tautomers in THF and in Water Using Different Radii

mode	THF		water	
	UA0	Bondi	UA0	Bondi
diketo form				
Rot C(5)H ₃	136.9	136.0	112.3	114.4
Rot C(1)H ₃	147.1	145.8	131.8	128.6
Stretch C(3)H (s)	3024.8	3030.0	3026.9	3033.9
Stretch C(3)H (as)	3102.2	3107.7	3086.0	3090.9
keto-enol form				
Rot C(5)H ₃	48.0	44.6	48.5	44.9
Rot C(1)H ₃	117.5	117.1	116.9	116.8
Bend OH	1646.5	1655.7	1636.6	1650.5
Stretch OH	3036.8	3029.7	3046.9	3032.1
Stretch C(3)H	3197.5	3216.3	3183.9	3212.2

on the cavity calculation method, and ΔE_{int} is invariably the dominating contribution to ΔG_{tot} . The basis set effect on the calculated ΔE_{int} is more than 2 kcal/mol, but for this system, the geometric effect is also important. ΔE_{int} values, calculated with the 6-311++G** basis set with two different geometries, differ by 0.4–0.8 kcal/mol. For example, the C₂C₃C₄O₂ torsion angle, reported in Table S3 (Supporting Information), differs by more than 10° in the optimized geometries obtained with the two basis sets.

The positive ΔG_{tot} values predict higher free energy in the diketo as compared to the keto-enol form at any considered level. ΔG_{tot} is less positive in water than in THF, showing increasing abundance of the diketo form with increasing solvent polarity. This is in qualitative agreement with the experimental results^{28c} which, however, indicate that the diketo form is the prevailing tautomer in the aqueous solution with a diketo/keto-enol ratio of 100:34. The geometry relaxation of the gas-phase structure is mainly related to the OCCC torsions in the diketo form that vary from 88/88 in vacuo to about 100/10 in solution. They, however, change insignificantly with the basis set and only moderately with the solvent polarity.

The calculated ΔG_{therm} is large for the acetylacetone tautomerism. In fact, addition of ΔG_{therm} to ΔG_{tot} reduces the latter by about 70% at the B3LYP/6-31G* level. Table 6 shows some calculated vibrational frequencies of special interest. The basic structural change in acetylacetone throughout the tautomeric transformation is that the O=C-CH₂ moiety isomerizes into the HO-C=CH substructure. The developed enolic OH group forms an intramolecular H bond with the remaining carbonyl oxygen. Thus, the C=O and C-H bonds of the diketo form disappear, and C=C and O-H bonds come into existence in the keto-enol form.

Interestingly enough, the IR spectra computed in solution (THF or water) using either the UA0 or Bondi cavities for the diketo form (Figures S1–S2 of the Supporting Information) are indistinguishable; just the maximum intensities change somewhat. The keto-enol form (Figures S3–S4, Supporting Information) shows much higher maximum intensities, especially using Bondi cavities. When the IR spectra for the keto-enol form in THF and in water are

compared using the same radii (either UA0 or Bondi), there is some shift with UA0 below 1000 cm^{-1} , while the peaks within 1000 and 1500 cm^{-1} in water are slightly taller than in THF. Conversely, using Bondi radii they are still indistinguishable. The tallest peak (about 1700 cm^{-1}), present in both **5** and **6**, corresponds to the bending of the overall structures. Methyl CH stretching modes are characterized by very small peaks. Conversely, the OH stretching mode displays tall peaks. Table 6 shows that both the C-H and O-H frequencies are about 3000 cm^{-1} , producing small differences in ΔZPE . The change in ΔZPE is also relatively small due to the disappearance and appearance of a C=O and a C=C bond, respectively. The large variation in ΔG_{tot} should, in our opinion, be due to the changes in the methyl torsional vibrations. Since each tautomer bears two methyl groups, we originally guessed that their vibrational frequencies would not considerably change. The frequency analysis, however, pointed out different results. Since in the harmonic oscillator approximation low-frequency motions provide the largest vibrational entropy contribution to ΔG_{therm} , the classification of the methyl group motion as a vibration or as a hindered rotation becomes a crucial problem if the motion is remarkably different in the keto-enol compared to the diketo form. The consideration of this particular problem is out of the scope of the present study but anyway underlines how difficult a reliable comparison between calculated and experimental relative free energies (and related equilibrium constants) is in those cases when methyl groups are present in the molecule. ΔG_{therm} was smaller for the isonicotinic acid as well as for the 4-pyridone/4-OH-pyridine tautomeric pairs, where the atomic motions could be reasonably related to vibrations.

B. Atomic Charge Derivation. Despite problems regarding the sign of some ΔG_{tot} terms in the above IEF-PCM calculations, the method may still allow the derivation of atomic charges usefully applicable in MC calculations. FEP/MC calculations automatically account for the ΔG_{drc} term; thus, the crucial problem for the choice of a proper cavity-formation method is overcome in an explicit solvent study. A better estimate for the $\Delta E_{\text{int}}/\Delta E_{\text{elst}}$ balance, thus the balance of the relative internal and solvation free energy, may produce total relative free energies in acceptable agreement with the experimental values. To this aim, a number of charge sets have been derived and applied in FEP/MC calculations. The full collection of derived charges and relevant dipole moments is supplied in Tables S4–S39 of the Supporting Information. An analysis of the charges in those tables reveals the roles of the solvent, selected cavity model, basis set, and fitting method in the derivation procedure.

Charges of specific interest for the isonicotinic acid tautomers are summarized in Tables 7 and 8. A concise conclusion is that the derived in-solution values are sensitive to the basis set applied in the IEF-PCM/B3LYP calculations and the procedure, CHELPG or RESP, used for fitting atomic charges to the generated ELPO. Setting these two conditions, the charges differ only slightly, generally up to 0.03 charge units depending on the solvent, the cavity model taken in the PCM calculations, and whether the geometry was optimized with the 6-31G* or 6-311++G** basis set. The

Table 7. Isonicotinic Acid Atomic Charges Fitted to the B3LYP Electrostatic Potentials in the Gas Phase and in Solution (IEF-PCM/B3LYP/6-31G* Optimization in the Solvent Indicated, UA0 Cavity)

	gas phase				THF			
	6-31G*		6-311++G**/6-31G*		6-31G*		6-311++G**/6-31G*	
	CHELPG	RESP	CHELPG	RESP	CHELPG	RESP	CHELPG	RESP
	ZW							
N	-0.226	-0.116	-0.211	-0.087	-0.217	-0.080	-0.197	-0.046
H	0.324	0.306	0.313	0.294	0.380	0.358	0.372	0.348
C _{carb}	0.551	0.512	0.688	0.626	0.590	0.549	0.776	0.714
O	-0.595	-0.573	-0.673	-0.644	-0.685	-0.663	-0.795	-0.768
	Neutral							
N	-0.580	-0.509	-0.644	-0.594	-0.649	-0.582	-0.739	-0.700
H	0.424	0.415	0.433	0.428	0.475	0.465	0.490	0.486
C _{carb}	0.530	0.477	0.593	0.509	0.561	0.505	0.642	0.555
=O	-0.480	-0.447	-0.526	-0.486	-0.533	-0.500	-0.596	-0.556
-O	-0.550	-0.511	-0.587	-0.546	-0.569	-0.528	-0.614	-0.572
	methanol				water			
	6-31G*		6-311++G**/6-31G*		6-31G*		6-311++G**/6-31G*	
	CHELPG	RESP	CHELPG	RESP	CHELPG	RESP	CHELPG	RESP
	ZW							
N	-0.209	-0.071	-0.186	-0.038	-0.208	-0.070	-0.185	-0.036
H	0.391	0.369	0.384	0.360	0.393	0.371	0.386	0.363
C _{carb}	0.594	0.551	0.788	0.723	0.595	0.552	0.791	0.725
O	-0.700	-0.678	-0.816	-0.788	-0.703	-0.681	-0.820	-0.792
	Neutral							
N	-0.665	-0.597	-0.763	-0.724	-0.668	-0.600	-0.767	-0.729
H	0.486	0.476	0.503	0.498	0.488	0.479	0.505	0.501
C _{carb}	0.567	0.512	0.652	0.566	0.566	0.513	0.651	0.567
=O	-0.543	-0.511	-0.610	-0.571	-0.544	-0.513	-0.611	-0.574
-O	-0.572	-0.532	-0.620	-0.578	-0.572	-0.533	-0.619	-0.580

Table 8. Isonicotinic Acid Atomic Charges Fitted to IEF-PCM/B3LYP Electrostatic Potentials for the Geometries Optimized in Solution at the IEF-PCM/B3LYP/6-31G* Level in the Solvent Indicated, Bondi Cavity

	water				methanol				THF			
	6-31G*		6-311++G**		6-31G*		6-311++G**		6-31G*		6-311++G**	
	CHELPG	RESP	CHELPG	RESP	CHELPG	RESP	CHELPG	RESP	CHELPG	RESP	CHELPG	RESP
	ZW											
N	-0.194	-0.080	-0.164	-0.046	-0.205	-0.066	-0.179	-0.030	-0.211	-0.075	-0.188	-0.040
H	0.369	0.357	0.359	0.349	0.367	0.344	0.357	0.333	0.361	0.338	0.351	0.327
C _{carb}	0.595	0.590	0.786	0.759	0.586	0.545	0.774	0.710	0.579	0.541	0.759	0.700
O	-0.693	-0.683	-0.809	-0.793	-0.688	-0.667	-0.802	-0.774	-0.674	-0.653	-0.781	-0.755
	Neutral											
N	-0.655	-0.588	-0.750	-0.712	-0.652	-0.585	-0.747	-0.708	-0.639	-0.571	-0.727	-0.686
H	0.458	0.449	0.470	0.465	0.457	0.448	0.469	0.464	0.452	0.442	0.463	0.458
C _{carb}	0.562	0.508	0.641	0.555	0.560	0.507	0.639	0.553	0.555	0.501	0.631	0.546
=O	-0.529	-0.496	-0.592	-0.552	-0.527	-0.495	-0.590	-0.550	-0.519	-0.487	-0.579	-0.539
-O	-0.561	-0.521	-0.601	-0.559	-0.560	-0.521	-0.601	-0.558	-0.559	-0.519	-0.600	-0.557

gas-phase charges differ from the in-solution values remarkably. The effect of the different charge sets may be assessed from the difference in the calculated $\Delta G(\text{soln})$ values. $\Delta G(\text{soln})$ differs by 1 kcal/mol if the 6-31G* and the 6-311++G** Bondi/RESP sets (Table 9) are applied for the isonicotinic acid tautomers in THF. The difference in $\Delta G(\text{soln})$ is, however, 7 kcal/mol in water despite the small deviation in the calculated dipole moments: 2.51 versus 2.60

D (neutral) and 18.53 versus 19.81 D (zwitterion) at the IEF-PCM/B3LYP/6-31G*/IEF-PCM/B3LYP/6-31G* and IEF-PCM/B3LYP/6-311++G**/IEF-PCM/B3LYP/6-31G* levels, respectively. (No experimental data are available.) Thus, whereas the global physical parameter, the dipole moment, shows a small variation with the basis set, the derived atomic charges lead to largely different relative solvation free energies. Dipole moments and $\Delta G(\text{soln})$ values were calcu-

Table 9. Relative Solvation Free Energies, $\Delta G(\text{sol})^a$

isonicotinic acid		UA0		Bondi	
ZW (2) – Neu (1)	CHELPG	RESP	CHELPG	RESP	
		THF			
6-31G ^{*b}	-19.73 ± 0.12	-19.78 ± 0.11	-19.48 ± 0.11	-19.46 ± 0.11	
RCUT = 12.0 + RF ^c				-18.06 ± 0.12	
custom THF ^d + RF				-19.39 ± 0.12	
Ewald ^e				-3.90 ± 0.12	
6-311++G ^{**f}				-20.56 ± 0.12	
		MeOH			
6-31G [*]			-38.78 ± 0.15	-38.82 ± 0.15	
RCUT = 12.0 + RF			-34.97 ± 0.15	-35.75 ± 0.14	
6-311++G ^{**}				-45.86 ± 0.15	
		Water			
6-31G [*]	-51.41 ± 0.14	-52.49 ± 0.15	-47.85 ± 0.13	-50.64 ± 0.14	
RCUT = 12.0 + RF				-43.89 ± 0.15	
Ewald				-26.37 ± 0.16	
6-311++G ^{**}	-58.08 ± 0.15	-59.60 ± 0.15	-54.89 ± 0.13	-57.45 ± 0.16	
4-pyridone		UA0		Bondi	
4-py (4) – 4-OH (3)	CHELPG	RESP	CHELPG	RESP	
		THF			
6-31G [*]	-4.35 ± 0.09	-4.26 ± 0.10	-3.92 ± 0.09	-4.21 ± 0.08	
		Water			
6-31G [*]	-10.40 ± 0.15	-11.87 ± 0.13	-10.76 ± 0.14	-10.96 ± 0.13	
6-311++G ^{**}	-14.24 ± 0.16	-14.77 ± 0.14	-13.53 ± 0.14	-13.80 ± 0.14	
acetylacetone		UA0		Bondi	
diketo (6) – enol (5)	CHELPG	RESP	CHELPG	RESP	
		THF			
6-31G [*]	-2.32 ± 0.07	-2.35 ± 0.06	-2.24 ± 0.06	-2.33 ± 0.06	
6-311++G ^{**}				-2.92 ± 0.06	
		Water			
6-31G [*]	-3.79 ± 0.19	-3.18 ± 0.15	-1.81 ± 0.18	-2.79 ± 0.15	
RCUT = 12.0 + RF				-4.77 ± 0.11	
6-311++G ^{**}			-3.19 ± 0.20	-2.52 ± 0.19	

^a Energies in kcal/mol. SCUT (solute–solvent cutoff) = 12.0 Å in all simulations. ^b IEF-PCM/B3LYP/6-31G^{*} charges. RCUT (solvent–solvent cutoff) = 9.75 Å. ^c RF: reaction field applied with dielectric constant for the pure solvent. ^d All-atom custom THF solvent, see the text. ^e Ewald summation, RCUT = 12.0 Å. ^f IEF-PCM/B3LYP/6-311++G^{**}//IEF-PCM/B3LYP/6-31G^{*} charges, RCUT = 9.75 Å.

lated in methanol very similarly to those in water (compare Tables S8 and S9 to S4 and S5, Supporting Information; Table 9). Since the relative internal energy at the QCISD-(T) level changes by up to 0.3 kcal/mol with the two basis sets (Table 10), the total relative free energy (eq 5) strongly depends on the applied charge set in highly polar solvents.

The last important factor to be considered is the fitting procedure. Polar atomic charges calculated with the CHELPG or the RESP procedure differ in every molecule. Differences of at least 0.05 units can be found both with the 6-31G^{*} and the 6-311++G^{**} basis set. An a priori decision between the two methods is difficult, since both procedures provide derived atomic charges reproducing the IEF-PCM/B3LYP/6-31G^{*} dipole moments as close as 0.01–0.02 D. Thus, FEP/MC calculations both with the CHELPG and RESP charge sets were performed for the tautomeric pairs, using different simulation parameters, as reported in Table 9.

C. Monte Carlo and Molecular Dynamics Simulations. Relative solvation free energies calculated with different

charge sets are compared in Table 9. In the case of the Bondi/RESP charges, several different simulation models were considered (see also the footnote for Table 9). In the simplest model, the system corresponds to an infinitely dilute solution where the solute–solute interaction has been disregarded. This may be a reasonable approach since experimental data refer to 0.0003–0.1 molar solutions.²⁸ Headers 6-31G^{*} and 6-311++G^{**} indicate the basis set used in calculating the IEF-PCM/B3LYP molecular electrostatic potential for charge derivation. The solvent–solvent cutoff was set to 9.75 Å in these simulations. The long-range solute–solvent electrostatic interaction was corrected in the PCM approximation by calculating the interaction energy of the solute embedded in a cavity carved in the corresponding dielectric (THF, water, and methanol). The cavity was created by overlapping spheres around the solute atoms with radii of 12 Å, equal to the solute–solvent cutoff. In the second infinitely dilute solution model, the solvent–solvent cutoff was set to 12.0 Å and a reaction field was applied (RCUT = 12.0 +

Table 10. Relative Internal Energies (kcal/mol) at the QCISD(T) and CCSD(T) Levels with Different Basis Sets^a

isonicotinic acid		THF		methanol		water	
$E_{ZW} - E_{Neu}$	UA0	Bondi	Bondi	UA0	Bondi	UA0	Bondi
QCISD(T)							
6-31G*	44.62	43.24	45.91	48.63	46.62		
6-311++G**	44.15	42.93	46.13	48.35	46.53		
cc-pVTZ		43.74			46.91		
CCSD(T)							
6-31G*	44.75	43.39		48.74	46.75		
4-pyridone		THF		water			
$E_{4-pyr} - E_{4-OH}$	UA0	Bondi	UA0	Bondi	UA0	Bondi	
QCISD(T)							
6-31G*	7.59	6.77	9.84	8.63			
6-311++G**	8.91	8.17	11.43	10.32			
cc-pVTZ		7.96		9.96			
CCSD(T)							
6-31G*	7.81	7.01	10.05	8.87			
cc-pVTZ		8.19		10.20			
acetylacetone		THF		water			
$E_{diketo} - E_{enol}$	UA0	Bondi	UA0	Bondi	UA0	Bondi	
QCISD(T)							
6-31G*	0.28	0.11	1.29	1.10			
6-311++G**	3.18	3.02	4.38	4.34			
cc-pVTZ		5.97		7.10			
CCSD(T)							
6-31G*	0.29	0.13	1.30	1.10			

^a Geometries from IEF-PCM/B3LYP/6-31G* optimizations.

RF calculations), whereas the atomic charges were derived by means of the IEF-PCM/B3LYP/6-31G* wavefunction.

$\Delta G(\text{solv})$ estimates with the most simulation conditions/charge parametrizations have been carried out for the isonicotinic system. The values in a row indicate the same simulation conditions with charge sets of different origin. Values in any row (also for the 4-pyridone/4-OH-pyridine and the acetylacetone system, in the lower part of Table 9) do not differ much in most cases. Thus, $\Delta G(\text{solv})$ does not significantly depend on the origin of the charge set. The simulation conditions (RCUT = 9.75 Å, RCUT = 12.0 Å + reaction field, and Ewald summation) have, however, a remarkable effect on $\Delta G(\text{solv})$ in most cases. The basis set, used in deriving the partial charges, also has a considerable effect on $\Delta G(\text{solv})$.

Special attention is to be paid to the calculated $\Delta G(\text{solv})$ values for the isonicotinic acid equilibria. Despite the large relative solvation free energies (in absolute value), the calculation is robust in aqueous solution: the forward and backward $\Delta G(\text{solv})$ values at the 6-31G*/CHELPG/Bondi level are -47.85 ± 0.13 and 48.33 ± 0.14 kcal/mol, respectively. When the 6-311++G** charges are used, $\Delta G(\text{solv})$ becomes more negative by 1–7 kcal/mol in the THF, methanol, and water series. In contrast, $\Delta G(\text{solv})$ was calculated less negative with the RCUT = 12.0 + RF simulation using the 6-31G* charges. The differences are 1.4, 3.1–3.8, and 6.7 kcal/mol in the three solvents, respectively.

For the calculation of the long-range electrostatic interactions, the Ewald summation is the most widely used method nowadays. By applying this method, $\Delta G(\text{solv})$ for the isonicotinic system was calculated to be much less negative than by means of other methods. It is implicit in the Ewald summation that the reference solution box is surrounded by an infinite number of its replica. In our systems, the equilibrium box edge was 24–33 Å, producing a nearly 0.1 molar solution concentration. Thus, the Ewald summation refers to an about 0.1 molar solution, which is much denser than the experimental one (about 0.0003 molar).^{28a} The unrealistically low $\Delta G(\text{solv})$ predicts a too positive ΔG_{tot} in eq 5 and thus predicts the preference for the neutral form in water, in contrast to the experiment. Therefore, we concluded that the Ewald summation is inapplicable for modeling the present dilute solutions and was not further considered in relation to the 4-pyridone/4-OH-pyridine system and for the keto–enol tautomerism for acetylacetone.

The $\Delta G(\text{solv})$ values for the isonicotinic system are -18 to -21 kcal/mol in THF if Ewald summation is not applied. Then, ΔG_{tot} is at least $+15$ kcal/mol, taking ΔE_{int} from Tables 3 or 10 (see below). The corresponding experimental value is 2.7 kcal/mol.^{28a} The large difference was attributed first to the inappropriateness of the five-point THF model,^{19b} and the authors developed an all-atom THF model for the pure liquid with the density and heat of vaporization close to the experimental values (Table 2). When a solvent box comprised of 264 custom THF molecules was used, RCUT = SCUT = 12 Å was set, and a reaction field with a dielectric constant of 7.43 was applied, $\Delta G(\text{solv})$ was calculated at -19.4 kcal/mol. This value does not differ significantly from previous values with simpler models.

The failure for predicting a $\Delta G(\text{solv})$ value close to the experimental one initiated the idea that isonicotinic acid forms dimers in THF. Thus, 2-ns-long MD simulations have been carried out for the tautomers in the recently developed THF box. The simulations started with stacked, antiparallel solute dimers. The neutral dimer dissociated, whereas the stacked zwitterionic dimer moved into a hydrogen-bonded linear form. Thus, MD simulations indicate that the neutral form is dissolved in THF, whereas the zwitterion forms at least dimers. (Only two solute molecules were considered.) This result means that a simple transformation of a single neutral tautomer into a zwitterion throughout FEP/MC simulations is not a correct model. Although on the basis of the present MD modeling the correct $\Delta G(\text{solv})$ has not been derived, the qualitative results indicate that the solvation in THF is more complicated than expected before. Further consideration of this problem is in progress.

An argument may be raised that the MD simulations produced an artifact for the linear zwitterionic dimer. Similar simulations have been performed for such dimers in water, and these MDs predicted largely separated neutral and zwitterionic molecules. Thus, both isonicotinic acid tautomers are present in monomeric form in water, and this feature allows the one-to-one transformation in the FEP/MC procedure. Good simulation parameters for this system provide $\Delta G(\text{solv})$ and, ultimately, ΔG_{tot} results close to experimental

Table 11. Calculated ΔG_{tot} Values for the Zwitterionic Isonicotinic Acid Relative to the Neutral Form in Aqueous Solution^a

Charges for $\Delta G(\text{solv})$		ΔG_{tot}		
		ΔE_{int} from B3LYP	ΔE_{int} from QCISD(T)	
UA0	6-31G*	6-31G*	6-31G*	
	CHELPG	-7.57	-1.90	
	RESP	-8.65	-2.98	
	6-311++G**	6-311++G**	6-311++G**	
	CHELPG	-16.21	-8.85	
	RESP	-17.73	-10.35	
Bondi	6-31G*	6-31G*	6-31G*	cc-pVTZ
	CHELPG	-5.95	-0.36	-0.07
	RESP	-8.74	-3.15	-2.86
	RCUT = 12.0 +RF	-1.99 ^b	3.60 ^b	3.89 ^b
	6-311++G**	6-311++G**	6-311++G**	cc-pVTZ
	CHELPG	-14.84	-7.49	-7.11
	RESP	-17.40	-10.05	-9.67
exp ^c		-2.59 ± 0.05		

^a $\Delta G_{\text{tot}} = \Delta E_{\text{int}} + \Delta G(\text{solv}) + \Delta G_{\text{therm}} + RT \ln 2$. Values in kcal/mol. Standard deviations as for the corresponding $\Delta G(\text{solv})$ value in Table 9. ^b RESP charges. ^c From ref 28a.

results. In contrast, the ΔG_{tot} calculation fails in THF when assuming the presence of monomeric zwitterions in solution.

$\Delta G(\text{solv})$ for the 4-pyridone/4-OH-pyridine system was calculated at the 6-31G* level both in THF and in water. Charge sets of different origin led to almost identical $\Delta G(\text{solv})$ values of about -4 kcal/mol in THF and varied in the -10.4 to -11.9 kcal/mol range in water. When the 6-311++G** level in water was used, the calculated $\Delta G(\text{solv})$ values scattered in the -13.5 to -14.8 kcal/mol range. The widths of the ranges in aqueous solution are very similar to each other for the two charge sets, but the 6-311++G** $\Delta G(\text{solv})$ values are significantly more negative by about 3 kcal/mol.

For the keto-enol tautomerism of acetylacetone, the calculated $\Delta G(\text{solv})$ in THF is nearly equal using either the 6-31G* or 6-311++G** charges. The relative solvation free energy in water varies in the -1.8 to -3.8 kcal/mol range irrespective of using the 6-31G*- or 6-311++G**-derived charges. These calculations were performed using an RCUT of 9.75 Å and PCM-based correction for the long-range electrostatic. At the RCUT = 12.0 + RF level (dielectric constant = 78.39), $\Delta G(\text{solv})$ was calculated at -4.8 kcal/mol, which is 2 kcal/mol more negative than the corresponding 6-31G*-level value.

Thus, the use of atomic charges from fits to the in-solution B3LYP/6-31G* and B3LYP/6-311++G** ELPOs produces considerably different $\Delta G(\text{solv})$ values in several cases, mainly in water. Still the tautomeric equilibrium constant could be predicted correctly if $\Delta G(\text{solv})$ is combined with an appropriate ΔE_{int} term providing $\Delta G_{\text{tot}} = \Delta E_{\text{int}} + \Delta G(\text{solv}) + \Delta G_{\text{therm}}$ (plus possible corrections due to molecular symmetry) close to the experimental ΔG . Table 10 shows that not only $\Delta G(\text{solv})$, reported in Table 9, but also $\Delta E_{\text{int}}(\text{corr})$ can change remarkably with the applied basis set. This problem will be examined in the next section.

D. Calculation of ΔG_{tot} . ΔG_{tot} (theoretical) shows a subtle interplay of its two main components, ΔE_{int} and $\Delta G(\text{solv})$.

A recent study^{27e} revealed that neither IEF-PCM/B3LYP/6-311++G** nor IEF-PCM/MP2/6-311++G** level calculations can account correctly for the relative internal energies in an enolimine-enaminone tautomeric equilibrium in different organic solvents. Good agreement with the experimental ΔG_{tot} was achieved, however, when ΔE_{int} was calculated as $\Delta E_{\text{int}}(\text{corr})$. Thus, $\Delta E_{\text{int}}(\text{corr})$ (eq 1) was calculated also in the present study with basis sets up to cc-pVTZ (Table 10).

Calculations of $\Delta E_{\text{int}}(\text{corr})$ at the QCISD(T) and CCSD(T) levels (the IEF-PCM reference for all internal energies will be omitted henceforth) lead to small differences up to 0.2 kcal/mol in comparable cases, that is, when the same basis set was used. The results show, however, different basis set dependence of $\Delta E_{\text{int}}(\text{corr})$ for the three studied tautomeric systems. $\Delta E_{\text{int}}(\text{corr})$ is fairly stable in all three solvents for the isonicotinic tautomers and shows a saturation trend for the 4-pyridone/4-OH-pyridine pair but increases monotonically for the ketone-enol system with increasing the basis set. The combinations for ΔG_{tot} were taken so that the basis considered in ΔE_{int} fitted to the one used in the charge derivation for calculating $\Delta G(\text{solv})$. Furthermore, the corresponding cavity model, UA0 or Bondi, was applied in relation to ΔE_{int} and the charge set used in determining $\Delta G(\text{solv})$. The presented ΔG_{tot} values include the ΔG_{therm} corrections from Tables 3–5. The optimized geometry for each of the zwitterionic isonicotinic acids and 4-pyridones shows 2-fold symmetry providing a symmetry number of 2 and reducing the rotational entropy by $R \ln 2$. This effect makes ΔG_{tot} less negative by $RT \ln 2 = 0.41$ kcal/mol.

The general conclusion from the analyses of Tables 11–16 is that we have succeeded in deriving ΔG_{tot} values close to the available experimental ones, but a unique and superior combination of the ΔE_{int} and $\Delta G(\text{solv})$ terms in calculating ΔG_{tot} has not been found. Part of the problem is the weak to strong basis set dependence of the $\Delta E_{\text{int}}(\text{QCISD(T),corr})$ term for the studied equilibria.

Table 12. Calculated ΔG_{tot} Values for the Zwitterionic Isonicotinic Acid Relative to the Neutral Form in Methanol^a

charges for $\Delta G(\text{solv})$		ΔE_{int} from B3LYP	ΔE_{int} from QCISD(T)
Bondi	6-31G*	6-31G*	6-31G*
	CHELPG	<u>2.46</u>	7.99
	RESP	<u>2.42</u>	7.95
	RCUT = 12.0 +RF	6-31G*	6-31G*
	CHELPG	5.41	11.80
	RESP	4.63	11.02
	6-311++G**	6-311++G**	6-311++G**
	RESP	-6.54	1.13
exp ^b	2.3		

^a $\Delta G_{\text{tot}} = \Delta E_{\text{int}} + \Delta G(\text{solv}) + \Delta G_{\text{therm}} + RT \ln 2$. Values in kcal/mol. Standard deviations as for the corresponding $\Delta G(\text{solv})$ value in Table 9. ^b ΔG_{tot} from ref 28a.

Table 13. Calculated ΔG_{tot} Values for 4-Pyridone Relative to 4-OH-Pyridine in Aqueous Solution^a

charges in $\Delta G(\text{solv})$		ΔE_{int} from B3LYP	ΔE_{int} from QCISD(T)	
UA0	6-31G*	6-31G*	6-31G*	
	CHELPG	<u>-4.22</u>	0.38	
	RESP	<u>-5.68</u>	-1.09	
	6-311++G**	6-311++G**	6-311++G**	
	CHELPG	-7.61	-1.87	
	RESP	-8.14	-2.40	
Bondi	6-31G*	6-31G*	6-31G*	cc-pVTZ
	CHELPG	<u>-5.39</u>	-1.27	0.07
	RESP	<u>-5.59</u>	-1.47	-0.14
	6-311++G**	6-311++G**	6-311++G**	cc-pVTZ
	CHELPG	-7.74	-2.35	-2.71
	RESP	-8.01	-2.62	-2.98
exp ^b	-4.5			

^a $\Delta G_{\text{tot}} = \Delta E_{\text{int}} + \Delta G(\text{solv}) + \Delta G_{\text{therm}} + RT \ln 2$. Values in kcal/mol. Standard deviations as for the corresponding $\Delta G(\text{solv})$ value in Table 9. ^b ΔG_{tot} from ref 28b

Table 14. Calculated ΔG_{tot} Values for 4-Pyridone Relative to 4-OH-Pyridine in THF^a

charges for $\Delta G(\text{solv})$		ΔE_{int} from B3LYP	ΔE_{int} from QCISD(T)	
UA0	6-31G*	6-31G*	6-31G*	
	CHELPG	0.36	4.10	
	RESP	0.45	4.19	
Bondi	6-31G*	6-31G*	6-31G*	cc-pVTZ
	CHELPG	0.25	3.65	4.84
	RESP	-0.04	3.36	4.55

^a $\Delta G_{\text{tot}} = \Delta E_{\text{int}} + \Delta G(\text{solv}) + \Delta G_{\text{therm}} + RT \ln 2$. Values in kcal/mol. Standard deviations as for the corresponding $\Delta G(\text{solv})$ value in Table 9.

Tables 11 and 12 summarize the calculated ΔG_{tot} values for the isonicotinic acid tautomers in water and methanol, respectively. No such table has been provided with the THF solvent where the zwitterionic tautomer forms a linear dimer according to the MD simulations. These authors ascribe the failure of FEP/MC calculations to this structural peculiarity. In contrast to the THF solutions, as stated above, the MD simulations have resulted in two, largely separated zwitterionic isonicotinic acid molecules in aqueous solution, either starting from a stacked or a linear dimer form. In aqueous

solution, four ΔG_{tot} values were calculated with no more than 0.6 kcal/mol deviation from the experimental value of -2.59 ± 0.05 kcal/mol.^{28a} (The best results are underscored in this and subsequent tables.) The three best results in the range of -2.86 to -3.15 kcal/mol were obtained by using the RESP/6-31G* parametrization and the $\Delta E_{\text{int}}(\text{QCISD(T),corr})$ values. The RESP charges are clearly superior in comparison with CHELPG charges for this equilibrium. When the UA0 versus Bondi cavity is used, the calculated ΔG_{tot} changes by less than 0.2 kcal/mol, -2.98 versus -3.15 kcal/mol. When the cc-pVTZ $\Delta E_{\text{int}}(\text{QCISD(T),corr})$ is used instead of the ΔE_{int} corrected at the QCISD(T)/6-31G* level, the improvement is 0.3 kcal/mol. The experimental ΔG_{tot} of -2.59 kcal/mol predicts a percentage zwitterionic/neutral composition of 98.8:1.2. Even with the calculated value of -3.15 kcal/mol, the corresponding ratio increases only to 99.5:0.5.

All the above values were calculated by using an RCUT = 9.75 Å and a followup LRE correction utilizing the PCM method. By applying an RCUT of 12.0 Å and a reaction field (RCUT = 12.0 + RF) throughout the calculations of the FEP increments, a value of -1.99 kcal/mol was calculated with the 6-31G*/Bondi/RESP charges. This value

Table 15. Calculated ΔG_{tot} Values for the Acetylacetone Diketone Relative to the Ketone-Enol Form in Aqueous Solution^a

charges for $\Delta G(\text{solv})$		ΔE_{int} from B3LYP	ΔE_{int} from QCISD(T)	
UA0	6-31G*	6-31G*	6-31G*	
	CHELPG	<u>-0.21</u>	-4.19	
	RESP	0.40	-3.58	
Bondi	6-31G*	6-31G*	6-31G*	cc-pVTZ
	CHELPG	1.74	-2.44	3.56
	RESP	0.76	-3.42	2.58
	RCUT = 12.0 +RF	<u>-1.22^b</u>	-5.40 ^b	0.60 ^b
	6-311++G**	6-311++G**	6-311++G**	cc-pVTZ
	CHELPG	2.75	<u>-0.58</u>	2.18
	RESP	3.42	0.09	2.85
exp ^c		-0.64		

^a $\Delta G_{\text{tot}} = \Delta E_{\text{int}} + \Delta G(\text{solv}) + \Delta G_{\text{therm}}$. Values in kcal/mol. Standard deviations as for the corresponding $\Delta G(\text{solv})$ value in Table 9. ^b RESP charges. ^c ΔG_{tot} from ref 28c.

Table 16. Calculated ΔG_{tot} Values for the Acetylacetone Diketone Relative to the Keto-Enol Form in THF^a

charges for $\Delta G(\text{solv})$		ΔE_{int} from B3LYP	ΔE_{int} from QCISD(T)	
UA0	6-31G*	6-31G*	6-31G*	
	CHELPG	<u>0.80</u>	-3.46	
	RESP	<u>0.77</u>	-3.49	
Bondi	6-31G*	6-31G*	6-31G*	cc-pVTZ
	CHELPG	<u>0.76</u>	-3.64	2.22
	RESP	<u>0.67</u>	-3.73	2.13
	6-311++G**	6-311++G**	6-311++G**	cc-pVTZ
	RESP	2.36	-1.41	<u>1.54</u>
exp ^b		1.1		

^a $\Delta G_{\text{tot}} = \Delta E_{\text{int}} + \Delta G(\text{solv}) + \Delta G_{\text{therm}}$. Values in kcal/mol. Standard deviations as for the corresponding $\Delta G(\text{solv})$ value in Table 9. ^b ΔG_{tot} from ref 28c.

was obtained by using the B3LYP/6-31G* ΔE_{int} in Table 3. When the corresponding $\Delta E_{\text{int}}(\text{QCISD(T),corr})$ value was applied in the $\Delta E_{\text{int}} + \Delta G(\text{solv})$ combination, a $\Delta G_{\text{tot}} = +3.60$ kcal/mol was produced without even the correct sign. Since the standard deviation for the experimental value is very small compared to the data, the value of +3.60 kcal/mol is definitely an incorrect prediction. [Standard deviations for the calculated ΔG_{tot} data in Tables 11–16 are equal to those estimated for the corresponding $\Delta G(\text{solv})$ in Table 9.] Thus, Table 11 suggests that a fairly good estimate of ΔG_{tot} may also be obtained by a combination of $\Delta E_{\text{int}}(\text{B3LYP/6-31G*}) + \Delta G(\text{solv}, \text{RCUT} = 12.0 + \text{RF})$. The predicted percentage composition was zwitterionic/neutral = 96.7:3.3.

Nevertheless, Table 12 contradicts this latter assumption. By applying the $\Delta E_{\text{int}}(\text{B3LYP/6-31G*}) + \Delta G(\text{solv}, \text{RCUT} = 12.0 + \text{RF})$ combination for the isonicotinic acid equilibrium in methanol, this approach overestimates the experimental ΔG_{tot} by 2–3 kcal/mol. [Although the published data in ref 28a prevent the estimation of the standard deviation for $\Delta G_{\text{exp}}(\text{methanol})$, since the same experimental technique was applied as in the case of the aqueous solution, the standard deviation has not been expected to be larger than a few tenths of a kilocalorie per mole.] Table 12 shows that a good agreement of the calculated and experimental ΔG_{tot} was reached by the combination of $\Delta E_{\text{int}}(\text{B3LYP/6-31G*}) +$

$\Delta G(6-31G*/\text{Bondi})$, either by applying the CHELPG or the RESP charges. By using the $\Delta E_{\text{int}}(\text{QCISD(T)/6-31G*}, \text{corr})$ relative internal energy, the predicted ΔG_{tot} has been strongly overestimated. In contrast, however, a somewhat underestimated value of 1.13 kcal/mol (including ΔLRE as in all cases where the RF term is not explicitly indicated) was calculated with the 6-311++G**/Bondi/RESP charge parametrization and using the $\Delta E_{\text{int}}(\text{QCISD(T)/6-311++G**}, \text{corr})$ relative internal energy. All calculations (except one, $\Delta G_{\text{tot}} = -6.54$ kcal/mol) predict the preference of the neutral form; the best result is neutral/zwitterionic = 98.4:1.6 compared to the experimental composition of 98:2.

The 4-pyridone/4-OH-pyridine calculations in aqueous solution (Table 13) predict a stable preference for 4-pyridone. This table also indicates that much better agreement with the experimental value can be reached by utilizing the $\Delta E_{\text{int}}(\text{B3LYP/6-31G*})$ instead of the $\Delta E_{\text{int}}(\text{QCISD(T)/6-31G*}, \text{corr})$ term in calculating ΔG_{tot} . The best result (B3LYP/6-31G*) was obtained with the CHELPG charges; results with the RESP charges are too negative by about 1 kcal/mol with respect to the experiment. The experimental 4-pyridone/4-OH-pyridine composition in aqueous solution is 2000:1 compared with our best calculated ratio of 1247:1.

The solvent effect is large for the 4-pyridone/4-OH-pyridine tautomeric equilibria. The gas-phase equilibrium constant $K < 0.1$ increases to 2000 in aqueous solution.^{28b}

The dipole moment measured for the system subject to tautomerization was reported as 6.0–6.3 D in dioxane.^{36a,b} Our calculated dipole moments for 4-pyridone and 4-OH-pyridine are 8.8 and 3.5 D, respectively, at the IEF-PCM/B3LYP/6-31G* level with the Bondi cavity in THF. The corresponding values are 9.4 and 3.7 D in aqueous solution (compare Tables S6 and S7 with S14 and S15 in the Supporting Information). The experimental dipole moment about halfway between the calculated values for the pure tautomers suggests a nearly 1:1 equilibrium composition in the low-dielectric constant solvent ($\epsilon = 2.21$ for dioxane²⁰). Thus, we may conclude that the equilibrium is shifted from a 4-OH-pyridine preference in the gas phase to a nearly equal concentration of both tautomers in dioxane and to an overwhelming preference for 4-pyridone in aqueous solution. Our computational results in the THF solvent ($\epsilon = 7.43$) fit in this series.

Table 14 indicates a total free energy of -0.04 to $+0.45$ kcal/mol for 4-pyridone relative to 4-OH-pyridine in THF, when the $\Delta E_{\text{int}}(\text{B3LYP}/6-31\text{G}^*)$ term has been used for calculating ΔG_{tot} . These values provide 4-pyridone/4-OH-pyridine ratios from 1:0.9 to 1:2.1. A very small 4-pyridone fraction is predicted in the equilibrium mixture when the ΔG_{tot} is calculated by accepting the (QCISD(T)/6-31G*,-corr) relative internal energy. With these ΔE_{int} values, the ratio varies between 1:292 and 1:3553. These latter values are exceedingly smaller than the gas-phase experimental ratio. By recognizing that the dielectric constant for THF is much closer to that for dioxane than to that for water, the authors consider a ratio between 1:1 and 1:2 as a realistic prediction in THF.

The acetylacetone tautomerization was studied experimentally in a large number of solvents.^{28c} The K (enol/diketone) equilibrium constant gradually decreases with increasing solvent polarity. The K values were determined (studying 0.01 molar solutions) as 48, 6.5, 3.3, and 0.34 in hexane, THF, methanol, and water, respectively. The derived ΔG_{exp} is -0.64 kcal/mol for the diketone form relative to the enol tautomer in aqueous solution (Table 15). The theoretical calculations for this equilibrium provide ΔG_{tot} values in the broad range of -5.40 to $+3.56$ kcal/mol. The large variation is a consequence of the different combinations of the ΔE_{int} scattering in a range of 6.6 kcal/mol (Tables 5 and 9) and the $\Delta G(\text{solv})$ term varying by up to 3.0 kcal/mol in different FEP/MC calculations. The large positive values for ΔG_{tot} are most likely incorrect by assuming that at least the sign of the ΔG_{exp} is correct. Unless there was a systematic error in the experiment, the equal equilibrium constants derived at two different concentrations (0.01 and 0.1 molar) suggest the stable composition with about 3:1 diketone/enol.

Table 9 shows that the $\Delta G(\text{solv})$ values are robust at the 6-31G*/RESP level and hardly change upon the charge set derived on the basis of calculations with UA0 or Bondi cavity. In contrast, $\Delta G(\text{solv})$ calculated with CHELPG charges changes by 2 kcal/mol whether the charge set was obtained with the UA0 or the Bondi cavity via the IEF-PCM calculations. Except a small negative value of -0.21 kcal/mol for ΔG_{tot} , the other three values calculated by means of the $\Delta E_{\text{int}}(\text{B3LYP}/6-31\text{G}^*)$ relative internal energy are all

positive. When the $\Delta E_{\text{int}}(\text{B3LYP}/6-311++\text{G}^{**})$ relative energies are used with the Bondi cavity and the corresponding charge sets, ΔG_{tot} becomes even more positive. Consistently negative ΔG_{tot} values were calculated by applying the $\Delta E_{\text{int}}(\text{QCISD}(\text{T})/6-31\text{G}^*,\text{corr})$ relative energies, but the derived relative free energies are too negative. The very good $\Delta G_{\text{tot}} = -0.58$ kcal/mol was obtained by a combination of the $\Delta E_{\text{int}}(\text{QCISD}(\text{T})/6-311++\text{G}^{**},\text{corr})$ and the 6-311++G**/Bondi/CHELPG $\Delta G(\text{solv})$ terms. The results, however, might be a consequence of fortuitous error cancellations; this specific combination led to poor or moderately good results at most for other systems (Tables 11 and 13).

The ΔG_{tot} value derived by means of the $\text{RCUT} = 12.0 + \text{RF}/\Delta G(\text{solv})$ term (Table 9) in combination with the $\Delta E_{\text{int}}(\text{B3LYP}/6-31\text{G}^*)$ relative internal energy is -1.22 kcal/mol, deviating only by 0.6 kcal/mol from the experimental value. A similar deviation (although in the opposite direction) was obtained by this combination for the isonicotinic acid equilibrium in aqueous solution. Also, the correct sign for ΔG_{tot} (but a deviation of 2–3 kcal/mol) was found, however, for the latter equilibrium in methanol. Further studies are necessary for exploring how robust this combination is for different equilibrium systems.

In calculating ΔG_{tot} , only the IEF-PCM/B3LYP/6-31G* minimum energy structures were considered both for the enol and the diketone. Although the IEF-PCM calculations predicted two structures (Figures 1a and 2) for the diketone differing only by 0.39 kcal/mol, FEP/MC calculations provided an increase of 2.37 ± 0.13 kcal/mol in $\Delta G(\text{solv})$ for the Figure 1a structure. Since this conformer population in the diketone mixture is less than 2%, it has been disregarded and will not be the subject of further consideration.

The computational results for the acetylacetone tautomeric equilibrium in THF solvent (Table 16) may be summarized so that the $\Delta E_{\text{int}}(\text{B3LYP}/6-31\text{G}^*) + 6-31\text{G}^*/\Delta G(\text{solv})$ combination provides good agreement with the experimental values, whereas other combinations either lead to exaggeration of the ΔG_{tot} or even provide the wrong sign for the relative free energy. Application of the $\Delta E_{\text{int}}(\text{QCISD}(\text{T})/6-31\text{G}^*,\text{corr})$ term leads to a ΔG_{tot} of -3.46 to -3.73 kcal/mol compared to the value of 0.8–1.1 kcal/mol derived from the experimental K value. When $\Delta E_{\text{int}}(\text{QCISD}(\text{T})/\text{cc-pVTZ},\text{corr})$ is used, the sign is correct, but ΔG_{tot} becomes too positive: 1.54–2.22 kcal/mol.

Our calculated dipole moments at the IEF-PCM/B3LYP/6-31G* level (using either UA0 or Bondi cavity) range from 3.7 D (keto–enol) and 4.3 D (diketone) in THF to 3.9 D (keto–enol) and 5.1 D (diketone) in water (Tables S16–S21 in the Supporting Information). Thus, on the basis of these theoretical values, the more polar tautomer becomes the preferred one in the larger polarity solvent. The same trend was found for the other two aforementioned tautomeric systems as well.

An experimental study by Ghanadzadeh et al.³⁷ alerts, however, that dipolar molecules may form dimers in nonpolar solvents, causing an increased measurable value for the dipole moment of the solute. The type of the association depends on the functional group; ketones form dimers with

parallel C=O moments. The diketone form of acetylacetone has C=O bond moments pointing away depending on the chemical environment: the conformation is largely different in the gas phase and in solution (see above). Association may also be possible for 4-pyridone and, as it has been explored, for the zwitterionic isonicotinic acid in THF. These authors have concluded from the present study that a fast molecular dynamics exploration of the association character (monomer vs dimer) is useful prior to the application of the DFT + FEP/MC method for calculating ΔG_{tot} in nonpolar solvents.

A general comparison of the polarizable continuum dielectric and explicit solvent approaches indicates that the IEF-PCM method saves the sign of the ΔG_{tot} term in different solvents (Tables 3–5), whereas the DFT/FEP/MC procedure is more flexible (Tables 9–16). The experimental results seemingly are precise enough for considering at least the switch of the sign for ΔG to be significant. The IEF-PCM method did not predict the switch of the tautomeric preference for isonicotinic acid in water. As discussed above, the positive ΔE_{int} terms from B3LYP calculations determine the sign of ΔG_{tot} in all three solvents (Table 3), although the trend of the tautomeric preference has been correctly predicted. By considering thermal and symmetry corrections for the 4-pyridone/4-OH-pyridine system, ΔG_{tot} was predicted to be $-3.79 + 0.86 = -2.93$ kcal/mol at the IEF-PCM/B3LYP/6-31G*/Bondi theoretical level, in qualitative agreement with the experimental value of -4.5 kcal/mol in aqueous solution. The IEF-PCM/B3LYP/6-31G* ΔG_{tot} values of 1.1–1.3 kcal/mol, including thermal corrections, for the acetylacetone equilibrium are in good agreement with the experimental 0.8–1.1 kcal/mol in THF. The calculated positive sign of ΔG_{tot} disagrees, however, with the sign of the experimental value for acetylacetone in aqueous solution.

In contrast, by calculating ΔG_{tot} as $\Delta G_{\text{tot}} = \Delta E_{\text{int}}(\text{IEF-PCM/B3LYP/6-31G}^*) + 6\text{-31G}^*/\Delta G(\text{solv}) + \Delta G_{\text{therm}} + (\text{symmetry correction})$ in the framework of the DFT/FEP/MC procedure, deviations from experimental values were generally less than 1 kcal/mol. With $\text{abs}(\Delta G_{\text{exp}}) > 2$ kcal/mol, the deviation was small and the predicted sign for ΔG was correct. In cases with $\text{abs}(\Delta G_{\text{exp}}) < 1$ kcal/mol, the sign was not correctly determined in some cases. The total relative free energy for the isonicotinic acid equilibrium in aqueous solution was very well estimated, however, only when the QCISD(T)/cc-pVTZ relative internal energy was used.

IV. Conclusions

A systematic study has been performed for estimating the relative free energies of the tautomeric pairs in equilibria for isonicotinic acid in THF, methanol, and water, and for the 4-OH-pyridine and acetylacetone in THF and aqueous solution. Isonicotinic acid can form a zwitterionic tautomer, whereas the tautomers are common organic structures in the other two equilibria considered. After performing in-solution geometry optimizations at the IEF-PCM/B3LYP level for the three pairs of tautomers, the effects of the solvent, basis set, cavity model, and charge fitting procedure on the resulting total relative free energy were studied. Correspond-

ing geometries obtained at the IEF-PCM/B3LYP/6-31G* and 6-311++G** levels differ moderately both in tetrahydrofuran and in water solvents, but some torsion angles may deviate considerably from their gas-phase values.

Atomic charges fitted to the in-solution ELPO show small variations in different solvents, if all other calculation conditions are kept unchanged. The UA0 cavity model, or a model with explicit consideration of the polar hydrogens and applying scaled Bondi radii, has a small effect on the derived atomic charges as well. In contrast, the fitting procedure, CHELPG or RESP, has a considerable effect on the calculated values. The CHELPG fit produces more separated atomic charges than those obtained with the RESP procedure, although both derivation methods reproduce well the overall in-solution dipole moment of the selected species. The charges increase up to 20% in absolute value when the 6-311++G** rather than the 6-31G* basis set is used in the IEF-PCM/B3LYP calculations. The relative solvation free energies from FEP/MC calculations lead to differences up to about 3 kcal/mol with $\text{RCUT} = 9.75 \text{ \AA}$ in aqueous solution, using different cavity and charge derivation methods and with either 6-31G* or 6-311++G** basis sets. In contrast, $\Delta G(\text{solv})$ terms are fairly insensitive to these simulation parameters in THF. When $\text{RCUT} = 12.0 \text{ \AA}$ and a reaction field throughout the FEP/MC process are considered, the calculated $\Delta G(\text{solv})$ deviates by 1.5–3.0 kcal/mol from the values above.

ΔG_{tot} , as calculated from $\Delta E_{\text{int}} + \Delta G(\text{solv}) + \Delta G_{\text{therm}} + (\text{symmetry correction})$, strongly depends on the accepted value for ΔE_{int} . In order to predict the relative free energy for the zwitterionic isonicotinic acid tautomer in close agreement with the experimental values in aqueous solution, ΔE_{int} had to be calculated at the IEF-PCM/QCISD(T)/cc-pVTZ//IEF-PCM/B3LYP/6-31G* level. Consideration of the IEF-PCM/B3LYP/6-31G* ΔE_{int} value suffices, however, in methanol. Molecular dynamics simulations pointed out that isonicotinic acid forms a dimeric zwitterion in THF, in contrast to what happens in aqueous solution, and this structural peculiarity has been interpreted by these authors as the reason for the considerable failure of the ab initio/DFT + FEP/MC method in this particular case.

In cases of the 4-OH-pyridine and acetylacetone tautomeric systems, the calculated ΔG_{tot} values, by considering the ΔE_{int} (IEF-PCM/B3LYP/6-31G*) and 6-31G*/ $\Delta G(\text{solv})$ contributions, are close to the available experimental values both in THF and in aqueous solution. The agreement is good in cases of $\text{abs}(\Delta G_{\text{exp}}) > 2$ kcal/mol, whereas the deviation of the calculated and experimental ΔG values may amount to about 1 kcal/mol with $\text{abs}(\Delta G_{\text{exp}}) < 1$ kcal/mol.

Acknowledgment. The authors are grateful to David Case for granting them permission to use the AMBER9 program. P.I.N. thanks the Ohio Supercomputer Center for the granted computer time used for some of the QCISD(T) and CCSD(T) calculations.

Supporting Information Available: Geometric parameters optimized in solution at the IEF-PCM/B3LYP level with two basis sets (6-31G* and 6-311++G**) for the systems considered (Tables S1–S3); CHELPG (Tables S4–

S21) and RESP (Tables S22–S39) charges and relevant dipole moments in solution on the geometries optimized in solution at the IEF-PCM/B3LYP6-31G* or 6-311++G** level using UA0 and Bondi cavities (single-point IEF-PCM/B3LYP/6-311++G**//IEF-PCM/B3LYP/6-31G* results are also included); in the last line of the CHELPG tables (S4–S21), the dipole moments computed from the real densities in solution are reported; IR spectra in THF and aqueous solution for the acetylacetone tautomers with both cavities (Figures S1–S4). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Hutter, J.; Alavi, A.; Deutsch, T.; Bernasconi, M.; Goedecker, S.; Marz, D.; Tuckermann, M.; Parrinello, M. *CPMD; MPI für Festkörperforschung, and IBM Zurich Research Laboratory*: Stuttgart, Germany, 1995–1999.
- (2) (a) Laio, A.; Vondedele, J. V.; Röthlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941–6947. (b) Andreoni, W. In *3D QSAR in Drug Design: Ligand–Protein Interactions and Molecular Similarity*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, The Netherlands, 1998; Vol. 2, pp 161–167. (c) Andreoni, W.; Curioni, A.; Mordasini, T. *IBM J. Res. Dev.* **2001**, *45*, 397–407 and references therein.
- (3) (a) Bonaccorsi, R.; Petrongolo, C.; Scrocco, E.; Tomasi, J. *Theor. Chim. Acta* **1971**, *20*, 331–342. (b) Alagona, G.; Cimiraglia, R.; Scrocco, E.; Tomasi, J. *Theor. Chim. Acta* **1972**, *25*, 103–119. (c) Kollman, P. A.; Hayes, D. M. *J. Am. Chem. Soc.* **1981**, *103*, 2955–2961. (d) Warshel, A.; Lappicirella, A. *J. Am. Chem. Soc.* **1981**, *103*, 4664–4673. (e) Alagona, G.; Desmeules, P.; Ghio, C.; Kollman, P. A. *J. Am. Chem. Soc.* **1984**, *106*, 3623–3632. (f) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986**, *7*, 718–730. (g) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700–733. (h) Ferenczy, G. G.; Rivail, J.-L.; Surján, P. R.; Náráy-Szabó, G. *J. Comput. Chem.* **1992**, *13*, 830–837. (i) Gao, J.; Xia, X. *Science* **1992**, *258*, 631–635. (j) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170–1179. (k) Stanton, R. V.; Hartsough, D. S.; Merz, K. M., Jr. *J. Comput. Chem.* **1995**, *16*, 113–128.
- (4) See, for example: (a) Nagy, P. I.; Dunn, W. J., III.; Alagona, G.; Ghio, C. *J. Am. Chem. Soc.* **1991**, *113*, 6719–6729. (b) Nagy, P. I.; Dunn, W. J., III.; Alagona, G.; Ghio, C. *J. Am. Chem. Soc.* **1992**, *114*, 4752–4758. (c) Nagy, P. I.; Dunn, W. J., III.; Alagona, G.; Ghio, C. *J. Phys. Chem.* **1993**, *97*, 4628–4642. (d) Nagy, P. I.; Alagona, G.; Ghio, C. *J. Am. Chem. Soc.* **1999**, *121*, 4804–4815. (e) Nagy, P. I.; Alagona, G.; Ghio, C.; Takács-Novák, K. *J. Am. Chem. Soc.* **2003**, *125*, 2770–2785.
- (5) (a) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652. (b) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785–789.
- (6) (a) Miertus, S.; Scrocco, E.; Tomasi, J. *J. Chem. Phys.* **1981**, *55*, 117–129. (b) Miertus, S.; Tomasi, J. *J. Chem. Phys.* **1982**, *65*, 239–245. (c) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027–2094. (d) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3094.
- (7) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.
- (8) 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.
- (9) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441–451.
- (10) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968–5975.
- (11) CCSD is described in: Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910–1918. The extension to noniterative inclusion of triple excitations appears in: Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.
- (12) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (13) McQuarrie, D. A. *Statistical Mechanics*; University Science Book: Sausalito, CA, 2000.
- (14) Zwanzig, J. *J. Chem. Phys.* **1954**, *22*, 1420–1426.
- (15) Jorgensen, W. L.; Ravimohan, C. *J. Chem. Phys.* **1985**, *83*, 3050–3054.
- (16) Jorgensen, W. L. *BOSS, version 4.7*; Yale University: New Haven, CT, 2006.
- (17) (a) Jorgensen, W. L.; Madura, J. D. *J. Am. Chem. Soc.* **1983**, *105*, 1407–1413. (b) Jorgensen, W. L.; Swenson, C. J. *J. Am. Chem. Soc.* **1985**, *107*, 1489–1496. (c) Jorgensen, W. L.; Gao, J. *J. Phys. Chem.* **1986**, *90*, 2174–2182. (d) Jorgensen, W. L.; Briggs, J. M.; Contreras, M. L. *J. Phys. Chem.* **1990**, *94*, 1683–1686.
- (18) (a) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935. (b) Jorgensen, W. L.; Madura, J. D. *Mol. Phys.* **1985**, *56*, 1381–1392.
- (19) (a) Jorgensen, W. L. *J. Phys. Chem.* **1986**, *90*, 1276–1284. (b) Briggs, J. M.; Matsui, T.; Jorgensen, W. L. *J. Comput. Chem.* **1990**, *11*, 958–971.
- (20) Riddick, J. A.; Bunger, W. B.; Sakano, T. K. *Organic Solvents*; Wiley: New York, 1986.
- (21) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (22) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361–373.
- (23) (a) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280. (b) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *J. Am. Chem. Soc.* **1993**, *115*, 9620–9631.
- (24) (a) Ewald, P. P. *Ann. Phys.* **1921**, *64*, 253–287. (b) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092. (c) Essmann, U.; Perera, L.; Berkowitz, M. L.;

- Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593. (d) Crowley, M. F.; Darden, T. A.; Cheatham, T. E., III; Deerfield, D. W., II. *J. Supercomput.* **1997**, *11*, 255–278. (e) Sagui, C.; Darden, T. A. In *Simulation and Theory of Electrostatic Interactions in Solution*; Pratt, L. R., Hummer, G., Eds.; AIP: Melville, NY, 1999; pp 104–113. (f) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (25) Essex, J. W.; Jorgensen, W. L. *J. Phys. Chem.* **1995**, *99*, 17956–17962.
- (26) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman P. A. *AMBER 9*; University of California: San Francisco, CA, 2006.
- (27) See, for example: (a) Kwiatkowski, J. S.; Zielinski, T. J.; Rein, R. *Adv. Quantum. Chem.* **1986**, *18*, 85–130. (b) Cieplak, P.; Bash, P.; Singh, U. C.; Kollman, P. A. *J. Am. Chem. Soc.* **1987**, *109*, 6283–6289 and references therein. (c) Fabian, W. M. F.; Antonov, L.; Nedeltcheva, D.; Kamounah, F. S.; Taylor, P. J. *J. Phys. Chem. A* **2004**, *108*, 7603–7612. (d) Nagy, P. I.; Tejada, F. R.; Messer, W. S., Jr. *J. Phys. Chem. B* **2005**, *109*, 22588–22602. (e) Nagy, P. I.; Fabian, W. M. F. *J. Phys. Chem. B* **2006**, *110*, 25026–25032.
- (28) (a) Nagy, P. I.; Takács-Novák, K. *J. Am. Chem. Soc.* **1997**, *119*, 4999–5006. (b) Beak, P. *Acc. Chem. Res.* **1977**, *10*, 186–192 and references therein. (c) Moriyasu, M.; Kato, A.; Hashimoto, Y. *J. Chem. Soc., Perkin Trans. II* **1986**, 515–520.
- (29) 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*, 1999–2012.
- (30) (a) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87–110. (b) Li, J.; Zhu, T.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **1998**, *102*, 1820–1831. (c) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **2003**, *24*, 1291–1304.
- (31) (a) Jorgensen, W. L.; Tirado-Rives, J. *J. Comput. Chem.* **2005**, *26*, 1689–1700. (b) Jorgensen, W. L.; Tirado-Rives, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665–6670.
- (32) Kaminski, G. A.; Jorgensen, W. L. *J. Phys. Chem. B* **1998**, *102*, 1787–1796.
- (33) Udier-Blagovic, M.; Morales de Tirado, P.; Pearlman, S. A.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 1322–1332.
- (34) (a) Alagona, G.; Pullman, A.; Scrocco, E.; Tomasi, J. *Int. J. Peptide Protein Res.* **1973**, *5*, 251–259. (b) Pullman, A.; Alagona, G.; Tomasi, J. *Theor. Chim. Acta* **1974**, *33*, 87–90. (c) Ghio, C.; Scrocco, E.; Tomasi, J. In *Environmental Effects on Molecular Structure and Properties*; Pullman, B., Ed.; Reidel: Dordrecht, Holland, 1975; pp 329–342.
- (35) Alagona, G.; Ghio, C.; Nagy, P. I. *J. Chem. Theory Comput.* **2005**, *1*, 801–816.
- (36) (a) Leis, D. G.; Curran, B. C. *J. Am. Chem. Soc.* **1945**, *67*, 79–81. (b) Albert, A.; Phillips, J. N. *J. Chem. Soc.* **1956**, 1294–1304.
- (37) Ghanadzadeh Gilani, A.; Mamaghani, M.; Anbir, L. *J. Sol. Chem.* **2003**, *32*, 625–636.

CT6002252

Factors Contributing to the Accuracy of Harmonic Force Field Calculations for Water

Michael H. Cortez,[§] Nicole R. Brinkmann,[§] William F. Polik,^{*,§} Peter R. Taylor,[†]
Yannick J. Bomble,[‡] and John F. Stanton[‡]

Department of Chemistry, Hope College, Holland, Michigan 49423, Department of Chemistry, University of Warwick, Coventry CV4 7AL, United Kingdom, and Institute for Theoretical Chemistry, Departments of Chemistry and Biochemistry, The University of Texas at Austin, Austin, Texas 78712

Received November 29, 2006

Abstract: An analysis of the major factors affecting the accuracy of harmonic force field computations of water is presented. By systematically varying the level of approximation in the basis set, treatment of electron correlation, core electron correlation, and relativistic correction, the underlying sources of error in the computation of harmonic vibrational frequencies for water were quantified. The convergence error due to wavefunction description with a cc-pVQZ basis set in the absence of electron correlation was 1.6 cm^{-1} , as determined from extending the Hartree–Fock computations to larger basis sets. The convergence error due to neglecting higher-order electronic correlation terms than are included at the CCSD(T) level using the cc-pVTZ basis set was estimated to be 4.7 cm^{-1} , as determined from frequency calculations up to CCSDTQ for water and literature results up to CCSDTQP for diatomic molecules. The convergence error due to omitting higher-order diffuse functions than included in aug-cc-pVQZ was found to be 3.7 cm^{-1} , as determined by adding more diffuse functions in larger basis sets. The error associated with neglecting core electron correlation effects (i.e., “freezing” core electrons) was 5.0 cm^{-1} and with neglecting relativistic effects was 2.2 cm^{-1} . Due to a cancellation among these various sources of error, the harmonic frequencies for H₂O computed using the CCSD(T)/aug-cc-pVQZ model chemistry were on average within 2 cm^{-1} of experimentally inferred vibrational frequencies.

Introduction

The infrared absorption spectrum of H₂O is important in a variety of applications. H₂O is the third most abundant gaseous species in the universe,^{1,2} plays a critical role in the earth’s atmospheric chemistry and radiation trapping,^{3–5} and is of great astrophysical interest.^{6–10} Even though the vibrational transitions of water have been studied both theoretically and experimentally,^{1,2,7,11} currently reported spectra of H₂O contain many unassigned lines and do not

continuously cover the entire spectrum at high resolution.^{6,7} Portions of the spectrum of H₂O have been precisely determined experimentally through a variety of techniques over small ranges. Cavity ringdown spectroscopy has been used to study the vibrational spectra of H₂O at 555–604 nm and 810–820 nm vibrational overtone transitions in atmospheric flames.^{1,12} Intracavity laser spectroscopy has been used to study the absorption of H₂O near 795 nm.¹¹ The vibrational overtone spectra of H₂O in the near-infrared, visible, and near-ultraviolet spectrum were also studied by Carleer et al.² using Fourier transform spectroscopy.

The potential energy surface (PES) of water has been computed both theoretically and from experimental data to aid in the construction of an absorption spectrum of

* Corresponding author e-mail: polik@hope.edu.

[§] Hope College.

[†] University of Warwick.

[‡] The University of Texas at Austin.

spectroscopic accuracy for H₂O. Both Jensen¹³ and Polyansky et al.¹⁴ constructed PESs for H₂O solely from available experimental data. Beardsworth et al.¹⁵ used a nonrigid bender Hamiltonian program to study the rotational–vibrational energy levels of triatomic molecules, including H₂O. Polyansky et al.¹⁰ computed the PES of H₂O using MRCI theory and the aug-cc-pVXZ basis sets (X = T, Q, 5, 6). Császár and Mills¹⁶ determined the quartic and sextic force field parameters for H₂O using CCSD(T)/aug-cc-pVXZ (X = T, Q). Pair potentials for the H₂O dimer have also been computed using symmetry adapted perturbation theory.^{17,18}

The quantification of a PES is an important yet difficult practice in chemistry. Chemical reactions may be modeled as occurring on potential energy surfaces, and with the knowledge of the PES of a molecule, thermodynamic stability, reactivity, and reaction pathways can be predicted before an experiment is performed. The conventional method of modeling a PES is to construct a Taylor series expansion in terms of the displacement from the equilibrium geometry. This PES may then be used to calculate vibrational energy values.¹⁹ Since the harmonic term of the Taylor expansion is a good approximation only at small displacements, higher-order anharmonic corrections are needed away from the equilibrium point.

Equation 1 is a Taylor expansion of the PES about the equilibrium point r_e

$$\begin{aligned}
 U(r) &= \sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{\partial^i U}{\partial r^i} \right)_{r=r_e} (r - r_e)^i \\
 &= U(r_e) + \left(\frac{\partial U}{\partial r} \right)_{r=r_e} (r - r_e) + \frac{1}{2} \left(\frac{\partial^2 U}{\partial r^2} \right)_{r=r_e} (r - r_e)^2 + \\
 &\quad \sum_{i=3}^{\infty} \frac{1}{i!} \left(\frac{\partial^i U}{\partial r^i} \right)_{r=r_e} (r - r_e)^i + \dots \quad (1)
 \end{aligned}$$

where $r - r_e$ is the displacement from the equilibrium point, and U is the potential energy. Assigning the zero of potential energy to the equilibrium structure, the expansion about a minimum of the PES [$(\partial U/\partial r)_{r=r_e} = 0$] results in the quadratic term being the first nonzero term of the expansion. The quadratic term in eq 1 represents the harmonic energy, and the summation of subsequent terms represents the anharmonic corrections.

The harmonic terms of the Taylor expansion dominate the total energy near the equilibrium point, making it critical to compute the harmonic terms accurately. Császár and Mills¹⁶ have observed that the harmonic force constants converge slower with respect to basis set size and theory level than the anharmonic constants, requiring more computational time to obtain the same accuracy. Thus, it is important to understand and determine the accuracy with which the harmonic force constants can be computed.

Previous studies^{10,16,20,21} have examined the factors affecting the accuracy of theoretically computed force constants. Császár et al.^{16,20} found that the majority of the error in their quadratic force constants for water at the CCSD(T)/aug-cc-pVQZ level of theory was due to core-core and core-valence correlation. They also found that the inclusion of relativistic

effects had marginal effects on the computed force constants. Polyansky et al.¹⁰ examined H₂O using the aug-cc-pVXZ (XD = Q, 5, 6) basis sets and MRCI theory and found the neglect of core electron correlation for oxygen resulted in 19 cm⁻¹ residual error in computed vibrational band origins. Partridge and Schwenke²¹ studied the effects of core electron correlation of H₂O at varying levels of theory and basis set size and found that corrections in core electron correlation are insensitive to increases in basis set size with the use of basis sets larger than the aug-cc-pVQZ basis set.

In the literature, only effects of individual factors on the accuracy of computed force constants have been reported. This work presents a comprehensive study of the major factors that affect the accuracy of such computations. By doing so, the underlying sources of error in theoretical computations of PESs can be evaluated and quantified at varying levels of approximation. In addition, insight is gained into the circumstances under which cancellation of errors may be present and the levels of theory and basis set sizes required to compute force constants to a specific accuracy.

The equilibrium geometry of a molecule serves only as the minimum energy reference point to the PES, while harmonic frequencies depend on the precise curvature of the PES near the equilibrium structure. Hence, the computation of vibrational frequencies is a more stringent test of the computational methods. Since the quadratic term is the largest in size and converges the slowest, it is the most important term of the Taylor expansion about the equilibrium geometry and will be the only term computed in this study.

Although diatomic molecules are computationally efficient species to study, they are missing features that yield greater insight into the intricacies of the computations and therefore would limit the applicability of results to arbitrary polyatomic molecules. H₂O, therefore, was chosen for this study. As a polyatomic molecule, it has more than one vibrational degree of freedom and has both stretching and bending vibrations. H₂O also exhibits anharmonicities and resonances, although these are not explicitly considered in the present work. H₂O is also composed of light atoms, making the use of higher methods and larger basis sets feasible. Finally, H₂O is a well-studied molecule both theoretically and experimentally, thus allowing for sufficient data comparison.

Computational Methods

All computations were carried out using the MOLPRO²² and ACES II²³ ab initio programs, except for the calculations including quadruple excitations (CCSDTQ) which were done with the string-based many-body program of Kállay.²⁴ The computations were performed on a Linux-based cluster of IA32 computers (2.4 GHz Pentium 4 CPU, 1GB RAM, 120 GB disk) and AMD64 computers (dual 2.2 GHz Opteron 248 CPUs, 8GB RAM, 250 GB disk).

Reference wavefunctions for the ground state of H₂O were calculated at the Hartree–Fock self-consistent field (HF–SCF) level of theory. Dynamical correlation effects were included using the coupled-cluster series, including all single and double (CCSD)²⁵ and perturbatively applied triple excitations [CCSD(T)].^{26,27} Explicit computation of the full set of triple excitations (CCSDT),^{28,29} perturbatively applied

Table 1. Convergence of the Hartree–Fock Wavefunction for Computed Equilibrium Geometry and Harmonic Frequencies^a

method	basis	r_e	θ_e	ω_1	ω_2	ω_3	$\langle \text{error} \rangle$	$\langle \Delta \rangle$	energy
HF	cc-pVDZ	0.946287	104.6130	4113.47	1775.69	4211.79	225.92		-76.02705
HF	cc-pVTZ	0.940604	106.0016	4126.74	1752.89	4226.63	227.69	16.97	-76.05777
HF	cc-pVQZ	0.939601	106.2222	4129.84	1750.47	4229.14	228.75	2.68	-76.06552
HF	cc-pV5Z	0.939572	106.3280	4130.26	1748.19	4230.64	228.63	1.40	-76.06778
HF	cc-pV6Z	0.939582	106.3361	4129.97	1748.03	4230.51	228.44	0.19	-76.06810
experiment ³⁸		0.9572	104.52	3832.2	1648.5	3942.5			

^a $\langle|\text{error}|\rangle$ represents the average absolute difference between the experimentally inferred and computed harmonic vibrational frequencies. $\langle|\Delta|\rangle$ represents the average absolute difference in values from the previous set of computed harmonic vibrational frequencies. Bond lengths are in Å; bond angles are in degrees; frequencies, $\langle|\text{error}|\rangle$, and $\langle|\Delta|\rangle$ are in cm^{-1} ; and total electronic energy is in E_h .

quadruple excitations [CCSDT(Q)],³⁰ and the full set of quadruple excitations (CCSDTQ)^{31,32} were also carried out where feasible. Relativistic effects were analyzed using the Cowan–Griffin³³ (CG) and Douglas–Kroll³⁴ (DK) methods.

Three families of basis sets were used in the study. The first was Dunning’s correlation-consistent polarized valence basis (cc-pVXZ)³⁵ sets. For H₂O, the number of basis functions in these basis sets ranged from 24 for cc-pVDZ to 322 for cc-pV6Z. The more extensive augmented correlation-consistent polarized valence basis (aug-cc-pVXZ)^{35,36} sets (45–443 basis functions for X = D to 6) and the augmented correlation-consistent polarized valence with core basis (aug-cc-pCVXZ)^{35–37} sets (45–341 basis functions for X = D to 5) were also used.

Each computation included a geometry optimization performed at the respective level of theory and basis set size for the computation. In the core electron correlation computations, all electrons are considered in post-SCF calculations. In all other coupled-cluster computations, the correlation of core electrons was neglected. Vibrational frequencies were computed using finite differences for all computations.

Results and Discussion

A. Hartree–Fock Wavefunction. The convergence of the Hartree–Fock (HF) wavefunction was determined by comparing vibrational frequencies of H₂O computed at the HF level of theory with basis sets from the cc-pVXZ family. The computed equilibrium geometries and harmonic vibrational frequencies are compared to the experimentally inferred equilibrium geometries and harmonic frequencies of Pliva, Spirko, and Papousek³⁸ in Table 1. The three vibrational modes for H₂O are the symmetric stretch (ω_1), the bend (ω_2), and the antisymmetric stretch (ω_3). r_e and θ_e are the respective equilibrium bond length for the H–O bond and the equilibrium H–O–H angle for the molecule. Computed vibrational frequencies are compared to experimentally inferred frequencies via an average absolute difference between the two sets of values denoted $\langle|\text{error}|\rangle$. Because experimentally inferred frequencies involve fitting data to an assumed Hamiltonian model, some caution is warranted when interpreting $\langle|\text{error}|\rangle$. For example, McCoy and Sibert³⁹ have demonstrated that harmonic frequencies of water change by approximately 2 cm^{-1} when using an eighth-order effective Hamiltonian rather than a second-order effective Hamiltonian. The convergence of computed vibrational frequencies is determined by the convergence of the average absolute difference between adjacent sets of vibrational frequencies denoted $\langle|\Delta|\rangle$. Decreasing values of $\langle|\Delta|\rangle$

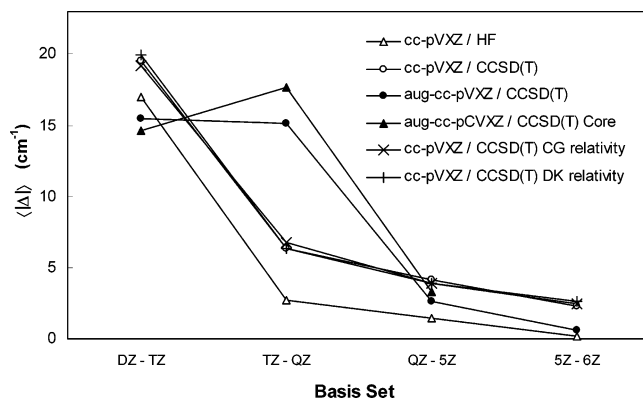


Figure 1. Average absolute difference ($\langle|\Delta|\rangle$) between computed vibrational frequencies with increasing basis set size. Computations were performed at the varying levels of theory.

for subsequent computations illustrate the convergence of the vibrational frequencies and consequently the convergence of the error associated with the approximation being varied.

When computing energies with the HF method, each electron is assumed to see an averaged distribution of the other electrons. As seen in Table 1, the exclusion of instantaneous electron correlation results in vibrational frequencies approximately 228 cm^{-1} in error of experimental frequencies (cc-pVQZ $\langle|\text{error}|\rangle = 228.75 \text{ cm}^{-1}$, cc-pV5Z $\langle|\text{error}|\rangle = 228.63 \text{ cm}^{-1}$, cc-pV6Z $\langle|\text{error}|\rangle = 228.44 \text{ cm}^{-1}$). As also seen in Figure 1 and Table 1, the average absolute difference, $\langle|\Delta|\rangle$, between vibrational frequencies computed with the cc-pVQZ and cc-pV5Z and between frequencies computed with the cc-pV5Z and cc-pV6Z basis sets is within 1.4 cm^{-1} and 0.2 cm^{-1} , respectively. Consequently, the computed values and the error associated with the HF wavefunction are converged to within 1.6 cm^{-1} of the limiting value with the use of the cc-pVQZ basis set ($\langle|\Delta|\rangle_{\text{QZ-5Z}} + \langle|\Delta|\rangle_{\text{5Z-6Z}} = 1.40 \text{ cm}^{-1} + 0.19 \text{ cm}^{-1} < 1.6 \text{ cm}^{-1}$). Therefore, use of basis sets of at least quadruple- ζ quality should be suitable for high accuracy harmonic frequency computations.

B. Electron Correlation Level. The effect of approximating the electron correlation for water was determined through the comparison of vibrational frequencies computed with theories from the coupled-cluster series, which include a systematic increase of electron correlation. The cc-pVTZ basis set was used since the average absolute difference between the vibrational frequencies computed with the HF/cc-pVTZ and HF/cc-pV6Z methods is less than 5 cm^{-1} , and the former is 1000 times faster than the latter. To include

Table 2. Convergence of Electron Correlation Treatment for Computed Equilibrium Geometry and Harmonic Vibrational Frequencies Computed Using the cc-pVTZ Basis Set^a

method	basis	r_e	θ_e	ω_1	ω_2	ω_3	$\langle \text{error} \rangle$	$\langle \Delta \rangle$	energy
HF	cc-pVTZ	0.940602	106.0016	4127.04	1753.02	4226.94	227.93		-76.05777
CCSD	cc-pVTZ	0.957118	103.8928	3875.94	1678.47	3979.07	36.76	191.17	-76.32456
CCSD(T)	cc-pVTZ	0.959426	103.5821	3840.93	1668.88	3945.54	10.72	26.04	-76.33222
CCSDT	cc-pVTZ	0.959390	103.5906	3841.36	1669.20	3945.31	10.89	0.33	-76.33229
CCSDT(Q)	cc-pVTZ	0.959722	103.5533	3835.06	1668.03	3940.06	8.28	4.24	-76.33265
CCSDTQ	cc-pVTZ	0.959677	103.5581	3835.94	1668.22	3940.73	8.41	0.58	-76.33261
experiment ³⁸		0.9572	104.52	3832.2	1648.5	3942.5			

^a $\langle|\text{error}|\rangle$, $\langle|\Delta|\rangle$, and units are as described in Table 1. In contrast to Tables 1 and 4–6, all calculations were performed with the ACES II and Kállay programs.

Table 3. Average Absolute Differences $\langle|\Delta|\rangle$ of Harmonic Vibrational Frequencies Using Various Methods and the cc-pVTZ Basis Set^a

$ \Delta $	HF	CCSD	CCSD(T)	CCSDT	CCSDT(Q)	CCSDTQ
HF	–					
CCSD	191.17	–				
CCSD(T)	217.22	26.04	–			
CCSDT	217.04	25.87	0.33	–		
CCSDT(Q)	221.28	30.11	4.07	4.24	–	
CCSDTQ	220.70	29.53	3.49	3.66	0.58	–

^a Units are cm^{-1} .

the CCSDT, CCSDT(Q), and CCSDTQ computations and to prevent the introduction of extraneous sources of error, all computations for this specific analysis were performed with the ACES II and Kállay programs.^{23,24} Table 2 compares the computed equilibrium geometries and harmonic frequencies to experimentally inferred values.³⁸

Table 3 describes the convergence of computed harmonic frequencies with the cc-pVTZ basis as a function of coupled-cluster method. The increase in theory from CCSD to CCSDT results in a 26 cm^{-1} average absolute difference in vibrational frequencies ($\langle|\Delta|\rangle_{\text{CCSD}-\text{CCSDT}} = 25.87 \text{ cm}^{-1}$) for water, while the increase from CCSDT to CCSDTQ results in only a 3.66 cm^{-1} difference. Ruden et al.⁴⁰ studied four diatomic molecules (HF, N_2 , F_2 , CO) and found that the average absolute difference in vibrational frequency computed with the cc-pVTZ basis set from CCSD to CCSDT was 65.7 cm^{-1} and from CCSDT to CCSDTQ was 10.9 cm^{-1} . Similar convergence trends were observed with the cc-pVDZ basis set, and the average convergence for these diatomics from CCSDTQ to CCSDTQP was computed to be only 1.2 cm^{-1} . Comparison of these results indicates that the harmonic frequencies of water converge more rapidly with method than these diatomic molecules, suggesting that little improvement in harmonic frequencies would be gained from a CCSDTQP calculation of water. It should be noted that inferences regarding convergence of vibrational frequencies that are based on cc-pVTZ calculations need to be tempered slightly for larger basis sets, as increasing the size of the basis set generally magnifies correlation effects.

Increasing theory from CCSD(T) to CCSDT to fully treat triple excitations results in only a 0.33 cm^{-1} difference in harmonic frequencies ($\langle|\Delta|\rangle_{\text{CCSD(T)}-\text{CCSDT}} = 0.33 \text{ cm}^{-1}$) for the triatomic water molecule. This is consistent with the conclusions of Feller and Sordo,⁴¹ who found no significant difference between the two theories when studying 13

diatomic hybrids, and with the previously mentioned study of Ruden et al. Similarly, the difference between CCSDT(Q) and CCSDTQ is only 0.58 cm^{-1} , again demonstrating that perturbative treatment of the next higher connected-excitation level is a very effective way to reduce computation time with minimal loss of accuracy.

The remaining average absolute error in harmonic frequencies due to electron correlation associated with the CCSD(T) level is estimated at 4.7 cm^{-1} , which arises from the computed difference between CCSD(T) and CCSDTQ, $\langle|\Delta|\rangle_{\text{CCSD(T)}-\text{CCSDTQ}} = 3.49 \text{ cm}^{-1}$, and an estimate from the diatomic data of 1.2 cm^{-1} for the remaining error.

C. Valence Electron Description. The convergence of the correlation consistent wavefunction was determined via the comparison of vibrational frequencies computed at the CCSD(T) level of theory with basis sets from the cc-pVXZ and aug-cc-pVXZ families. While the diffuse functions were developed specifically to provide increased basis set flexibility for charged species, they are often employed for neutral systems in order to yield more accurate results. Table 4 compares the computed vibrational frequencies to the experimentally inferred harmonic frequencies, $\langle|\text{error}|\rangle$, of Pliva et al.³⁸

As seen in Figure 2, the cc-pVXZ computations are initially more accurate than their augmented counterparts (cc-pVDZ $\langle|\text{error}|\rangle = 22.59 \text{ cm}^{-1}$, aug-cc-pVDZ $\langle|\text{error}|\rangle = 31.29 \text{ cm}^{-1}$, cc-pVTZ $\langle|\text{error}|\rangle = 10.48 \text{ cm}^{-1}$, aug-cc-pVTZ $\langle|\text{error}|\rangle = 15.81 \text{ cm}^{-1}$). With the use of the two largest ($X = 5, 6$) augmented basis sets though, the computed vibrational frequencies are more accurate than the largest cc-pVXZ basis set (cc-pV6Z $\langle|\text{error}|\rangle = 3.98 \text{ cm}^{-1}$, aug-cc-pV5Z and aug-cc-pV6Z $\langle|\text{error}|\rangle < 2 \text{ cm}^{-1}$), supporting the findings of Martin and Taylor⁴² who reported that augmented basis sets yield more accurate harmonic frequencies than nonaugmented basis sets for HF and H_2O . In addition, the augmented set, although initially converging slower, converges closer to the basis set limit than the nonaugmented set (cc-pVXZ family $\langle|\Delta|\rangle_{\text{QZ}-5Z} = 4.14 \text{ cm}^{-1}$ and $\langle|\Delta|\rangle_{5Z-6Z} = 2.31 \text{ cm}^{-1}$, aug-cc-pVXZ family $\langle|\Delta|\rangle_{\text{QZ}-5Z} = 2.64$ and $\langle|\Delta|\rangle_{5Z-6Z} = 0.56 \text{ cm}^{-1}$), as seen in Figure 1. From their study of diatomics, Ruden et al.⁴⁰ estimated the remaining basis set error beyond the aug-cc-pV6Z basis set to be conservatively within 0.5 cm^{-1} . Thus, with the use of the aug-cc-pVQZ basis set, the error associated with the valence electrons has converged within 3.7 cm^{-1} ($\langle|\Delta|\rangle_{\text{QZ}-5Z} + \langle|\Delta|\rangle_{5Z-6Z} + \langle|\Delta|\rangle_{6Z-\infty} = 2.64 \text{ cm}^{-1} + 0.55 \text{ cm}^{-1} + 0.5 \text{ cm}^{-1} = 3.7 \text{ cm}^{-1}$).

Table 4. Convergence of Valence Electron Description for Computed Equilibrium Geometry and Harmonic Vibrational Frequencies Computed Using the CCSD(T) Theory^a

method	basis	r_e	θ_e	ω_1	ω_2	ω_3	$\langle \text{error} \rangle$	$\langle \Delta \rangle$	energy
CCSD(T)	cc-pVDZ	0.966280	101.9127	3821.33	1690.19	3927.29	22.59		-76.24131
CCSD(T)	cc-pVTZ	0.959428	103.5821	3840.65	1668.76	3945.24	10.48	19.57	-76.33222
CCSD(T)	cc-pVQZ	0.957891	104.1159	3844.19	1659.17	3951.11	10.42	6.33	-76.35980
CCSD(T)	cc-pV5Z	0.958041	104.3723	3839.78	1653.24	3949.03	6.28	4.14	-76.36904
CCSD(T)	cc-pV6Z	0.958181	104.4221	3837.00	1651.20	3946.93	3.98	2.31	-76.37202
CCSD(T)	aug-cc-pVDZ	0.966514	103.9366	3786.66	1638.08	3904.59	31.29		-76.27390
CCSD(T)	aug-cc-pVTZ	0.961581	104.1796	3810.41	1645.62	3919.75	15.81	15.48	-76.34233
CCSD(T)	aug-cc-pVQZ	0.958931	104.3646	3830.81	1649.97	3940.39	1.66	15.13	-76.36359
CCSD(T)	aug-cc-pV5Z	0.958416	104.4273	3834.37	1649.97	3944.74	1.96	2.64	-76.37030
CCSD(T)	aug-cc-pV6Z	0.958344	104.4472	3834.70	1649.22	3945.32	2.01	0.56	-76.37256
experiment ³⁸		0.9572	104.52	3832.2	1648.5	3942.5			

^a $\langle|\text{error}|\rangle$, $\langle|\Delta|\rangle$, and units are as described in Table 1.

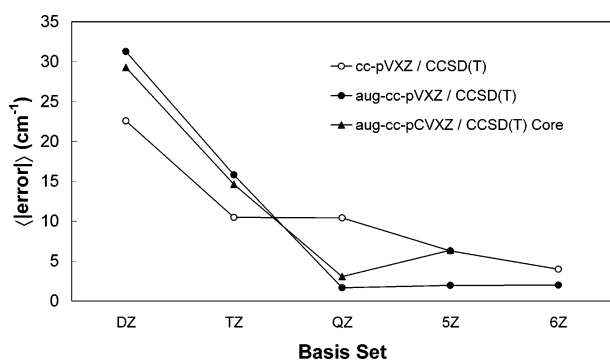


Figure 2. Average absolute difference ($\langle|\text{error}|\rangle$) between experimentally inferred harmonic frequencies and vibrational frequencies computed using CCSD(T) and increasing basis sets. All electrons were correlated in post-SCF calculations in the CCSD(T) core computation. Correlation of the core electrons was neglected in the other computations.

Comparing computed frequencies for each vibrational mode in Table 4 indicates that the three vibrational modes converge at different rates. As observed by Martin and Taylor,⁴² for both the augmented and the nonaugmented basis set families, the vibrational mode corresponding to the symmetric stretch converges faster than the bend or anti-symmetric stretch.

D. Core Electron Correlation. The effect of neglecting core electron correlation effects was determined by first characterizing the effect of using basis sets from the aug-cc-pCVXZ family and then characterizing the effect of neglecting core-core and core-valence correlation. The first comparison was made via the average absolute differences, $\langle|\text{diff}|\rangle$, between frequencies computed with the aug-cc-pVXZ and aug-cc-pCVXZ families. The contribution due to correlating the core electrons was determined by calculating the average absolute difference, $\langle|\text{diff}|\rangle$, between frequencies computed without correlating core electrons (i.e., “frozen core”) [aug-cc-pVXZ/CCSD(T)] and frequencies computed with correlated core electrons (i.e., “all electron”) [aug-cc-pCVXZ/CCSD(T);core]. The convergence of the all electron frequencies, $\langle|\Delta|\rangle$, and their error from experimentally inferred frequencies, $\langle|\text{error}|\rangle$, was also determined. All values listed above are included in Table 5.

The difference between the frozen core electron computations using basis sets from the aug-cc-pVXZ and aug-cc-

pCVXZ families was less than 1 cm^{-1} for most of the computations ($X = \text{D}$ $\langle|\text{diff}|\rangle = 0.36 \text{ cm}^{-1}$, $X = \text{Q}$ $\langle|\text{diff}|\rangle = 0.56 \text{ cm}^{-1}$, $X = 5$ $\langle|\text{diff}|\rangle = 0.13 \text{ cm}^{-1}$). The change in basis from the aug-cc-pVXZ to the aug-cc-pCVXZ family include the addition of functions that do not describe valence correlation, and hence there is little effect on the computed vibrational frequencies.

The difference between the frozen core and the all electron computations converges to 5.0 cm^{-1} with the use the aug-cc-pCVQZ basis set (aug-cc-pCVQZ;core $\langle|\text{diff}|\rangle = 4.72 \text{ cm}^{-1}$, aug-cc-pCV5Z;core $\langle|\text{diff}|\rangle = 4.99 \text{ cm}^{-1}$). This independence of basis set when using basis sets of at least quadruple- ζ quality is consistent with the previous studies^{21,40–44} of the contribution from the correlation of core electrons on harmonic frequencies. Since the frequencies from the frozen core and all electron computations converge at virtually the same rate [aug-cc-pCVXZ/CCSD(T) $\langle|\Delta|\rangle_{\text{TZ-QZ}} = 17.13 \text{ cm}^{-1}$, aug-cc-pCVXZ/CCSD(T);core $\langle|\Delta|\rangle_{\text{TZ-QZ}} = 17.65 \text{ cm}^{-1}$, aug-cc-pCVXZ/CCSD(T) $\langle|\Delta|\rangle_{\text{QZ-5Z}} = 3.17 \text{ cm}^{-1}$, aug-cc-pCVXZ/CCSD(T);core $\langle|\Delta|\rangle_{\text{QZ-5Z}} = 3.34 \text{ cm}^{-1}$], the error associated with neglecting core electron correlation is predicted to be 5.0 cm^{-1} .

As seen in Figure 2, when using large basis sets, the error from experiment, $\langle|\text{error}|\rangle$, of the vibrational frequencies computed with all electrons correlated is greater than the error of the frequencies computed with frozen core electrons (aug-cc-pCVQZ;core 3.06 cm^{-1} , aug-cc-pVQZ 1.66 cm^{-1} , aug-cc-pCV5Z;core 6.37 cm^{-1} , aug-cc-pV5Z 1.96 cm^{-1}). The greater accuracy of the computations with frozen core electrons is presumably due to a cancellation of errors between core electron correlation and inadequacies in correlation treatment.⁴² As the error associated with neglecting core electron correlation effects is increasingly accounted for, the other errors associated with the computation become observable. Ruden et al.⁴⁰ observed that core correlation computations at the CCSD(T) theory overestimate harmonic frequencies in diatomic molecules. In our study, the correlated core computations significantly overestimate the frequencies of the symmetric and antisymmetric stretches. Previous studies^{42,43} suggest this error in the frequency computations of H₂O and diatomics is due to n-particle space imperfections and contraction errors. The combination of a 30% increase in computational time and a decrease in accuracy when core electrons are correlated results in the

Table 5. Magnitude of the Core Electron Correlation Effects at CCSD(T) on Computed Equilibrium Geometry and Harmonic Vibrational Frequencies^a

method	basis	r_e	θ_e	ω_1	ω_2	ω_3	$\langle \text{error} \rangle$	$\langle \text{diff} \rangle$	$\langle \Delta \rangle$	energy
CCSD(T)	aug-cc-pVDZ	0.966514	103.9366	3786.66	1638.08	3904.59	31.29			-76.27390
CCSD(T)	aug-cc-pVTZ	0.961581	104.1796	3810.41	1645.62	3919.75	15.81		15.48	-76.34233
CCSD(T)	aug-cc-pVQZ	0.958931	104.3646	3830.81	1649.97	3940.39	1.66		15.13	-76.36359
CCSD(T)	aug-cc-pV5Z	0.958416	104.4273	3834.37	1649.97	3944.74	1.96		2.64	-76.37030
CCSD(T)	aug-cc-pCVDZ	0.966347	103.9175	3786.43	1638.65	3904.32	31.27	0.36		-76.27686
CCSD(T)	aug-cc-pCVTZ	0.961330	104.1912	3807.57	1645.91	3914.62	18.37	2.75	12.90	-76.34551
CCSD(T)	aug-cc-pCVQZ	0.958969	104.3669	3830.30	1649.07	3940.13	1.61	0.56	17.13	-76.36498
CCSD(T)	aug-cc-pCV5Z	0.958403	104.4296	3834.45	1649.74	3944.82	1.94	0.13	3.17	-76.37092
CCSD(T);core	aug-cc-pCVDZ	0.965881	103.9527	3789.06	1639.13	3907.21	29.27	2.02		-76.31517
CCSD(T);core	aug-cc-pCVTZ	0.960572	104.2891	3813.29	1645.40	3920.66	14.62	1.34	14.65	-76.39966
CCSD(T);core	aug-cc-pCVQZ	0.958098	104.4805	3836.80	1648.46	3947.05	3.06	4.72	17.65	-76.42464
CCSD(T);core	aug-cc-pCV5Z	0.957501	104.5485	3841.21	1649.11	3952.00	6.37	4.99	3.34	-76.43227
experiment ³⁸		0.9572	104.52	3832.2	1648.5	3942.5				

^a $\langle|\text{error}|\rangle$, $\langle|\Delta|\rangle$, and units are as described in Table 1. $\langle|\text{diff}|\rangle$, in cm^{-1} , represents the average absolute difference between the vibrational frequencies for each computation and the corresponding aug-cc-pVXZ computation.

Table 6. Magnitude of the Relativistic Correction at CCSD(T) for the Equilibrium Geometry and Harmonic Vibrational Frequencies Computed Using the Cowan–Griffin and Douglas–Kroll Relativistic Correction Methods^a

method	rel	basis	r_e	θ_e	ω_1	ω_2	ω_3	$\langle \text{error} \rangle$	$\langle \Delta \rangle$	$\langle \text{diff} \rangle$	energy
CCSD(T)		cc-pVDZ	0.966280	101.9127	3821.33	1690.19	3927.29	22.59			-76.24131
CCSD(T)		cc-pVTZ	0.959428	103.5821	3840.65	1668.76	3945.24	10.48	19.57		-76.33222
CCSD(T)		cc-pVQZ	0.957891	104.1159	3844.19	1659.17	3951.11	10.42	6.33		-76.35980
CCSD(T)		cc-pV5Z	0.958041	104.3723	3839.78	1653.24	3949.03	6.28	4.14		-76.36904
CCSD(T)		cc-pV6Z	0.958181	104.4221	3837.00	1651.20	3946.93	3.98	2.31		-76.37202
CCSD(T)	CG	cc-pVDZ	0.966222	101.8617	3818.34	1690.70	3924.24	24.77		2.18	-76.29280
CCSD(T)	CG	cc-pVTZ	0.959407	103.5122	3837.80	1670.27	3941.91	9.32	19.19	2.56	-76.38380
CCSD(T)	CG	cc-pVQZ	0.957942	104.0503	3841.62	1660.41	3948.46	9.10	6.74	2.15	-76.41153
CCSD(T)	CG	cc-pV5Z	0.958067	104.3095	3837.61	1654.49	3946.73	5.21	3.89	1.91	-76.42078
CCSD(T)	CG	cc-pV6Z	0.958200	104.3602	3834.69	1652.38	3944.49	2.79	2.42	1.98	-76.42382
CCSD(T)	DK	cc-pVDZ	0.966269	101.8564	3817.82	1690.69	3923.78	25.10		2.51	-76.28939
CCSD(T)	DK	cc-pVTZ	0.959341	103.5163	3838.50	1670.26	3942.55	9.37	19.96	2.11	-76.38064
CCSD(T)	DK	cc-pVQZ	0.957938	104.0529	3841.60	1660.35	3948.47	9.07	6.31	2.14	-76.40820
CCSD(T)	DK	cc-pV5Z	0.958071	104.3119	3837.65	1654.22	3946.84	5.17	3.90	1.77	-76.41745
CCSD(T)	DK	cc-pV6Z	0.958195	104.3637	3834.39	1652.37	3944.08	2.55	2.62	2.21	-76.42055
experiment ³⁸			0.9572	104.52	3832.2	1648.5	3942.5				

^a $\langle|\text{error}|\rangle$, $\langle|\Delta|\rangle$, $\langle|\text{diff}|\rangle$, and units are as described in Table 5.

conclusion that core electron correlation should be included for only the most rigorous computations.

E. Relativistic Effects. The error associated with neglecting relativistic effects was determined by comparing the vibrational frequencies from noncorrected computations to frequencies corrected using either the Cowan–Griffin (CG) method or the Douglas–Kroll (DK) method. The CG approach uses first-order perturbation theory to calculate expectation values for one-electron Darwin and mass-velocity integrals. The DK method performs a free-particle transformation on the Dirac Hamiltonian to produce a no-pair DK operator.

The convergence of the relativistically corrected vibrational frequencies, $\langle|\Delta|\rangle$, and the error from the experimentally inferred values of Pliva et al.,³⁸ $\langle|\text{error}|\rangle$, are found in Table 6. The relativistically corrected vibrational frequencies are also compared to the noncorrected frequencies by means of the average absolute difference between the two sets of values, $\langle|\text{diff}|\rangle$.

As seen in Figure 3, the difference between the corrected and experimental frequencies is less than the difference

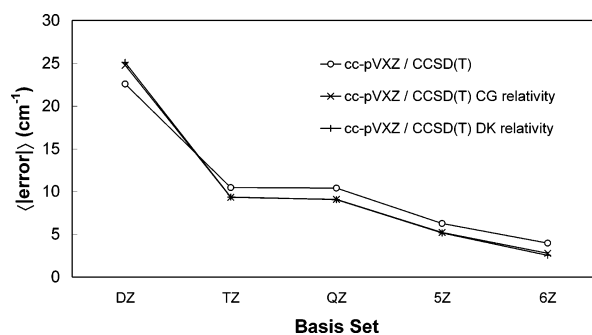


Figure 3. Average absolute difference ($\langle|\text{error}|\rangle$) between experimentally inferred harmonic frequencies and relativistically corrected vibrational frequencies computed using CCSD(T) with increasing basis sets. CG computations were performed using the Cowan–Griffin method. DK computations were performed using the Douglas–Kroll method.

between the noncorrected experimental frequencies ($\langle|\text{error}|\rangle = 6.28 \text{ cm}^{-1}$ and 3.98 cm^{-1} for cc-pV5Z and cc-pV6Z, respectively; $\langle|\text{error}|\rangle = 5.21 \text{ cm}^{-1}$ and 2.79 cm^{-1} for cc-pV5Z/CG and cc-pV6Z/CG, respectively; $\langle|\text{error}|\rangle = 5.17$

Table 7. Comparison of Contributing Factors of Uncertainty^a

factor	cc-pVTZ/ CCSD(T)	aug-cc-pVQZ/ CCSD(T)
HF wavefunction	4.3	1.6
electron correlation method	4.7	4.7
valence electron description	12.8	3.7
core electron correlation	5.0	5.0
relativistic effects	2.2	2.2
RSS	15.3	8.3
$\langle \text{error} \rangle$	10.5	1.7

^a RSS represents the square root of the sum of the squares of the contributing factors, which estimates the expected error in computed vibrational frequency. $\langle \text{error} \rangle$ represents the average absolute difference between the experimentally inferred³⁸ and computed harmonic vibrational frequencies, which estimates the actual error. All units are cm^{-1} .

cm^{-1} and 2.55 cm^{-1} for cc-pV5Z/DK and cc-pV6Z/DK, respectively). Also, the convergence of the computed frequencies with respect to changes in the basis set is approximately the same for all three sets of computations [for CCSD(T): $\langle |\Delta| \rangle_{\text{QZ-5Z}} = 4.14 \text{ cm}^{-1}$, $\langle |\Delta| \rangle_{\text{5Z-6Z}} = 2.31 \text{ cm}^{-1}$; for CCSD(T)/CG: $\langle |\Delta| \rangle_{\text{QZ-5Z}} = 3.98 \text{ cm}^{-1}$, $\langle |\Delta| \rangle_{\text{5Z-6Z}} = 2.42 \text{ cm}^{-1}$; for CCSD(T)/CG: $\langle |\Delta| \rangle_{\text{QZ-5Z}} = 3.90 \text{ cm}^{-1}$, $\langle |\Delta| \rangle_{\text{5Z-6Z}} = 2.62 \text{ cm}^{-1}$], as seen in Figure 1. The similarity in convergence of $\langle |\Delta| \rangle$ and $\langle \text{error} \rangle$, respectively, between the relativistically corrected and noncorrected computations can be seen in Figures 1 and 3 where the GC and DK data parallel the noncorrected data. As a result, the difference, $\langle \text{diff} \rangle$, between the corrected and the noncorrected vibrational frequencies is consistently in the range of 1.77 cm^{-1} to 2.56 cm^{-1} , and it is concluded that, on average, there is a 2.2 cm^{-1} error associated with neglecting relativistic effects.

F. Comparison of Different Computations. The magnitudes of the effect each factor has on the accuracy of two computations are compared in Table 7. The computations were done at the CCSD(T) level of theory with the cc-pVTZ and aug-cc-pVQZ basis sets. In both computations, core electrons were frozen, and relativistic effects were neglected. As previously discussed, the errors associated with using the CCSD(T) theory, neglecting core electron correlation, and neglecting relativistic effects are 4.7 cm^{-1} , 5.0 cm^{-1} , and 2.2 cm^{-1} , respectively. The error due to the convergence of the Hartree–Fock wavefunction for each computation was determined by the sum of the $\langle |\Delta| \rangle$ values found in Table 1 of each subsequent computation (for cc-pVTZ: $\langle |\Delta| \rangle_{\text{TZ-QZ}} + \langle |\Delta| \rangle_{\text{QZ-5Z}} + \langle |\Delta| \rangle_{\text{5Z-6Z}} = 2.68 \text{ cm}^{-1} + 1.40 \text{ cm}^{-1} + 0.19 \text{ cm}^{-1} < 4.3 \text{ cm}^{-1}$). Similarly, the error due to the convergence of valence electron description was determined by the sum of $\langle |\Delta| \rangle$ values (Table 4) of each subsequent computation (for cc-pVTZ: $\langle |\Delta| \rangle_{\text{TZ-QZ}} + \langle |\Delta| \rangle_{\text{QZ-5Z}} + \langle |\Delta| \rangle_{\text{5Z-6Z}} = 6.33 \text{ cm}^{-1} + 4.14 \text{ cm}^{-1} + 2.31 \text{ cm}^{-1} < 12.8 \text{ cm}^{-1}$). The total expected error from experiment of the vibrational frequencies for each computation is represented by the square root of the sum of the squares (RSS) of the errors associated with the contributing factors.

While the cc-pVTZ computation is available for a wide variety of atoms and yields results 175 times faster than the aug-cc-pVQZ computation (minutes vs hours), the expected error associated with the cc-pVTZ calculation is nearly

double (cc-pVTZ RSS = 15.3 cm^{-1} , aug-cc-pVQZ RSS = 8.3 cm^{-1}). As seen in Table 7, the actual error from experimental frequencies is 6 times as great for the cc-pVTZ computation (cc-pVTZ $\langle \text{error} \rangle = 10.5 \text{ cm}^{-1}$, aug-cc-pVQZ $\langle \text{error} \rangle = 1.7 \text{ cm}^{-1}$). The larger magnitude of the root sum of squares (RSS) of the expected errors in both calculations implies the presence of a cancellation of errors. While cancellation of error is expressed in both computations, the proportionally greater cancellation of error of the aug-cc-pVQZ computation in conjunction with a smaller expected error results in substantially more accurate harmonic frequencies.

Conclusions

It is important to understand and determine the accuracy to which harmonic force constants can be computed. This work presents a comprehensive study of the major contributing factors affecting the accuracy of such computations through the determination of the underlying sources of error and the evaluation and quantification of the error at varying levels of approximation.

When using basis sets larger than cc-pVQZ, the error associated with the Hartree–Fock wavefunction has converged to 1.6 cm^{-1} . Consequently, the associated error can be neglected in corresponding computations. Due to the small difference in values between the CCSD(T) and CCSDT methods ($\langle |\Delta| \rangle = 0.3 \text{ cm}^{-1}$), the CCSD(T) level of theory was chosen for computational efficiency and because of the widespread support of the CCSD(T) method in modern quantum chemistry programs. However, this use of the CCSD(T) theory over the CCSDT(Q), CCSDTQ, or CCSDTQP theories to approximate electron correlation results in a 5 cm^{-1} error. It was also found that just as CCSD(T) yields comparable results to CCSDT but at a lower computation cost, CCSDT(Q) is more computationally affordable than CCSDTQ but gives comparable results. Therefore, the possibility of efficiently increasing the accuracy of computations through increased correlation is possible via perturbatively applied quadruples.

The quality of a basis set is one of the most important factors affecting the error associated with computations of vibrational frequencies. Smaller basis sets decrease computational time and demand of computer hardware at the expense of a significant increase in error due to valence electron description. With the use of either the cc-pV5Z or the aug-cc-pVQZ basis set, the error associated with valence electron description has converged within 3.7 cm^{-1} , although the augmented basis set yields more accurate results.

The error associated with neglecting core electron correlation effects is determined to be 5 cm^{-1} when using basis sets of at least quadruple- ζ quality. While correlating the core electrons decreases the expected uncertainty in the computation, computational time increases, and the accuracy of the computed frequencies decreases due to the decrease in fortuitous cancellation of errors. Consequently, core electron correlation should be included in only the most rigorous computations. Similarly, while neglecting relativistic effects introduces a 2 cm^{-1} error, the decrease in expected error does not outweigh the increase in computational time.

While computations of CCSD(T)/cc-pVTZ quality can be done significantly faster than those of performed at CCSD(T)/aug-cc-pVQZ quality, the substantially smaller expected uncertainty results in vibrational frequencies with a 2 cm^{-1} accuracy for CCSD(T)/aug-cc-pVQZ computations.

Acknowledgment. The authors thank Michael L. Poulblon for constructing and maintaining the computer cluster used for these computations. M.H.C., N.R.B., W.F.P., and the computer cluster were supported by a Cottrell College Science Award of Research Corporation and by the Scholar/Fellow Program of the Camille and Henry Dreyfus Foundation. P.R.T. was supported by the Wolfson Foundation through the Royal Society. Y.J.B. and J.F.S. were supported by the National Science Foundation and the Robert A. Welch Foundation.

References

- (1) Naus, H.; Ubachs, W.; Levelt, P. F.; Polyansky, O. L.; Zobov, N. F.; Tennyson, J. *J. Mol. Spectrosc.* **2001**, *205*, 117.
- (2) Carleer, M.; Jenouvrier, A.; Vandaele, A. C.; Bernath, P. F.; Mérienne, M. F.; Colin, R.; Zobov, N. F.; Polyansky, O. L.; Tennyson, J.; Savin, V. A. *J. Chem. Phys.* **1999**, *111*, 2444.
- (3) Wayne, R. P. *Chemistry of Atmospheres*, 3rd ed.; Oxford University Press: Oxford, 2000; pp 50–58.
- (4) Callegari, A.; Theulé, P.; Muentner, J. S.; Tolchenov, R. N.; Zobov, N. F.; Polyansky, O. L.; Tennyson, J.; Rizzo, T. R. *Science* **2002**, *297*, 993.
- (5) Ramanathan, V.; Vogelmann, A. M. *Ambio* **1997**, *26*, 38.
- (6) Wallace, L.; Bernath, P.; Livingston, W.; Hinkle, K.; Busler, J.; Guo, B.; Zhang, K. *Science* **1995**, *268*, 1155.
- (7) Polyansky, O. L.; Zobov, N. F.; Viti, S.; Tennyson, J.; Bernath, P. F.; Wallace, L. *Science* **1997**, *277*, 346.
- (8) Griffith, C. A.; Yelle, R. V.; Marley, M. S. *Science* **1998**, *282*, 2063.
- (9) Oppenheimer, B. R.; Kulkarni, S. R.; Matthews, K.; Nakajima T. *Science* **1995**, *270*, 1478.
- (10) Polyansky, O. L.; Császár, A. G.; Shirin, S. V.; Zobov, N. F.; Barletta, P.; Tennyson, J.; Schwenke, D. W.; Knowles, P. J. *Science* **2003**, *299*, 539.
- (11) Kalmar, B.; O'Brien, J. J. *J. Mol. Spectrosc.* **1998**, *192*, 386.
- (12) Xie, J.; Paldus, B. A.; Wahl, E. H.; Martin, J.; Owano, T. G.; Kruger, C. H.; Harris, J. S.; Zare, R. N. *Chem. Phys. Lett.* **1998**, *284*, 387.
- (13) Jensen, P. *J. Mol. Spectrosc.* **1989**, *133*, 438.
- (14) Polyansky, O. L.; Jensen, P.; Tennyson, J. *J. Chem. Phys.* **1996**, *105*, 6490.
- (15) Beardsworth, R.; Bunker, P. R.; Jensen P.; Kraemer, W. P. *J. Mol. Spectrosc.* **1986**, *118*, 50.
- (16) Császár, A. G.; Mills, I. M. *Spectrochimica* **1997**, *53*, 1101.
- (17) Groenenboom, G. C.; Wormer, P. E. S.; van der Avoird, A.; Mas, E. M.; Bukowski, R.; Szalewicz, K. *J. Chem. Phys.* **2000**, *113*, 6702.
- (18) Mas, E. M.; Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; Wormer, P. E. S.; van der Avoird, A. *J. Chem. Phys.* **2000**, *113*, 6687.
- (19) Carney, G. D.; Curtiss, L. A.; Langhoff, S. R. *J. Mol. Spectrosc.* **1976**, *61*, 371.
- (20) Császár, A. G.; Allen, W. D. *J. Chem. Phys.* **1996**, *104*, 2746.
- (21) Partridge H.; Schwenke, D. W. *J. Chem. Phys.* **1997**, *106*, 4618.
- (22) Werner, H.-J.; Knowles, P. J. *MOLPRO 2002.6*; University of Birmingham: Birmingham, U.K., 2002.
- (23) Stanton, J. F.; Gauss, J. *ACES II*; University of Texas: Austin, TX, 2005.
- (24) Kállay, M.; Surján, P. R. *J. Chem. Phys.* **2001**, *115*, 2945.
- (25) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910.
- (26) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
- (27) Bartlett, R. J.; Watts, J. D.; Kucharski, S. A.; Noga, J. *Chem. Phys. Lett.* **1990**, *165*, 513; erratum: **1990**, *167*, 609.
- (28) Noga, J.; Bartlett, R. J. *J. Chem. Phys.* **1987**, *86*, 7041; erratum: **1988**, *89*, 3041.
- (29) Scuseria, G. E.; Schaefer, H. F. *Chem. Phys. Lett.* **1988**, *152*, 382.
- (30) Bomble, Y. J.; Stanton, J. F.; Kállay, M.; Gauss, J. *J. Chem. Phys.* **2005**, *123*, 54101.
- (31) Kucharski, S. A.; Bartlett, R. J. *J. Chem. Phys.* **1992**, *97*, 4282.
- (32) Oliphant, N.; Adamowicz, L. *J. Chem. Phys.* **1991**, *95*, 6645.
- (33) Cowan, R. D.; Griffin, D. C. *J. Opt. Soc. Am.* **1976**, *66*, 1010.
- (34) Douglas, M.; Kroll, N. M. *Ann. Phys. (N.Y.)* **1974**, *82*, 89.
- (35) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (36) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.
- (37) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358.
- (38) Pliva, J.; Spirko, V.; Papousek, D. *J. Mol. Spectrosc.* **1967**, *23*, 331.
- (39) McCoy, A. B.; Sibert, E. L., III *J. Chem. Phys.* **1990**, *92*, 1893.
- (40) Ruden, T. A.; Helgaker, T.; Jørgensen, P.; Olsen, J. *J. Chem. Phys.* **2004**, *121*, 5874.
- (41) Feller, D.; Sordo, J. A. *J. Chem. Phys.* **2000**, *112*, 5604.
- (42) Martin, J. M. L.; Taylor, P. R. *Chem. Phys. Lett.* **1994**, *225*, 473.
- (43) Martin, J. M. L. *Chem. Phys. Lett.* **1995**, *242*, 343.
- (44) Pawłowski, F.; Halkier, A.; Jørgensen, P.; Bak, K. L.; Helgaker, T.; Klopper, W. *J. Chem. Phys.* **2003**, *118*, 2539.

CT600347E

Comparison of Semiempirical ZILSH and DFT Calculations of Exchange Constants in the Single Molecule Magnet $[\text{Fe}_8\text{O}_2(\text{OH})_{12}(\text{tacn})_6]^{8+}$

Ted A. O'Brien* and Brian J. O'Callaghan

Department of Chemistry and Chemical Biology, Indiana University—Purdue University Indianapolis, 402 North Blackford Street, Indianapolis, Indiana 46202

Received March 11, 2007

Abstract: The exchange constants describing magnetic interactions between high spin Fe^{3+} ions in the complex $[\text{Fe}_8\text{O}_2(\text{OH})_{12}(\text{tacn})_6]^{8+}$ have been estimated with the semiempirical ZILSH method, and the results compared to those of DFT calculations and experimental magnetic studies. The ZILSH method provides more accurate estimates of the exchange constants than the DFT calculations, particularly for the “body–body” interaction within the central Fe_4 “butterfly” unit of the complex. This interaction is found to be antiferromagnetic, which contrasts with the DFT description but is in agreement with experimental studies on smaller Fe_4 butterfly complexes and known empirical correlations between exchange constants and structural parameters. The ground-state wavefunction obtained by diagonalizing the Heisenberg spin Hamiltonian matrix has a spin of ten, in agreement with previous experimental and theoretical studies. Spin alignments in the ground state demonstrate how spin frustration can lead to nonzero spin in complexes with exclusively antiferromagnetic exchange interactions.

1. Introduction

Polynuclear complexes containing transition-metal ions with unpaired spins are a subject of great interest in both nanoscale electronics and biological reaction chemistry. Coupling of the local spins of the magnetic centers can lead to high total spin ground states and the possibility of single molecule magnetism,^{1–7} which could be used in nanoscale digital memory storage applications.⁸ Changes in the total spin of metal clusters comprising enzyme active sites have also been implicated in important biological reactions, such as the conversion of water to free dioxygen by the water oxidizing center of the photosynthetic reaction center.^{9–12} There has thus been strong motivation for the study of magnetic polynuclear transition-metal complexes, including synthesis and characterization of a growing number of single molecule magnets (SMMs) and smaller analogs of enzyme active sites.

Focusing on the SMMs, these complexes display slow reversal of magnetization at low temperature due to negative zero-field splitting of the components of a high spin ground

state.⁶ The size of the energy barrier for spin reversal is thus related to both the zero-field splitting parameter of the complex and the total spin of the ground state. The spin states are typically described in terms of the Heisenberg spin Hamiltonian (HSH)

$$\hat{H} = - \sum_{A < B} J_{AB} \hat{S}_A \cdot \hat{S}_B \quad (1)$$

in which the local spin moments of the transition-metal ions (described by the local spin operators \hat{S}_A^2 and \hat{S}_B^2) couple to form states of composite total spin. The parameters $\{J_{AB}\}$ in eq 1, the exchange constants, describe the magnitude and preferred direction of magnetic coupling between paramagnetic centers labeled “A” and “B”.

From an experimental perspective, estimates of the exchange constants of a complex are obtained by fitting the magnetic susceptibility measured for the complex over a range of temperatures, assuming Boltzmann-weighted populations of the spin states specified by the HSH. While certainly useful in understanding the magnetic interactions in a complex that dictate the ground-state spin and other

* Corresponding author e-mail: teobrien@indiana.edu.

properties, this approach suffers from the problem that the great number of fitting parameters for larger complexes leads to nonunique sets of fitting parameters. There is thus need for independent methods for estimating exchange constants for magnetic polynuclear transition-metal complexes. Quantum chemistry immediately suggests itself for this purpose, since it can in principle provide information on the energetics of spin interactions from first principles.

Despite the great theoretical and computational difficulties presented by complexes containing multiple open-shell transition-metal ions, theoretical methods have increasingly played a role in the study of magnetic interactions within such complexes. It might safely be said, however, that these methods are not yet of great utility in a practical sense. Many applications of density functional theory (DFT) methods to particularly the smaller SMMs have started to appear, but these calculations are still limited in the size of complex that can be treated. A realistic practical limit is on the order of ten transition-metal ions without resorting to dividing larger complexes into smaller model fragments. Given that many larger complexes have been reported (e.g., complexes with 22 iron ions¹³ and 84 manganese ions¹⁴), there is still a need for more efficient semiempirical methods that can accurately estimate the size and direction of magnetic couplings in large complexes.

O'Brien and Davidson recently introduced the semiempirical ZILSH formalism for treating magnetic interactions in polynuclear transition-metal complexes.¹⁵ The formalism combines the INDO/S method of Zerner^{16–23} popularized in the ZINDO program package²⁴ with Davidson's local spin operator^{25,26} to obtain estimates of the exchange constant J_{AB} appearing in the Heisenberg spin model. ZILSH has been successfully applied to 20 iron and manganese complexes with nuclearities ranging from 2 to 12^{15,27–31} but has not been systematically tested by direct comparison to experimental results for a large number of complexes. Furthermore, few comparisons with ostensibly more accurate DFT calculations are available. Larger complexes are of particular interest in this regard, as the semiempirical ZILSH method has the potential to treat complexes of much greater size than more expensive DFT calculations.

Ruiz et al. recently reported results of DFT calculations on the SMM known as Fe_8 , $[\text{Fe}_8\text{O}_2(\text{OH})_{12}(\text{tacn})_6]^{8+}$.³² This complex was one of the first SMMs to be characterized.³³ It has been studied in great detail both experimentally^{33–36} and theoretically,^{32,37} making it a good candidate for a comparative study of quantum methods. Here we report results of ZILSH calculations of the exchange constants, state energies, and spin distribution in the ground state of Fe_8 . Comparisons are made with experiment and with results of the recent DFT calculations of Ruiz et al.³²

2. Summary of Experimental Studies of Fe_8

The structure of Fe_8 ³³ is shown in Figure 1 (panel a) along with a schematic diagram of the exchange pathways with significant magnetic interactions (panel b). The complex is asymmetric but exhibits virtual D_2 symmetry.³⁴ The exchange constants can thus be approximately grouped into J_{bb} , J_A , J_B , and J_C as shown in Figure 1. All previous treatments

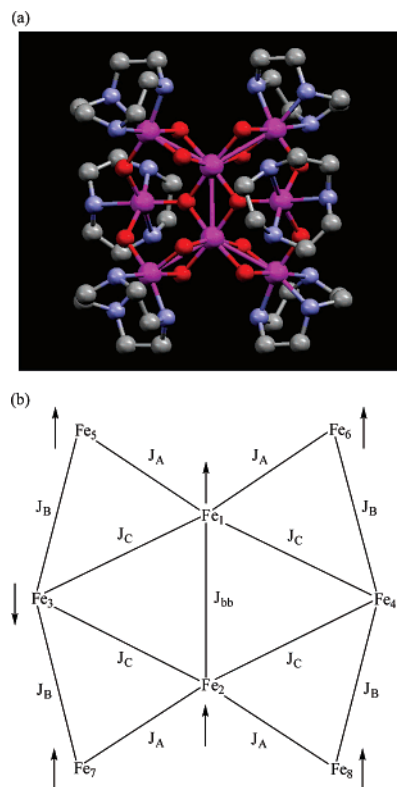


Figure 1. Structure of $[\text{Fe}_8\text{O}_2(\text{OH})_{12}(\text{tacn})_6]^{8+}$: (a) structural diagram (hydrogen atoms omitted for clarity). The structure was obtained from ref 33. (b) Schematic representation with labeling scheme for iron ions and exchange constants.

have assumed this approximation. The bridging ligands mediating these various pathways are listed in Table 2. The central tetranuclear unit consisting of ions Fe_1 – Fe_4 closely resembles the core of the well-known Fe_4 butterfly complexes.^{38–42} The “body–body” interaction J_{bb} (analogous to $J_{12} = J_{bb}$ in Figure 1) in these complexes is on the order of -20 cm^{-1} , while the “wingtip–body” interactions (analogous to J_C in Figure 1) are on the order of -100 cm^{-1} .

The magnetic properties of Fe_8 have been extensively studied with both dc and ac variable temperature magnetic susceptibility ($\chi_M T$ vs T) measurements,³⁴ magnetization vs magnetic field measurements at different temperatures,³⁴ and high-frequency electron paramagnetic resonance (EPR) measurements on both powder^{34,35} and single crystal³⁵ samples. In early work, Delfs et al.³⁴ suggested a ground-state spin of $S = 10$ based on the magnetic susceptibility and magnetization curves. They interpreted the temperature dependence of the $\chi_M T$ curve at low temperature and splitting of the EPR signal at 4.2 K to indicate the presence of an excited state with a spin of $S = 9$ at very low energy. Calculations of the magnetic susceptibility were performed by diagonalizing the HSH matrix for two choices of exchange constants J_{bb} , J_A , J_B , and J_C (Figure 1). Values of $J_{bb} = -102 \text{ cm}^{-1}$, $J_A = -15 \text{ cm}^{-1}$, $J_B = -35 \text{ cm}^{-1}$, and $J_C = -120 \text{ cm}^{-1}$ were found to reproduce the experimental data while predicting the presence of the excited state with a spin of $S = 9$ within 0.5 cm^{-1} of the ground state.

Barra et al. later reconsidered the interpretation described above on the basis of single-crystal high-frequency EPR

measurements.³⁵ They demonstrated that the low-temperature magnetic behavior of the complex is due to unusually large magnetic anisotropy in the ground state rather than the presence of a low-lying excited state. Similar behavior has also been reported for a structural analog of Fe₈.³⁶ Given this conclusion, exchange parameters that also reproduce the experimental χ_{MT} vs T curve but do not place the first excited state within a few cm^{-1} of the ground state were suggested. These values, $J_{\text{bb}} = -25 \text{ cm}^{-1}$, $J_{\text{A}} = -18 \text{ cm}^{-1}$, $J_{\text{B}} = -41 \text{ cm}^{-1}$, and $J_{\text{C}} = -140 \text{ cm}^{-1}$, represent the best reflection of the experimental data and are taken as “experimental” values in the following discussion. With these exchange constants, the first excited state with a spin of $S = 9$ has an energy of 24.5 cm^{-1} .

The ground-state wavefunction obtained by diagonalization of the HSH with the experimental values of the exchange constants gives local z -components of spin of close to $+5/2$ for ions Fe₁, Fe₂, and Fe₅–Fe₈ (see Figure 1 for labeling scheme) and $-5/2$ for Fe₃ and Fe₄.³⁷ These spin alignments are depicted in panel b of Figure 1. This arrangement of local spin components indicates that the J_{bb} and J_{A} pathways of Fe₈ are spin frustrated, as might be expected given the much larger antiferromagnetic couplings in the J_{B} and J_{C} pathways. Both the DFT calculations of Ruiz et al.³² and the ZILSH calculations reported here agree with this picture of the spin alignments in Fe₈.

3. Theory and Methods

3.1. ZILSH Calculations. The ZILSH calculations on Fe₈ used the procedure described in ref 15. In summary, the procedure uses unrestricted Hartree Fock (UHF) molecular orbital (MO) calculations with the intermediate neglect of differential overlap Hamiltonian parametrized for optical spectroscopy (INDO/S) of Zerner.^{16–23} These calculations provide single determinant wavefunctions for various spin components defined by particular alignments of the spins of the metal ions in the complex. A semiempirical application¹⁵ of Davidson’s local spin operator^{25,26} is used to obtain spin couplings $\langle \hat{S}_{\text{A}} \cdot \hat{S}_{\text{B}} \rangle^{\text{UHF}}$ from the unrestricted wavefunctions. The exchange constants are then obtained assuming an effective Hamiltonian operator of the Heisenberg spin form

$$\hat{H}_{\text{eff}} = \hat{H}_0 - \sum_{\text{A} < \text{B}} J_{\text{AB}} \langle \hat{S}_{\text{A}} \cdot \hat{S}_{\text{B}} \rangle^{\text{UHF}} \quad (2)$$

where \hat{H}_0 contains all spin-independent terms such as electron–nuclear attraction. The expectation value of \hat{H}_{eff} for the i th spin component is then

$$E^{\text{UHF},i} = E_0 - \sum_{\text{A} < \text{B}} J_{\text{AB}} \langle \hat{S}_{\text{A}} \cdot \hat{S}_{\text{B}} \rangle^{\text{UHF},i} \quad (3)$$

where E_0 contains all spin-independent contributions to the energy. Given energies and spin couplings for the appropriate number of spin components ($1/2 N_m(N_m - 1) + 1$, where N_m is the number of metal ions in the complex), the parameters E_0 and $\{J_{\text{AB}}\}$ are solved for simultaneously. Performing calculations on spin components with all unpaired spins parallel (“high spin”) and with unpaired spins reversed on all unique combinations of two metal ions provides the

correct number of equations for simultaneous solution for all parameters. This procedure is similar to those developed by Noodleman^{43–46} and Yamaguchi,^{47–49} except it calculates expectation values $\langle \hat{S}_{\text{A}} \cdot \hat{S}_{\text{B}} \rangle^{\text{UHF}}$ from the wavefunctions rather than assuming formal values.

The DFT calculations of Ruiz et al.³² used either the hybrid B3LYP functional^{50,51} or the PBE functional,⁵² with the TZVP basis set of Ahlrichs⁵³ for the iron ions and the DZVP basis of Ahlrichs⁵⁴ for lighter atoms. A different procedure than that described above was used to obtain estimates of the exchange constants.³² Calculations were performed for five spin components, and differences in the component energies were used to solve simultaneously for J_{bb} , J_{A} , J_{B} , and J_{C} (see Figure 1, panel b). This displays an important difference between the methods—the ZILSH calculations consider 29 spin components rather than five and solve for exchange constants for all unique pairwise magnetic interactions in the complex. No assumptions are made based on symmetry (e.g., setting J_{13} , J_{14} , J_{23} , and J_{24} equal to J_{C} ; see Figure 1, panel b), and none of the interactions are arbitrarily assumed to be zero regardless of whether the two metals involved are directly bridged by ligands or not. The latter allows evaluation of second-neighbor interactions, for example, which are generally assumed to be zero. Nonzero values for second-neighbor couplings have been suggested on the basis of both experimental and theoretical evidence; see, e.g., refs 55 and 41, respectively.

3.2. Spin Eigenstates – Diagonalization of the Heisenberg Spin Hamiltonian. The spin eigenstates of the complex are obtained for a given set of exchange constants by substituting them into the HSH (eq 1) and diagonalizing the operator in the basis of spin components $\phi_i = |M_1 M_2 \dots M_N\rangle_i$, where M_{A} is the local z -component of spin of the metal center labeled “A”. The resulting spin state wavefunctions $|\psi_S\rangle^I$ are linear combinations of these components

$$|\psi_S\rangle^I = \sum_i C_i \phi_i = \sum_i |M_1 M_2 \dots M_N\rangle_i \quad (4)$$

where the expansion runs over components for which the local z -components of spin add to the total spin S of the state. For smaller complexes the entire Hamiltonian matrix can readily be diagonalized to obtain energies and wavefunctions for all spin states, but this procedure becomes increasingly expensive for larger complexes. In this work we use an implementation of the Davidson algorithm⁵⁶ that efficiently provides the energy and wavefunction for the lowest energy state of each spin.

Several useful quantities can be calculated from the wavefunction of the ground (or any other) state, including the local z -component of spin of each metal, $\langle \hat{S}_{z\text{A}} \rangle$, and spin couplings $\langle \hat{S}_{\text{A}} \cdot \hat{S}_{\text{B}} \rangle$ between metal spins

$$\langle \hat{S}_{z\text{A}} \rangle = \langle \psi_S | \hat{S}_{z\text{A}} | \psi_S \rangle = \sum_i C_i^2 \langle M_{\text{A}} \rangle_i \quad (5)$$

$$\langle \hat{S}_{\text{A}} \cdot \hat{S}_{\text{B}} \rangle = \langle \psi_S | \hat{S}_{\text{A}} \cdot \hat{S}_{\text{B}} | \psi_S \rangle = \sum_{ij} C_i C_j \langle \phi_i | 1/2 \hat{S}_{\text{A}}^+ \cdot \hat{S}_{\text{B}}^- + 1/2 \hat{S}_{\text{A}}^- \cdot \hat{S}_{\text{B}}^+ + \hat{S}_{z\text{A}} \cdot \hat{S}_{z\text{B}} | \phi_j \rangle \quad (6)$$

where \hat{S}_{A}^+ and \hat{S}_{A}^- are the standard raising and lowering

Table 1. Energies and Local Spin Densities Computed from ZILSH UHF Wavefunctions for 24 Various Spin Components of the Fe₈ Complex [Fe₈O₂(OH)₁₂(tacn)₆]⁸⁺

spin reversal ^a	energy ^b (cm ⁻¹)	Fe ₁	Fe ₂	Fe ₃	Fe ₄	Fe ₅	Fe ₆	Fe ₇	Fe ₈
high spin	4551.64	4.39	4.39	4.26	4.27	4.20	4.19	4.19	4.19
1,2	552.46	-4.33	-4.34	4.22	4.23	4.19	4.20	4.19	4.20
1,3	1812.50	-4.36	4.37	-4.22	4.25	4.18	4.20	4.18	4.19
1,4	1820.26	-4.36	4.37	4.24	-4.23	4.19	4.19	4.19	4.18
1,5	2265.69	-4.34	4.40	4.23	4.25	-4.18	4.20	4.19	4.19
1,6	2275.51	-4.34	4.40	4.24	4.24	4.19	-4.19	4.19	4.19
1,7	1934.21	-4.34	4.39	4.23	4.25	4.19	4.20	-4.18	4.19
1,8	1981.63	-4.34	4.39	4.24	4.24	4.19	4.20	4.19	-4.19
2,3	1772.47	4.37	-4.36	-4.22	4.25	4.18	4.20	4.18	4.20
2,4	1897.81	4.37	-4.36	4.24	-4.23	4.19	4.19	4.19	4.19
2,5	1995.94	4.39	-4.34	4.23	4.25	-4.18	4.20	4.19	4.20
2,6	2037.88	4.39	-4.34	4.24	4.24	4.19	-4.19	4.19	4.20
2,7	2351.62	4.39	-4.35	4.23	4.25	4.19	4.20	-4.18	4.20
2,8	2380.07	4.39	-4.35	4.24	4.24	4.19	4.20	4.19	-4.18
3,4	0.00	4.34	4.35	-4.20	-4.21	4.18	4.19	4.18	4.18
3,5	2346.02	4.36	4.37	-4.21	4.27	-4.19	4.20	4.18	4.19
3,6	1745.03	4.36	4.37	-4.20	4.26	4.18	-4.19	4.18	4.19
3,7	2345.46	4.37	4.37	-4.21	4.27	4.18	4.20	-4.19	4.19
3,8	1775.53	4.37	4.37	-4.20	4.26	4.18	4.20	4.18	-4.19
4,5	1809.51	4.36	4.37	4.25	-4.21	-4.18	4.19	4.19	4.18
4,6	2443.86	4.36	4.37	4.26	-4.22	4.19	-4.20	4.19	4.18
4,7	1834.72	4.37	4.37	4.25	-4.21	4.19	4.19	-4.18	4.18
4,8	2426.64	4.37	4.37	4.26	-4.22	4.19	4.19	4.19	-4.20
5,6	3587.36	4.38	4.39	4.25	4.26	-4.18	-4.19	4.19	4.19
5,7	3567.76	4.38	4.39	4.24	4.27	-4.18	4.20	-4.18	4.19
5,8	3614.46	4.38	4.39	4.25	4.26	-4.18	4.20	4.19	-4.19
6,7	3607.94	4.38	4.39	4.25	4.26	4.19	-4.19	-4.18	4.19
6,8	3659.42	4.38	4.39	4.26	4.25	4.19	-4.19	4.19	-4.19
7,8	3642.71	4.39	4.38	4.25	4.26	4.19	4.20	-4.18	-4.19

^a All spins on the indicated metals reversed relative to others; see Figure 1 panel b for the numbering scheme. ^b Relative to energy of component with all unpaired spins on Fe₃ and Fe₄ reversed.

operators for the *z*-component of spin of center A, etc. The quantity $\langle \hat{S}_{zA} \rangle$ describes the spin alignments in the state being considered. The ground-state spin couplings $\langle \hat{S}_A \cdot \hat{S}_B \rangle$ are particularly useful for identifying exchange pathways that are spin frustrated. The spin coupling indicates the actual alignment of the spin components M_A and M_B in the state, while the exchange constant J_{AB} indicates the preferred alignment. Under the $-J$ convention, then, any pathway with $\langle \hat{S}_A \cdot \hat{S}_B \rangle$ and J_{AB} of different signs is frustrated. The contribution made by an exchange pathway to the total energy of the ground state is simply $\Delta E = -J_{AB} \langle \hat{S}_A \cdot \hat{S}_B \rangle$, so a frustrated pathway increases the ground-state energy. This occurs because the resulting distribution of spins throughout the complex allows compensatory, larger decreases in energy in other pathways that are not frustrated.

4. Results and Discussion

Following the ZILSH procedure given in ref 15, an initial set of molecular orbitals (MOs) was obtained with the configuration averaged Hartree Fock (CAHF) procedure of Zerner.⁵⁷ This calculation was followed by a restricted open shell Hartree Fock (ROHF) calculation using the CAHF MOs as the starting guess. The open shell MOs obtained from the ROHF calculation, which consisted largely of iron 3d atomic orbitals, were localized with the procedure of Boys.^{58–60} The resulting MOs were then used as starting

guesses for unrestricted Hartree Fock (UHF) calculations on components with all unpaired spins aligned, and all cases where the unpaired spins of two metals were reversed relative to the others. These UHF calculations converged readily, each executing in minutes on the IBM JS20 processors of the Libra cluster at Indiana University.

The energies and local spin densities of the metal ions found for the 29 UHF components are presented in Table 1. The local spin densities are computed within the population analysis scheme of Löwdin⁶¹ and are equal to twice the number of unpaired electrons on each metal. Their signs indicate alignments of the local *z*-components of spin. All values obtained are close to the formal values of ± 5 expected for high spin Fe³⁺ ions and are very similar to values obtained with ZILSH for other complexes of Fe³⁺ (refs 15, 30, and 31). The absolute values of the spin densities range from 4.18 to 4.40, which are comparable to those obtained in the DFT calculations of Ruiz et al., 4.10–4.17 found with natural bond order analysis and 4.18–4.25 found with Mulliken analysis.³² The lowest energy component is that with the spins of Fe₃ and Fe₄ reversed relative to the others (see panel b of Figure 1), in agreement with the results of Ruiz et al.³² The component with all spins aligned has the highest energy by a considerable margin, indicating that the magnetic interactions in the complex are predominantly antiferromagnetic.

Table 2. Exchange Constants Obtained for the Fe₈ Complex from ZILSH and DFT Calculations and from Experimental Magnetic Susceptibility Data and Ground-State Spin Couplings Obtained from Diagonalization of the Heisenberg Spin Hamiltonian with ZILSH Exchange Constants

parameter ^a	type of interaction	J_{PBE}^b (cm ⁻¹)	J_{B3LYP}^b	J_{ZILSH}^c	$\langle \hat{S}_A \cdot \hat{S}_B \rangle^d$	J_{exp}^e
J_{bb}	J_{12} (O ²⁻) ₂	+28.9	+5.1	-9.4	+6.25	-25
J_{A}	J_{15} (OH ⁻) ₂	-9.2	-10.4	-19.3	+5.58	-18
	J_{16}			-17.5	+5.56	
	J_{27}			-17.9	+5.61	
	J_{28}			-16.8	+5.58	
	J_{35} OH ⁻	-14.4	-34.1	-36.4	-6.00	-41
J_{B}	J_{37}			-34.0	-5.99	
	J_{46}			-33.2	-5.97	
	J_{48}			-30.6	-5.89	
	J_{13} O ²⁻	-55.8	-66.5	-94.7	-7.24	-140
J_{C}	J_{14}			-89.0	-7.22	
	J_{23}			-87.8	-7.17	
	J_{24}			-88.4	-7.25	

^a See Figure 1, panel b for numbering scheme. ^b Reference 32. ^c This work. ^d Computed from the ground-state wavefunction obtained by diagonalizing the HSH with ZILSH exchange constants. ^e Reference 36.

Spin couplings $\langle \hat{S}_A \cdot \hat{S}_B \rangle^{\text{UHF}}$ computed for the 29 spin components listed in Table 1 are given as Supporting Information. All values are close to ± 4.5 , which are typical of values obtained with ZILSH for other complexes of Fe³⁺ (refs 15, 30, and 31). Along with the component energies of Table 1, these spin couplings allow all exchange constants in the complex to be obtained through simultaneous solution of eq 3. All nonzero exchange constants found are presented in Table 2, along with those found with DFT by Ruiz et al.³² and those fit to the experimental magnetic susceptibility data.³⁵

Considering first the ZILSH exchange constants for all pathways, some minor variations in values were found within the subgroups J_{A} , J_{B} , and J_{C} , reflecting the actual lack of symmetry in the complex. These variations are uniformly small, indicating that the assumption of equivalent interactions is a good approximation for this complex. Turning to a comparison of calculated values with experiment, it is apparent from Table 2 that the ZILSH exchange constants compare more favorably with the experimental values than the DFT values. Comparing results obtained with the two functionals, it appears that the hybrid B3LYP functional performs better than the PBE functional, particularly for the hydroxide-mediated interaction J_{B} and the oxide-mediated interaction J_{bb} .

The DFT calculations suggest that J_{bb} , the “body–body” interaction within the central Fe₄ butterfly cluster of the complex, might be weakly ferromagnetic. This is not supported by experimental results for known Fe₄ butterfly complexes, which all have small but antiferromagnetic couplings ranging between -11 cm⁻¹ and -21 cm⁻¹.^{38,39,62–64} It should be pointed out, however, that the quality of fits of magnetic susceptibility data for these complexes is relatively insensitive to the value of J_{bb} . In the case of the complex [Fe₄O₂(O₂CMe)₇(bpy)₂]⁺, for example, McCusker et al.³⁹

could only conclude that J_{bb} is more positive than -15 cm⁻¹ and likely to be antiferromagnetic. The experimental results can thus not be assumed to be definitive for these complexes regarding the sign of J_{bb} .

Further insight regarding the value and sign of the exchange constant J_{bb} in Fe₈ can be gained by looking at existing relationships between exchange constants and structural parameters such as Fe–O distances (r) and Fe–O–Fe angles (ϕ) in bridging pathways. Several such “magnetostructural correlations” have been presented in the literature. Gorun and Lippard⁶⁵ considered 36 complexes with two or three Fe³⁺ ions and both substituted and unsubstituted oxide bridging ligands and found a correlation between Fe–O distance and J

$$-J = (1.7526 \times 10^{12} \text{ cm}^{-1}) \exp(-12.633 \text{ \AA}^{-1} \cdot P) \quad (7)$$

where P is “half the shortest distance of the superexchange pathway between two metals”. Weihe and Güdel⁶⁶ considered 32 oxide-bridged Fe³⁺ dimer complexes with exchange constants ranging from -160 cm⁻¹ to -265 cm⁻¹ and found a relationship between J and both Fe–O distance and Fe–O–Fe angle

$$-J = (1.337 \times 10^8 \text{ cm}^{-1}) (3.536 + 2.488 \cos(\phi) + \cos^2(\phi)) \exp(-7.909 \text{ \AA}^{-1} \cdot \bar{r}) \quad (8)$$

where \bar{r} is the average Fe–O distance for the exchange pathway. Cañada-Vilalta et al.³¹ considered a correlation between J and Fe–O distance and Fe–O–Fe angle in four hexanuclear Fe³⁺ complexes with substituted and unsubstituted oxide bridging interactions, obtaining

$$-J = (2 \times 10^7 \text{ cm}^{-1}) (0.2 - \cos(\phi) + \cos^2(\phi)) \exp(-7 \text{ \AA}^{-1} \cdot \bar{r}) \quad (9)$$

where \bar{r} and ϕ are the average Fe–O distance and Fe–O–Fe angle for the shortest bridging pathway between two metals.

Estimates of the exchange constant J_{bb} of Fe₈ obtained from eqs 7–9 are collected in Table 3 along with the various geometric parameters used. Values between -16 and -77 cm⁻¹ are obtained. Among these, the value of -77 cm⁻¹ obtained from eq 8 seems too large in magnitude and might be considered least reliable given that the correlation included only dimer complexes with $|J|$ greater than 160 cm⁻¹, and \bar{r} and ϕ values quite different from those found for the J_{bb} pathway in Fe₈.⁶⁶ The correlation of Gorun and Lippard⁶⁵ included interactions mediated by unsubstituted oxide ligands with exchange constants ranging from -52 cm⁻¹ to -264 cm⁻¹ as well as interactions mediated by substituted oxide ligands with J as low as -14 cm⁻¹ so is likely more applicable to J_{bb} in Fe₈. The correlation of Cañada-Vilalta et al.³¹ is perhaps the most reliable, as it is based on interactions in polynuclear complexes that are structurally similar to Fe₈, with exchange constants for Fe–O²⁻–Fe interactions as low as -8 cm⁻¹. The latter two correlations predict values of -31 cm⁻¹ and -16 cm⁻¹ for J_{12} in Fe₈, respectively. Given that J_{bb} found to closely reproduce the experimental $\chi_{\text{M}}T$ curve of Fe₈ is -25 cm⁻¹,³⁵ the ZILSH calculations estimate a value of -10 cm⁻¹ (Table 2), and

Table 3. Exchange Constant J_{bb} (See Figure 1, Panel b for the Labeling Scheme) of the Fe_8 Complex Estimated with Various Magnetostructural Correlations^f

formula	geometric parameter(s)	J_{12} (cm^{-1})	ref
$-J = (1.7526 \times 10^{12} \text{ cm}^{-1}) \exp(-12.633 \text{ \AA}^{-1} \cdot P)$	$P^a = 1.9565 \text{ \AA}$	-30.5	67
$-J = (1.337 \times 10^8 \text{ cm}^{-1}) (3.536 + 2.488 \cos(\phi) + \cos^2(\phi)) \exp(-7.909 \text{ \AA}^{-1} \cdot \bar{\tau})$	$\bar{\tau}^b = 1.96525 \text{ \AA}$ $\phi^c = 96.81^\circ$	-77.3	68
$-J = (2 \times 10^7 \text{ cm}^{-1}) (0.2 - \cos(\phi) + \cos^2(\phi)) \exp(-7 \text{ \AA}^{-1} \cdot \bar{\tau})$	$\bar{\tau}^d = 1.9565 \text{ \AA}$ $\phi^e = 97.38^\circ$	-15.6	31

^a "Half the shortest distance of the superexchange pathway between two metals." (ref 67). ^b Average Fe–O distance in the exchange pathway. ^c Average Fe–O–Fe angle in the exchange pathway. ^d Average Fe–O distance in the shortest bridging pathway. ^e Average Fe–O–Fe angle in the shortest bridging pathway. ^f Structural parameters were obtained from ref 33.

Table 4. Local z-Components of Spin $\langle \hat{S}_{zA} \rangle$ and Energy Difference between Ground and First Excited State Computed from the Ground-State Wavefunction Obtained by Diagonalizing the HSH with Exchange Constants Obtained from Various Methods

z-component	PBE ^a	B3LYP ^a	ZILSH ^b	(exp) ^c
$\langle \hat{S}_{z1} \rangle^d$	3.72	4.46	4.15	4.06
$\langle \hat{S}_{z2} \rangle$	3.72	4.46	4.18	4.06
$\langle \hat{S}_{z3} \rangle$	-3.65	-4.28	-4.04	-3.98
$\langle \hat{S}_{z4} \rangle$	-3.65	-4.28	-4.04	-3.98
$\langle \hat{S}_{z5} \rangle$	4.97	4.91	4.95	4.96
$\langle \hat{S}_{z6} \rangle$	4.97	4.91	4.94	4.96
$\langle \hat{S}_{z7} \rangle$	4.97	4.91	4.94	4.96
$\langle \hat{S}_{z8} \rangle$	4.97	4.91	4.94	4.96
$\Delta E, S = 10 \rightarrow S = 9$ (cm^{-1})	4.5	30.5	17.4	24.5

^a Reference 32. ^b This work. ^c Reference 36. ^d See Figure 1, panel b for the numbering scheme.

J_{bb} in Fe_4 butterfly complexes with Fe–O–Fe structural parameters similar to those for the J_{bb} pathway of Fe_8 ranges from -11 cm^{-1} to -21 cm^{-1} (see above), it seems unlikely that J_{bb} is ferromagnetic, as estimated by the DFT calculations reported in ref 32.

The above discussion suggests that the DFT calculations, particularly those with the PBE functional, are overestimating the ferromagnetic contribution to the exchange constant J_{bb} in Fe_8 . This might indicate a general tendency of DFT—a similar result was reported for the complex $[\text{Fe}_4\text{O}_2(\text{O}_2\text{CMe})_7(\text{bpy})_2]^+$, for which B3LYP calculations using the TZVP basis set⁵³ for the iron ions and the DZVP basis set⁵⁴ on lighter atoms (the same basis set used in ref 32 for Fe_8) gave $J_{bb} = +8.3 \text{ cm}^{-1}$,⁴¹ versus -18.8 cm^{-1} from a fit of the experimental χ_{MT} curve measured for the complex.³⁹ ZILSH calculations on this complex gave $J_{bb} = -12.0 \text{ cm}^{-1}$.¹⁵ The ZILSH method thus appears to slightly overestimate the ferromagnetic contribution to J_{bb} in Fe_8 and $[\text{Fe}_4\text{O}_2(\text{O}_2\text{CMe})_7(\text{bpy})_2]^+$ but not to the extent that the DFT methods do. More testing with additional complexes will be needed to confirm these conclusions about the performance of the methods.

Wavefunctions for the lowest energy state of each spin of the complex were obtained by substituting the exchange constants of Table 2 into the HSH (eq 1) and diagonalizing in the basis of spin components (eq 4). Results obtained from these calculations are presented in Table 4. The ground state has a spin of $S = 10$, in agreement with previous experimental^{34,35} and theoretical^{32,37} studies. The first excited state has a spin of $S = 9$, also as suggested previously,^{32,34,35,37} and is 17.4 cm^{-1} above the ground state in energy. This compares favorably with both experiment (24.5 cm^{-1} ; ref 35) and DFT calculations with the B3LYP functional (30.5

cm^{-1}). The PBE functional predicts a smaller energy difference (4.5 cm^{-1}).³²

The ground-state wavefunction obtained by diagonalizing the HSH with the ZILSH exchange constants consists primarily of the component in which the spins of Fe_3 and Fe_4 are reversed relative to the others. Its form is

$$|\psi_{S=10}\rangle = 0.70 \left| \frac{5}{2}, \frac{5}{2}, -\frac{5}{2}, -\frac{5}{2}, \frac{5}{2}, \frac{5}{2}, \frac{5}{2}, \frac{5}{2} \right\rangle - 0.24 \left| \frac{5}{2}, \frac{3}{2}, -\frac{5}{2}, -\frac{3}{2}, \frac{5}{2}, \frac{5}{2}, \frac{5}{2}, \frac{5}{2} \right\rangle - 0.24 \left| \frac{5}{2}, \frac{3}{2}, -\frac{3}{2}, -\frac{5}{2}, \frac{5}{2}, \frac{5}{2}, \frac{5}{2}, \frac{5}{2} \right\rangle - 0.24 \left| \frac{3}{2}, \frac{5}{2}, -\frac{5}{2}, -\frac{3}{2}, \frac{5}{2}, \frac{5}{2}, \frac{5}{2}, \frac{5}{2} \right\rangle + \dots \quad (10)$$

The rest of the wavefunction is comprised of a large number of components with much smaller coefficients. This wavefunction has the same leading component as that obtained by Barra et al.³⁵ with the experimental exchange constants but differs slightly in the relative magnitudes of the coefficients for those components making smaller contributions. The two wavefunctions predict very similar properties, however—both the local z-components of spin $\langle \hat{S}_{zA} \rangle$ and the energy difference between the ground state and the first excited state are very similar for the two wavefunctions (see the right two columns of Table 4). Wavefunctions obtained with the DFT exchange constants also give similar values for the local z-components of spin.

The wavefunctions obtained with the experimental and various theoretical exchange constants all display the same ground-state spin alignments, with the local spins of Fe_3 and Fe_4 reversed relative to the others. This leads to the total ground-state spin of $S = 10$. According to the ZILSH and the experimental exchange constants, the J_{bb} and four J_A pathways are spin frustrated, as seen from the spin couplings $\langle \hat{S}_A \cdot \hat{S}_B \rangle$ computed from the ground-state wavefunction (sixth column of Table 2): $\langle \hat{S}_A \cdot \hat{S}_B \rangle$ and J_{AB} differ in sign for these pathways, while they carry the same sign for the other pathways. This occurs because of the relatively small magnitudes of J_{bb} and J_A relative to those of J_B and J_C .

It is interesting to note that $\langle \hat{S}_1 \cdot \hat{S}_2 \rangle$, the spin coupling between metal ions in the J_{bb} pathway in the ground state, takes on a value of exactly $+6.25$ (Table 2). This is the formal value expected for two noninteracting particles with local spin quantum numbers of $S_A = S_B = 5/2$. This occurs because the very small J_{bb} interaction between Fe_1 and Fe_2 of -9.4 cm^{-1} is completely overwhelmed by the four much larger J_C interactions of ca. -100 cm^{-1} in the central butterfly unit. The interaction between Fe_1 and Fe_2 is thus rendered insignificant, and the two ions display the spin coupling

of two noninteracting particles with $S_A = S_B = 5/2$. The J_A pathways are also frustrated but with spin couplings that deviate from the value expected for noninteracting particles with $S_A = S_B = 5/2$. Viewed in this way, the J_{bb} pathway could be said to be completely frustrated in Fe_8 , while the J_A pathways are largely (but not completely) frustrated.

In general, the Fe_8 complex is a good example of how a large ground-state spin can occur in a complex in which all magnetic interactions are antiferromagnetic. The mechanism of this is spin frustration caused by competing antiferromagnetic exchange interactions of different magnitudes. In the case of Fe_8 , the topology of the complex leads to five adjacent pathways that are spin frustrated (the J_{bb} and four J_A pathways; see Figure 1), so that the six local spin components of Fe_1 , Fe_2 , and $\text{Fe}_5\text{--Fe}_8$ are aligned parallel in the ground state. The locations of the spin frustrated pathways are thus crucial to building up an appreciable spin moment in a complex with exclusively antiferromagnetic interactions—if the J_{bb} and J_C pathways were spin frustrated rather than the J_{bb} and J_A pathways, for example, the spins of $\text{Fe}_1\text{--Fe}_4$ would be aligned parallel to each other and antiparallel to the spins of $\text{Fe}_5\text{--Fe}_8$, and the spin of the ground state would be zero.

5. Conclusions

The exchange constants describing magnetic interactions between high spin Fe^{3+} ions in the single molecule magnet $[\text{Fe}_8\text{O}_2(\text{OH})_{12}(\text{tacn})_6]^{8+}$ have been estimated with the semi-empirical ZILSH method. The resulting values were compared to those obtained from DFT calculations³² as well as those fit to reproduce experimental magnetic data.³⁵ The ZILSH calculations were performed for 29 spin components of the complex, allowing exchange constants for all pairwise interactions to be solved for. This contrasts with the DFT calculations, which grouped together exchange constants that are approximately equivalent by symmetry, neglected others presumed to have zero magnitude, and considered only five spin components.³² Spin densities obtained for the metal ions for these spin components with ZILSH and DFT were very similar. The component with all unpaired spins aligned parallel was found to be considerably higher in energy than all others with both methods, indicating that the magnetic interactions are predominantly antiferromagnetic.

A number of conclusions can be drawn by comparing the exchange constants obtained from ZILSH and DFT calculations and from fitting to reproduce the experimental magnetic data. First, the exchange constants obtained with ZILSH show that the approximation of grouping exchange constants together into J_{bb} , J_A , J_B , and J_C (see Figure 1 and Table 1) is a good approximation for this complex. Second, the exchange constants obtained with ZILSH are consistently closer to those obtained from the experimental data than those obtained from the DFT calculations. Comparing the two functionals used in ref 32, the hybrid B3LYP functional performs better than the PBE functional for this complex.

Both DFT functionals indicate that J_{bb} , the “body–body” interaction within the central Fe_4 butterfly unit of Fe_8 , is

weakly ferromagnetic.³² The ZILSH calculations, by contrast, indicate that this interaction is weakly antiferromagnetic. This is supported by experimental magnetic data for known Fe_4 butterfly complexes^{38,39,62–64} and established correlations between exchange constants and structural parameters within the bridging pathways.^{31,65,66} There could be a general tendency of DFT methods to overestimate ferromagnetic contributions to exchange constants of small magnitude, as a similar result was reported for the butterfly complex $[\text{Fe}_4\text{O}_2(\text{O}_2\text{CMe})_7(\text{bpy})_2]^+$ (ref 41). Further testing of both ZILSH and DFT for other complexes is needed to further investigate this question.

Substitution of the exchange constants obtained with ZILSH into the Heisenberg spin Hamiltonian and diagonalization gives a ground state with a spin of $S = 10$ and a first excited state with spin of $S = 9$ that is 17.4 cm^{-1} above the ground state in energy. This is in close agreement with both experiment³⁵ and the B3LYP calculations.³² Spin alignments in the ground state are similar for all sets of exchange constants (ZILSH, DFT,³² and experimental^{35,37} values) and are arranged as shown in panel b of Figure 1. The J_{bb} and J_A pathways are spin frustrated, leading to parallel alignment of the spins of Fe_1 , Fe_2 , and $\text{Fe}_5\text{--Fe}_8$ and hence the ground-state spin of $S = 10$.

The ability of quantum chemical methods such as ZILSH and DFT to provide detailed analysis of magnetic interactions in large complexes could eventually be useful in the rational design of single molecule magnets with desirable properties such as high blocking temperatures for spin reversal. The ZILSH method is very efficient and could in principle be applied to much larger complexes than DFT methods. We are currently performing ZILSH calculations on complexes with nuclearities as high as 22 to demonstrate this capability.

Acknowledgment. This work was supported in part by Shared University Research grants from IBM, Inc. to Indiana University.

Supporting Information Available: Spin couplings $\langle \hat{S}_A \cdot \hat{S}_B \rangle$ computed from UHF wavefunctions for spin components with all unpaired spins aligned and all cases with unpaired spins on pairs of metal ions reversed (Table S1). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Sessoli, R.; Tsai, H. L.; Schake, A. R.; Wang, S. Y.; Vincent, J. B.; Folting, K.; Gatteschi, D.; Christou, G.; Hendrickson, D. N. *J. Am. Chem. Soc.* **1993**, *115*, 1804–1816.
- (2) Sessoli, R.; Gatteschi, D.; Caneschi, A.; Novak, M. A. *Nature* **1993**, *365*, 141–143.
- (3) Sangregorio, C.; Ohm, T.; Paulsen, C.; Sessoli, R.; Gatteschi, D. *Phys. Rev. Lett.* **1997**, *78*, 4645–4648.
- (4) Eppley, H. J.; Tsai, H. L.; Devries, N.; Folting, K.; Christou, G.; Hendrickson, D. N. *J. Am. Chem. Soc.* **1995**, *117*, 301–317.
- (5) Caneschi, A.; Gatteschi, D.; Sessoli, R.; Barra, A. L.; Brunel, L. C.; Guillot, M. *J. Am. Chem. Soc.* **1991**, *113*, 5873–5874.

- (6) Beltran, L. M. C.; Long, J. R. *Acc. Chem. Res.* **2005**, *38*, 325–334.
- (7) Barra, A. L.; Debrunner, P.; Gatteschi, D.; Schulz, C. E.; Sessoli, R. *Europhys. Lett.* **1996**, *35*, 133–138.
- (8) Leuenberger, M. N.; Loss, D. *Nature* **2001**, *410*, 789–793.
- (9) Yagi, M.; Kaneko, M. *Chem. Rev.* **2001**, *101*, 21–35.
- (10) Yachandra, V. K.; Sauer, K.; Klein, M. P. *Chem. Rev.* **1996**, *96*, 2927–2950.
- (11) Ruttinger, W.; Dismukes, C. G. *Chem. Rev.* **1997**, *97*, 1–24.
- (12) Manchanda, R.; Brudvig, G. W.; Crabtree, R. H. *Coord. Chem. Rev.* **1995**, *144*, 1–38.
- (13) Foguet-Albiol, D.; Abboud, K. A.; Christou, G. *Chem. Commun.* **2005**, 4282–4284.
- (14) Tasiopoulos, A. J.; Vinslava, A.; Wernsdorfer, W.; Abboud, K. A.; Christou, G. *Angew. Chem., Int. Ed. Engl.* **2004**, *43*, 2117–2121.
- (15) O'Brien, T. A.; Davidson, E. R. *Int. J. Quantum Chem.* **2003**, *92*, 294–325.
- (16) Zerner, M. C.; Loew, G. H.; Kirchner, R. F.; Muellerwesterhoff, U. T. *J. Am. Chem. Soc.* **1980**, *102*, 589–599.
- (17) Ridley, J. E.; Zerner, M. C. *Theor. Chim. Acta* **1973**, *32*, 111–134.
- (18) Kotzian, M.; Rosch, N.; Zerner, M. C. *Theor. Chim. Acta* **1992**, *81*, 201–222.
- (19) Culberson, J. C.; Knappe, P.; Rosch, N.; Zerner, M. C. *Theor. Chim. Acta* **1987**, *71*, 21–39.
- (20) Cory, M. G.; Kostlmeier, S.; Kotzian, M.; Rosch, N.; Zerner, M. C. *J. Chem. Phys.* **1994**, *100*, 1353–1365.
- (21) Bacon, A. D.; Zerner, M. C. *Theor. Chim. Acta* **1979**, *53*, 21–54.
- (22) Anderson, W. P.; Cundari, T. R.; Zerner, M. C. *Int. J. Quantum Chem.* **1991**, *39*, 31–45.
- (23) Anderson, W. P.; Cundari, T. R.; Drago, R. S.; Zerner, M. C. *Inorg. Chem.* **1990**, *29*, 1–3.
- (24) Zerner, M. C.; Ridley, J. E.; Bacon, A. D.; Edwards, W. D.; Head, J. D.; McKelvey, J.; Culberson, J. C.; Knappe, P.; Cory, M. G.; Weiner, B.; Baker, J. D.; Parkinson, W. A.; Kannis, D.; Yu, J.; Rosch, N.; Kotzian, M.; Tamm, T.; Karelson, M. M.; Zheng, X.; Pearl, G. M.; Broo, A.; Albert, K.; O'Brien, T. A.; Cullen, J. M.; Cramer, C. J.; Truhlar, D. G.; Li, J.; Hawkins, G. D.; Liotard, D. A. *ZINDO-A Semi-Empirical Program Package*; University of Florida: Gainesville, FL, 2000.
- (25) Davidson, E. R.; Clark, A. E. *Mol. Phys.* **2002**, *100*, 373–383.
- (26) Clark, A. E.; Davidson, E. R. *J. Chem. Phys.* **2001**, *115*, 7382–7392.
- (27) Tasiopoulos, A. J.; O'Brien, T. A.; Abboud, K. A.; Christou, G. *Angew. Chem., Int. Ed. Engl.* **2004**, *43*, 345–349.
- (28) Foguet-Albiol, D.; O'Brien, T. A.; Wernsdorfer, W.; Moulton, B.; Zaworotko, M. J.; Abboud, K. A.; Christou, G. *Angew. Chem., Int. Ed. Engl.* **2005**, *43*, 897–901.
- (29) Canada-Vilalta, C.; Streib, W. E.; Huffman, J. C.; O'Brien, T. A.; Davidson, E. R.; Christou, G. *Inorg. Chem.* **2004**, *43*, 101–115.
- (30) Canada-Vilalta, C.; O'Brien, T. A.; Pink, M.; Davidson, E. R.; Christou, G. *Inorg. Chem.* **2003**, *42*, 7819–7829.
- (31) Canada-Vilalta, C.; O'Brien, T. A.; Brechin, E. K.; Pink, M.; Davidson, E. R.; Christou, G. *Inorg. Chem.* **2004**, *43*, 5505–5521.
- (32) Ruiz, E.; Cano, J.; Alvarez, S. *Chem. Eur. J.* **2005**, *11*, 4767–4771.
- (33) Wieghardt, K.; Pohl, K.; Jibril, I.; Huttner, G. *Angew. Chem., Int. Ed. Engl.* **1984**, *23*, 77–78.
- (34) Delfs, C.; Gatteschi, D.; Pardi, L.; Sessoli, R.; Wieghardt, K.; Hanke, D. *Inorg. Chem.* **1993**, *32*, 3099–3103.
- (35) Barra, A. L.; Gatteschi, D.; Sessoli, R. *Chem. Eur. J.* **2000**, *6*, 1608–1614.
- (36) Barra, A. L.; Bencini, F.; Caneschi, A.; Gatteschi, D.; Paulsen, C.; Sangregorio, C.; Sessoli, R.; Sorace, L. *ChemPhysChem* **2001**, *2*, 523–531.
- (37) Raghu, C.; Rudra, I.; Sen, D.; Ramasesha, S. *Phys. Rev. B* **2001**, *64*, 064419.
- (38) Wemple, M. W.; Coggin, D. K.; Vincent, J. B.; McCusker, J. K.; Streib, W. E.; Huffman, J. C.; Hendrickson, D. N.; Christou, G. *J. Chem. Soc., Dalton Trans.* **1998**, 719–725.
- (39) McCusker, J. K.; Vincent, J. B.; Schmitt, E. A.; Mino, M. L.; Shin, K.; Coggin, D. K.; Hagen, P. M.; Huffman, J. C.; Christou, G.; Hendrickson, D. N. *J. Am. Chem. Soc.* **1991**, *113*, 3012–3021.
- (40) Gorun, S. M.; Lippard, S. J. *Inorg. Chem.* **1988**, *27*, 149–156.
- (41) Cauchy, T.; Ruiz, E.; Alvarez, S. *J. Am. Chem. Soc.* **2006**, *128*, 15722–15727.
- (42) Armstrong, W. H.; Roth, M. E.; Lippard, S. J. *J. Am. Chem. Soc.* **1987**, *109*, 6318–6326.
- (43) Zhao, X. G.; Richardson, W. H.; Chen, J. L.; Li, J.; Noodleman, L.; Tsai, H. L.; Hendrickson, D. N. *Inorg. Chem.* **1997**, *36*, 1198–1217.
- (44) Noodleman, L.; Norman, J. G. *J. Chem. Phys.* **1979**, *70*, 4903–4906.
- (45) Noodleman, L.; Davidson, E. R. *Chem. Phys.* **1986**, *109*, 131–143.
- (46) Noodleman, L. *J. Chem. Phys.* **1981**, *74*, 5737–5743.
- (47) Yamaguchi, K.; Fukui, H.; Fueno, T. *Chem. Lett.* **1986**, 625–628.
- (48) Yamaguchi, K. *Chem. Phys. Lett.* **1975**, *33*, 330–335.
- (49) Soda, T.; Kitagawa, Y.; Onishi, T.; Takano, Y.; Nagao, H.; Yoshioka, Y.; Yamaguchi, K. *Chem. Phys. Lett.* **2000**, *319*, 223–230.
- (50) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (51) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (52) Perdew, J.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (53) Schaefer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.
- (54) Schaefer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–2577.

- (55) Barra, A. L.; Caneschi, A.; Cornia, A.; Fabrizio de Biani, F.; Gatteschi, D.; Sangregorio, C.; Sessoli, R.; Sorace, L. *J. Am. Chem. Soc.* **1999**, *121*, 5302–5310.
- (56) Davidson, E. R. *J. Comput. Phys.* **1975**, *17*, 87–94.
- (57) Zerner, M. C. *Int. J. Quantum Chem.* **1989**, *35*, 567–575.
- (58) Boys, S. F. *Rev. Mod. Phys.* **1960**, *32*, 305–307.
- (59) Boys, S. F. *Rev. Mod. Phys.* **1960**, *32*, 300–302.
- (60) Boys, S. F. *Rev. Mod. Phys.* **1960**, *32*, 296–299.
- (61) Löwdin, P.-O. *J. Chem. Phys.* **1950**, *18*, 365–375.
- (62) Yan, B.; Chen, Z. D. *Inorg. Chem. Commun.* **2001**, *4*, 138–141.
- (63) Boudalis, A. K.; Tangoulis, V.; Raptopoulous, C. P.; Terzis, A.; Tuchagues, J. P.; Perlepes, S. P. *Inorg. Chim. Acta* **2004**, *357*, 1345–1354.
- (64) Boudalis, A. K.; Lalioti, N.; Spyroulias, G. A.; Raptopoulous, C. P.; Terzis, A.; Bousseksou, A.; Tangoulis, V.; Tucagues, J. P.; Perlepes, S. P. *Inorg. Chem.* **2002**, *41*, 6474–6487.
- (65) Gorun, S. M.; Lippard, S. J. *Inorg. Chem.* **1991**, *30*, 1625–1630.
- (66) Weihe, H.; Güdel, H. U. *J. Am. Chem. Soc.* **1997**, *119*, 6539–6543.

CT7000599

Lone-Pair Orientation Effect of an α -Oxygen Atom on $^1J_{CC}$ NMR Spin–Spin Coupling Constants in *o*-Substituted Phenols. Experimental and DFT Study

Oscar E. Taurian,^{†,‡} Rubén H. Contreras,[§] Dora G. De Kowalewski,[§]
Jorge E. Pérez,[†] and Cláudio F. Tormena^{*,||}

Department of Physics, FCEFQyN, National University of Río Cuarto, Ruta Nacional No. 36, Km 601, 5800 Río Cuarto, Argentina, Physical Chemistry Section, School of Chemistry, Biochemistry and Pharmacy, National University of San Luis, Chacabuco and Pedernera, 5700 San Luis, Argentina, Department of Physics, FCEyN, University of Buenos Aires and CONICET, Ciudad Universitaria, Pab. 1, (C1428EHA) Buenos Aires, Argentina, and Chemistry Institute, State University of Campinas, CP 6154, CEP: 13084-971, Campinas, SP, Brazil

Received February 14, 2007

Abstract: The well-known N lone-pair orientation effect on $^1J_{CC}$ spin–spin coupling constants (SSCCs) in oximes and their derivatives was used to study how negative hyperconjugative interactions of type $LP_1(O) \rightarrow \sigma^*_{CC}$ depend on ortho interactions involving the OH group. This study demanded the following analyses: (i) a qualitative estimation of how $^1J_{CC}$ SSCCs are affected by hyperconjugative interactions, (ii) a study of similar stereochemical effects to those in oximes, but in $^1J_{C_1C_2}$ and $^1J_{C_1C_6}$ in a series of 2-substituted phenols, and (iii) a quantitative estimation, with the natural bond order approach, of some key electron delocalization interactions. A few unexpected results are quoted. $LP_1(O) \rightarrow \sigma^*_{CC}$ interactions are affected by proximity interactions as follows: (a) they are enhanced by hydrogen bonds transferring charge into the $(O-H)^*$ antibonding orbital; (b) they are enhanced by proximity interactions of type $LP_1(O) \cdots H-C$; (c) they are inhibited by interactions of type $LP(O_1) \cdots H-O$. Consequences of these observations are discussed.

1. Introduction

Extensive experimental^{1–8} and theoretical^{9–15} studies of the $^1J_{^{13}C^{13}C}$ (hereafter, $^1J_{CC}$) spin–spin coupling constants, SSCCs, in oximes and their derivatives allowed determination of the stereospecificity of these couplings toward the orientation of the nitrogen lone pair. As an example, in Figure 1, the cis and trans $^1J_{CC}$ couplings in the acetone oxime are shown;³ it is observed that $^1J_{trans} - ^1J_{cis} = 7.9$ Hz. According

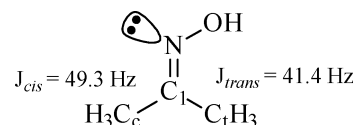


Figure 1. Experimental nitrogen lone-pair stereospecific effect on $^1J_{CC}$ couplings in acetone oxime (taken from ref 3).

to Wray,¹ the respective SSCCs are 48.42 and 40.51 Hz, and their difference amounts to 7.91 Hz.

So far, the most detailed analysis of the lone-pair orientation effect on $^1J_{CC}$ SSCCs in acetone oxime was presented by Barone et al.,⁹ and it was based on the Natural J coupling, NJC,¹⁶ dissection of SSCCs into localized molecular orbital, LMO, contributions. In short, the three main contributions discussed by Barone et al.⁹ are (a) the nitrogen lone pair, (b) the carbon–carbon bonds containing the coupling carbon

* Corresponding author phone: +55-19-3521-2092, fax: +55-19-3521-3023, e-mail: tormena@iqm.unicamp.br.

[†] National University of Río Cuarto.

[‡] National University of San Luis.

[§] University of Buenos Aires and CONICET.

^{||} State University of Campinas.

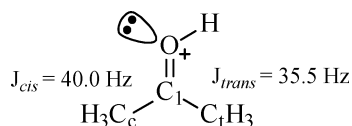


Figure 2. Experimental oxygen lone-pair stereospecific effect on $^1J_{CC}$ couplings in protonated acetone (taken from ref 19). The difference $^1J_{cis} - ^1J_{trans}$ amounts to 4.5 Hz.

atoms, and (c) the carbon inner core orbitals.¹⁷ This last contribution seems to be exaggerated within the NJC approach due to the localization procedure employed in the NJC dissection of SSCCs.¹⁸

A similar stereospecific effect of the oxygen lone pair with sp character in dicoordinated oxygen atoms in the α position to an sp^2 hybridized carbon atom was also studied.^{5,8,19–21} It is interesting to compare values displayed in Figure 1 with the corresponding experimental couplings measured in protonated acetone,¹⁹ which are displayed in Figure 2, where it is observed that the oxygen lone-pair effect is smaller than that corresponding to an N atom. Since the oxygen atom bears two nonbonding electron pairs, this stereoelectronic property cannot be used as straightforwardly as that of the N lone pair in oximes. However, its potential for studying structural problems is envisaged as excellent if a deeper insight into several aspects of this oxygen stereochemical behavior is obtained. The aim of this work is to achieve such a deeper insight. To this end, an adequate set of model compounds was sought, where $^1J_{CC}$ SSCCs could be measured at the ^{13}C natural abundance and the orientation of both lone-pair oxygen atoms should be defined by intramolecular interactions. Following such criteria, in this work was chosen a set of nine 2-X-phenol derivatives (**1**, X = H; **2**, X = CH_3 ; **3**, X = CH_2CH_3 ; **4**, X = CN; **5**, CHO; **6**, X = F; **7**, X = Cl; **8**, X = Br; **9**, X = *t*-butyl). In these phenol derivatives, $^1J_{CC}$ SSCCs were measured and a series of calculations within the density functional theory (DFT) framework were carried out, considering in all cases two conformations for the O–H group, namely, the **a** conformation, defined as that where the O–H bond eclipses the $\sigma_{C_1C_2}$ bond, and the **b** conformation, defined by the O–H bond eclipsing the $\sigma_{C_1C_6}$ bond.

Calculations were carried out for different stable side-chain conformations including the dielectric solvent effect, which was taken into account using the polarization continuum model, PCM. The following two cases were considered: (i) an isolated molecule, that is, $\epsilon = 1$, and an infinitely diluted dimethylsulfoxide (DMSO) solution, that is, $\epsilon = 46.7$. To study the dielectric solvent influence on SSCCs, their calculations were carried out employing the respective optimized geometries. The results of this study are described in terms of a pictorial representation that is expected to be useful for many scientists interested in using these results in a qualitatively predictive way. For this reason, discussions are presented in terms of the NBO approach, and interactions that define the oxygen lone pairs' stereochemical behavior are presented in terms of hyperconjugative and conjugative interactions, as it is frequently found in the chemistry literature. However, SSCC calculations are calculated within the coupled perturbed density functional theory (CP–DFT)

approach as it is implemented in the Gaussian 03 package of programs. Following this line, it can be said that the main contributions⁹ to the in-plane oxygen lone pair, hereafter $LP_1(\text{O})$, orientation effect on $^1J_{C_1C_2}$ and $^1J_{C_1C_6}$ originate in the negative hyperconjugative interactions $LP_1(\text{O}) \rightarrow \sigma_{CC}^*$, where σ_{CC} is the σ -bond orbital involving the coupling nuclei. These interactions, among other factors, should depend on both the electron acceptor ability of the σ_{CC}^* antibonding orbital and on the donor ability of the $LP_1(\text{O})$. This LP orientation effect should depend on, among other factors, the resonance interaction between both side chains and on the electrostatic interactions between them, or, more precisely, the proximity effects of both substituents placed ortho to each other. In this work, quantitative estimations of negative hyperconjugative as well as conjugative interactions were obtained using the Weinhold et al.'s natural bond orbital, NBO, approach.^{22,23}

However, an important point to be taken into account when studying the in-plane $LP_1(\text{O})$ orientation effect in 2-X-phenols is the different substituent effects that affect the $^1J_{C_1C_2}$ and $^1J_{C_1C_6}$ SSCCs. In fact, while in compound **1** the difference between these two SSCCs can be attributed almost entirely to the $LP_1(\text{O})$ orientation effect, when a substituent is placed at ring position 2, different substituent effects are introduced on $^1J_{C_1C_2}$ and $^1J_{C_1C_6}$ SSCCs. It must be recalled that the influence of substituents on $^1J_{CC}$ SSCCs in benzene derivatives was extensively studied,^{5,7,24,25} and at present, it is accepted that the inductive effect is the main substituent interaction affecting such couplings. This effect decays rapidly when the σ_{CC} bond containing the coupling nuclei departs from the ipso carbon atom bonded to the substituent. This suggests that the 2-X-inductive substituent effect on $^1J_{C_1C_2}$ is stronger than on $^1J_{C_1C_6}$. However, the OH inductive effect on both SSCCs is expected to be approximately the same.

1.1. Qualitative Theoretical Analysis of Hyperconjugative Effects on $^1J_{CC}$ SSCCs. In a recent paper,²⁶ it was shown how the CLOPPA method (Contribution from Localized Orbitals within the Polarization Propagator Approach)²⁷ can provide a qualitative prediction of how hyperconjugative interactions affect $^1J_{CH}$ SSCCs. Those considerations can easily be extended to get a qualitative estimation of how such interactions are expected to affect $^1J_{CC}$ SSCCs. An approach of this type is expected to be useful for rationalizing the stereospecific oxygen lone-pair effect on $^1J_{CC}$ SSCCs in the phenol derivatives studied in this work. In previous papers,⁹ it was observed that, of the four Ramsey terms of $^1J_{CC}$ SSCCs, Fermi contact (FC), paramagnetic spin–orbit (PSO), spin-dipolar (SD), and diamagnetic spin–orbit (DSO), only the first one determines the orientation effect of the N lone pair on $^1J_{CC}$ SSCCs. For this reason, this qualitative description is based only on the FC term, which can be written as a sum of contributions from LMOs, eq 1

$$^1J_{CC}^{\text{FC}} = \sum_{ia,jb} ^1J_{ia,jb}(\text{C}_m\text{C}_n) \quad (1)$$

where i and j are occupied LMOs, while a and b are vacant LMOs. As shown previously, the LMO contributions to the

FC term can be written as in eq 2.

$${}^1J_{ia,jb}^{\text{FC}}(C_m C_n) = W_{ia,jb} [U_{ia}(C_m) U_{jb}(C_n) + U_{ia}(C_n) U_{jb}(C_m)] \quad (2)$$

where $U_{ia}(C_m)$ [$U_{jb}(C_n)$] are the “perturbators”, that is, the matrix elements of the FC operator between the occupied i (j) and vacant a (b) LMOs evaluated at the C_m (C_n) site of the coupling nuclei, and they give a measure of the strength of the $i \rightarrow a$ ($j \rightarrow b$) virtual excitation due to that operator; $W_{ia,jb}$ are the polarization propagator matrix elements, and they correspond to the response of the electronic molecular system to the presence of the magnetic electron-nucleus FC interaction, connecting two virtual excitations $i \rightarrow a$ and $j \rightarrow b$. These matrix elements decrease when increasing the $\epsilon_{i \rightarrow a}$ and $\epsilon_{j \rightarrow b}$ energy gaps between these occupied and vacant LMOs involved in each virtual excitation. For this reason, any hyperconjugative interaction that increases any of these energy gaps should decrease the corresponding term in eq 2. On the other hand, the sum in eq 2 is largely dominated by the following two different types of terms: (1) The first is when $i = j$ corresponds to the LMO localized on the σ_{CC} bond involving the coupling nuclei, C_m and C_n , and $a = b$ corresponds to the vacant LMO localized at that σ_{CC} bond. The corresponding term in eq 2 is dubbed the “bond contribution”, J^b . For this type of coupling, this contribution is always positive. (2) The second type of term is where either i or j corresponds to the occupied LMO on the $\sigma_{C_m C_n}$ bond containing the coupling nuclei, and j or i corresponds to an occupied LMO on either other $\sigma_{C_m X}$ or $\sigma_{C_n Y}$ bonds involving either the C_m or C_n coupling nucleus, and $a = b$ corresponds to a localized vacant MO placed at that $\sigma_{C_m C_n}$ bond containing the coupling nuclei. The corresponding term in eq 2 is dubbed “other bond contribution”, J^{ob} ; for two sp^3 hybridized carbon atoms, there are six of these contributions, and for two sp^2 hybridized carbon atoms, there are four of them, two for each coupling carbon atom. However, it should be stressed that the J^{ob} terms involve also the $\sigma_{C_m C_n}$ bond and antibond containing the coupling nuclei. For this type of coupling, ${}^1J_{CC}$, these J^{ob} contributions are negative, and their absolute values are notably smaller than that of the corresponding J^b term.

It is stressed that here only a qualitative description of the effect of hyperconjugative interactions on ${}^1J_{C_1 C_2}$ and on ${}^1J_{C_1 C_6}$ SSCCs in *o*-substituted phenols is sought. Qualitatively, the effect of such interactions on both occupied and vacant LMOs, eq 1, can be described by the simple “perturbed molecular orbital theory”.²⁸ Thus, hyperconjugative interactions from the $\sigma_{C_m C_n}$ bond yield a decrease on the FC term of the ${}^1J_{C_m C_n}$ SSCC. A similar effect produces a hyperconjugative interaction into the $\sigma_{C_m C_n}^*$ antibonding orbital since both types of hyperconjugative interactions increase the energy gap between the $\sigma_{C_m C_n}$ bond and its antibond, $\sigma_{C_m C_n}^*$. On the other hand, hyperconjugative interactions from “other bonds” increase the energy gap relevant for the J^{ob} contributions since such interactions do not affect the antibonding $\sigma_{C_m C_n}^*$ orbital energy, while the “other bond” orbital energy is pushed down. As two different examples, the negative hyperconjugative interaction $\text{LP}(X) \rightarrow \sigma_{C_m C_n}^*$ or the hyperconjugative interaction from nearby σ bonds into $\sigma_{C_m C_n}^*$ can

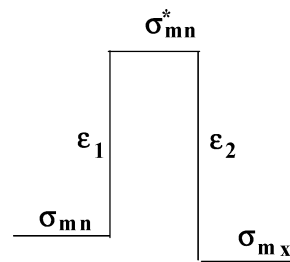


Figure 3. Schematic representation of LMOs involved in the J^b and J^{ob} contributions to ${}^1J_{C_m C_n}$ not perturbed by any hyperconjugative interaction. A $\sigma_{C_m C_n}$ hyperconjugative interaction from this bond into some antibonding orbital (not shown in this scheme, for instance σ_{YZ}^*) causes a lowering of its orbital energy, increasing the ϵ_1 energy gap, and decreasing the ${}^1J^b$ contribution, while a hyperconjugative interaction into the $\sigma_{C_m C_n}^*$ antibonding orbital, like $\text{LP}(X) \rightarrow \sigma_{C_m C_n}^*$, increases its energy, causing also an increase in the ϵ_1 energy gap and concomitantly, producing also a decrease, in the absolute value of ${}^1J^b$ contribution.

be mentioned. These two types of interactions decrease the absolute value of ${}^1J^{\text{ob}}$, and therefore, since this contribution is negative, they cause the ${}^1J_{C_m C_n}$ SSCC to increase. The effects of hyperconjugative interactions on the relevant energy gaps is displayed schematically in Figure 3.

As an example, the above qualitative considerations can be applied to rationalize the difference between both ${}^1J_{CC}$ SSCCs in acetone oxime, Figure 1. For the ${}^1J_{C_1 C_c}$ SSCC, the $\sigma_{C_1 C_c}$ bond plays the role of “the bond contribution”, J^b , while the $\sigma_{C_1 C_t}$ bond plays the role of one of the “other bond” contributions, J^{ob} , while for ${}^1J_{C_1 C_t}$, both roles are interchanged. The main hyperconjugative interaction defining the orientation of the N lone-pair stereospecific effect on ${}^1J_{CC}$ SSCCs is the $\text{LP}(N) \rightarrow \sigma_{C_1 C_t}^*$ negative hyperconjugation, which produces both a decrease in the J^b contribution to the ${}^1J_{C_1 C_t}$ SSCC and a decrease in the absolute value of the J^{ob} contribution to ${}^1J_{C_1 C_c}$. Therefore, such an interaction yields a decrease on the ${}^1J_{C_1 C_c}$ SSCC and an increase on the ${}^1J_{C_1 C_t}$ SSCC, both of them contributing to increase the ${}^1J_{C_1 C_c} - {}^1J_{C_1 C_t}$ difference.

This approach is applied in this work to rationalize the calculated and observed differences between ${}^1J_{C_1 C_2}$ and ${}^1J_{C_1 C_6}$ SSCCs in the chosen set of 2-substituted phenols mentioned above.

2. Results and Discussion

In compound **1**, employing the respective optimized geometries, ${}^1J_{CC}$ SSCCs were calculated for $\epsilon = 1$ and $\epsilon = 46.7$. For the latter dielectric constant, total ${}^1J_{CC}$ couplings are shown schematically in Figure 4, while in Table 1, the four contributions to all six ${}^1J_{CC}$ SSCCs in **1** are explicitly shown for $\epsilon = 1$ and $\epsilon = 46.7$, and they are compared with the experimental values obtained as part of this work. In Figure 4, it is observed that the difference between the calculated (for $\epsilon = 46.7$) ${}^1J_{C_1 C_6}$ and ${}^1J_{C_1 C_2}$ SSCCs is 4.6 Hz, a value quite close to the ${}^1J_{C_1 C_c} - {}^1J_{C_1 C_t}$ difference measured in protonated acetone, Figure 2, that is, 4.5 Hz.¹⁹ It is noted that, in the calculated values shown in Figure 4, a slight asymmetry is also observed for ${}^1J_{C_2 C_3}$ and ${}^1J_{C_5 C_6}$ and for ${}^1J_{C_3 C_4}$

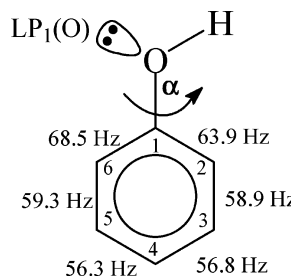


Figure 4. In phenol, **1**, for $\epsilon = 46.7$ and using the optimized geometry obtained at the B3LYP-6-311G** level, $^1J_{CC}$ SSCCs were calculated at the B3LYP-EPR III level (in Hz). The difference $^1J_{C_1C_6} - ^1J_{C_1C_2}$ amounts to 4.6 Hz, a value which is quite close to the corresponding experimental difference measured in protonated acetone.

Table 1. All Four Ramsey Contributions to the Six $^1J_{CC}$ SSCCs (in Hz) in **1** Calculated at the B3LYP/EPR-III Level Considering Both $\epsilon = 1$ and $\epsilon = 46.7^a$

SSCC	ϵ	FC	SD	PSO	DSO	total	av.	exptl.
$^1J_{C_1C_2}$	1	70.7	1.4	-6.8	0.3	65.5	67.8	65.4
	46.7	69.1	1.3	-6.8	0.3	63.9	66.2	
$^1J_{C_1C_6}$	1	75.3	1.4	-6.8	0.3	70.2	67.8	65.4
	46.7	73.6	1.3	-6.8	0.3	68.5	66.2	
$^1J_{C_2C_3}$	1	66.2	1.3	-7.1	0.2	60.6	60.8	58.0
	46.7	64.6	1.3	-7.2	0.2	58.9	59.1	
$^1J_{C_3C_4}$	1	64.4	1.3	-7.2	0.2	58.7	58.4	56.5
	46.7	62.5	1.3	-7.2	0.2	56.8	56.6	
$^1J_{C_4C_5}$	1	63.6	1.3	-7.1	0.2	58.0	58.4	56.5
	46.7	62.0	1.2	-7.1	0.2	56.3	56.6	
$^1J_{C_5C_6}$	1	66.7	1.4	-7.3	0.2	61.0	60.8	58.0
	46.7	65.1	1.3	-7.3	0.2	59.3	59.1	

^a Since both planar conformers are equivalent, individual SSCCs are not amenable to measurement. Total couplings, properly averaged, are compared with the corresponding experimental values measured as part of this work.

and $^1J_{C_4C_5}$, respectively. Experimentally, these differences, as it also happens with $^1J_{C_1C_6}$ and $^1J_{C_1C_2}$ SSCCs, cannot be observed since, due to the equivalence of both planar rotamers of compound **1**, only the average of the respective values is amenable to measurement.

Trends of noncontact terms displayed in Table 1 are similar to those observed for other aromatic compounds;^{21,29,30} the most important of them is the PSO term, although the SD term cannot be neglected. The “LP₁(O) orientation effect”, that is, the $^1J_{C_1C_6} - ^1J_{C_1C_2}$ difference is 4.7 Hz for $\epsilon = 1$, while it is 4.6 Hz for $\epsilon = 46.7$. In these results converge two different trends; that is, negative hyperconjugative interactions are slightly inhibited by a highly polar solvent³¹ and the effect of electrostatic interactions of type C₆-H₆···LP₁(O) and C₂-H₂···O-H³²; a highly polar solvent shields them.

For all $^1J_{CC}$ SSCCs in **1**, Table 1, the dielectric solvent effect causes a decrease within the range 1.6–1.8 Hz, which originates mainly in the respective FC contribution. Even though the inclusion of the dielectric solvent effect improves the agreement between total calculated and experimental SSCCs, their total calculated values are slightly overestimated in about 1 Hz. However, the experimental trends are correctly reproduced.

Two important questions to be answered for 2-X-phenols are these: how does the X side-chain affect the O–H conformation, and how much are the relative populations of the **a** and **b** conformations affected by a highly polar solvent? To answer these questions, the geometries of compounds **2–8** were optimized for different side-chain conformations (see Figure 5 for conformations considered when X is a nonlinear substituent). In Table 2 are shown the relative energies for the different conformers considered in this work. For X = CH₃, **2**; X = Et, **3**; and X = CN, **4**, **b** is the preferential conformation, even for $\epsilon = 1$. For a highly polar solvent in these three compounds, the **a** conformation is notably more destabilized. For conformation **a** of compound **4**, using the NBO method, a charge-transfer interaction $\pi_1(\text{CN}) \rightarrow (\text{O}-\text{H})^* = 0.8$ kcal/mol is calculated [$\pi_1(\text{CN})$ stands for the π symmetry LMO with lowest energy]. For X = CHO, **5**, as expected, the preferential conformation for the OH group is **a** due to the strong intramolecular hydrogen bond that takes place between both side-chain groups. For a polar solvent, **a** is still the preferential conformation, although the energy of **5-(b-2)** is only 2.5 kcal/mol above that of the **a** conformation, see Figure 5. It is noted that for the **a** conformation the NBO analysis yields the H proton of the OH group as a separate unit. This is interpreted as being a very strong intramolecular hydrogen bond, but unfortunately, the magnitude of the LP₁(O) $\rightarrow \sigma^*_{C_1C_2}$ interaction cannot be considered to be reliable.

For $\epsilon = 1$, 2-X-phenols (X = F, Cl, Br) **6**, **7**, and **8** show as preferential the **a** conformation, suggesting that an intramolecular hydrogen bond of type O–H···X is operating. Such hydrogen bonds are expected to be mainly electrostatic in character.³³ It must be emphasized that in this work only the optimized geometries of a few obvious conformers were sought to study certain aspects of the stereospecificity of LP₁(O) on the $^1J_{C_1C_6} - ^1J_{C_1C_2}$ difference. Detailed studies of conformers of 2-substituted phenols were reported recently.³⁴ It is interesting to note that, according to the NBO approach, for the **6-(a)** conformation, the LP₂(F) $\rightarrow (\text{O}-\text{H})^*$ interaction is 0.8 kcal/mol (LP₂ stands for the in-plane F lone pair with important p character), which is of similar strength to that of the $\pi_1(\text{CN}) \rightarrow (\text{O}-\text{H})^*$ charge-transfer interaction calculated for **4-(a)**. For X = Cl and Br, the analogous charge-transfer interactions are weaker than 0.5 kcal/mol. In Table 3, the H···X distances for the **a** conformation of compounds **6**, **7**, and **8** are compared for $\epsilon = 1$ and $\epsilon = 46.7$. In all three cases, an important increase of the H···X distance is observed for $\epsilon = 46.7$. For a high polar solvent, all three of these compounds show as preferential the **b** conformation, although for X = F the **6-(a)** conformation is only 0.25 kcal/mol above the **6-(b)** conformation. Probably, in this case, the observed $^1J_{CC}$ SSCC shows a non-negligible contribution from conformation **6-(b)**. When calculating $^1J_{C_1C_6}$ and $^1J_{C_1C_2}$ SSCCs in these compounds, some caution should be exercised, since both ortho substituents contain an α electron-rich atom, and it is known that under such conditions DFT-calculated SSCCs could yield unreliable results for SSCCs.¹⁷

In Table 4, for different conformations of compounds **2–8** (see Table 2 and Figure 5), total calculated couplings for $\epsilon = 1$ and $\epsilon = 46.7$ are compared with the experimental values

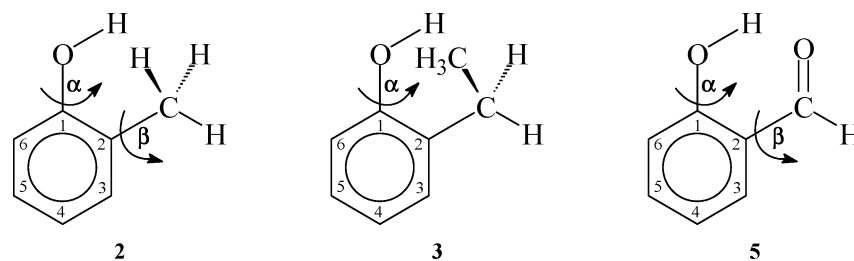


Figure 5. Conformations of compounds **2**, **3**, and **5**. For **2**, (**a-1**) $\alpha = 0^\circ$, $\beta = 0^\circ$; (**b-1**) $\alpha = 180^\circ$, $\beta = 0^\circ$; (**a-2**) $\alpha = 0^\circ$, $\beta = 180^\circ$; (**b-2**) $\alpha = 180^\circ$, $\beta = 180^\circ$. For **3**, (**a**) $\alpha = 0^\circ$ and the dihedral angle $C_{Me}-C_{CH_2}-C_2-C_3 = 77.8^\circ$; (**b**) $\alpha = 180^\circ$ and the dihedral angle $C_{Me}-C_{CH_2}-C_2-C_3 = 78.3^\circ$. For **5**, (**a**) $\alpha = 0^\circ$, $\beta = 0^\circ$; (**b-1**) $\alpha = 180^\circ$, $\beta = 0^\circ$; (**b-2**) $\alpha = 180^\circ$, $\beta = 180^\circ$.

Table 2. Relative Energies (kcal/mol) of Conformers **a** and **b** of Compounds **2–8**, 2-X-Phenols, Calculated at the B3LYP/6-311G** Level of Approximation Considering an Isolated Molecule, $\epsilon = 1$, and an Infinitely Diluted DMSO Solution, $\epsilon = 46.7^a$

	X	$\epsilon = 1$	$\epsilon = 46.7$
2-(a-1)	CH ₃ -(a-1)	2.05	4.24
2-(b-1)	CH ₃ -(b-1)	0.89	0.97
2-(a-2)	CH ₃ -(a-2)	0.52	3.02
2-(b-2)	CH ₃ -(b-2)	0.0	0.0
3-(a)	Et-(a)	0.64	2.57
3-(b)	Et-(b)	0.0	0.0
4-(a)	CN-(a)	1.30	2.43
4-(b)	CN-(b)	0.0	0.0
5-(a)	CHO-(a)	0.0	0.0
5-(b)	CHO-(b)	11.60	4.25
5-(b-2)	CHO-(b-2)	8.62	2.52
6-(a)	F-(a)	0.0	0.25
6-(b)	F-(b)	3.02	0.0
7-(a)	Cl-(a)	0.0	1.41
7-(b)	Cl-(b)	3.06	0.0
8-(a)	Br-(a)	0.0	9.86
8-(b)	Br-(b)	3.12	0.0

^a For the definition of the different conformations see Figure 5.

Table 3. The **a** Conformation of Compounds **6–8**, for Which the Optimized H–X Distances (in Å) for $\epsilon = 1$ and $\epsilon = 46.7$ Yield an Idea of the Electrostatic Character of These Interactions

compound	H...X	$\epsilon = 1$	$\epsilon = 46.7$
6	H...F	2.205	2.320
7	H...Cl	2.417	2.526
8	H...Br	2.511	2.617

measured as part of this work. For the sake of completeness, in the same table are shown also the experimental ${}^1J_{CC}$ SSCCs measured in 2-*t*-butyl-phenol as part of this work. No calculations were performed in this compound. In each case, the calculated preferential conformation is marked with an asterisk (*). It is observed that the best agreement between calculated and experimental values is obtained for the preferential conformation considering $\epsilon = 46.7$. Calculated SSCCs are overestimated in an amount somewhat above 1 Hz, except for X = F, Cl, and Br, that is, for compounds where the α atoms of both substituents are electron-rich atoms. In this case, DFT-B3LYP-calculated SSCCs are expected to be less reliable than in other types of compounds.

For compounds **2–6**, the difference $\Delta = {}^1J_{C_1C_6} - {}^1J_{C_1C_2}$ between these two SSCCs for the preferential conformation is also in better agreement with its experimental value. That difference Δ depends markedly on the side-chain conformations of both the OH and the X moieties. Since the inductive effect of both substituents is not expected to be sensitive to their conformation, the sensitivity of Δ to the side-chain conformations suggests that the negative hyperconjugative interactions $LP_1(O) \rightarrow \sigma^*_{C_1C_2}$ and $LP_1(O) \rightarrow \sigma^*_{C_1C_6}$ are sensitive to proximity interactions between both side chains.

In order to get a rationalization of how changes in the negative hyperconjugative interactions take place when changing the side-chain conformations, in Table 5 are displayed the aromatic $C_1=C_2$ and $C_1=C_6$ bond lengths as well as, for the OH group, the $LP_1(O) \rightarrow \sigma^*_{C_1C_2}$ and $LP_1(O) \rightarrow \sigma^*_{C_1C_6}$ negative hyperconjugative interactions and conjugative interactions of type $LP_2(O) \rightarrow \pi^*$, where $LP_2(O)$ stands for the oxygen lone pair of pure π character. Values displayed in Table 5 correspond to calculations performed considering $\epsilon = 46.7$, since in this way experimental conditions are better reproduced. For compounds **2**, **3**, **4**, and **5**, the $C_1=C_2$ bond length is a bit longer than the $C_1=C_6$ bond, irrespective of the **a** or **b** conformation of the OH group. This suggests that the different strengths of the $LP_1(O) \rightarrow \sigma^*_{C_1C_2}$ interactions are not much affected by that bond length effect; indeed, when the corresponding bond lengths in compound **1** are compared, it is estimated that such interaction lengthens the antiperiplanar C=C bond with respect to $LP_1(O)$ only by about 0.001 Å. For compound **2**, it is observed that the strength of the $LP_1(O) \rightarrow \sigma^*_{CC}$ negative hyperconjugative interaction is notably enhanced for the **b** conformation. In fact, while for the **2-(a-1)** and **2-(a-2)** conformations they are, respectively, 7.6 and 7.4 kcal/mol (i.e., they are very close to the corresponding interaction in compound **1**, 7.4 kcal/mol), for the **2-(b-1)** and **2-(b-2)** conformations, they are 10.2 and 10.3 kcal/mol, respectively. It is evident that this marked difference between both types of hyperconjugative interactions is an important factor for determining as preferential the **b** conformation. At first sight, this marked difference for the negative hyperconjugative interaction involving the $LP_1(O)$ for the two planar conformations of the OH group seems to indicate that the $\sigma^*_{C_1C_6}$ antibonding orbital is a better acceptor than the $\sigma^*_{C_1C_2}$ antibonding orbital since the former corresponds to a bond shorter than the latter. However, in compound **4**, those bond lengths show a larger bond length difference than in

Table 4. Comparison between Calculated (for $\epsilon = 1$ and $\epsilon = 46.7$) and Experimental $^1J_{CC}$ SSCCs (Hz) in 2-X-Phenols Studied in This Work^a

X		$^1J_{C_1C_2}$	$^1J_{C_1C_6}$	Δ^b	$^1J_{C_2C_3}$	$^1J_{C_3C_4}$	$^1J_{C_4C_5}$	$^1J_{C_5C_6}$
CH ₃ -(a-1)	1	66.2	71.6	5.4	61.9	60.0	57.2	61.3
	46.7	70.5	66.9	-3.6	61.4	59.3	57.9	61.1
CH ₃ -(b-1)	1	70.5	66.9	-3.6	61.4	59.3	57.9	61.1
	46.7	69.1	65.0	-4.1	60.3	57.6	56.1	59.4
CH ₃ -(a-2)	1	64.3	71.9	6.6	62.9	59.2	57.8	61.0
	46.7	63.8	70.6	6.8	61.2	57.3	56.0	59.3
CH ₃ -(b-2)(*)	1	68.8	67.3	-1.5	62.4	58.5	58.6	60.7
	46.7	67.4	65.5	-1.9	61.3	56.8	56.7	59.0
	exptl.	66.1	64.3	-1.8	59.3	56.4	56.0	57.5
Et-(a)	1	64.4	71.7	7.3	62.2	59.2	57.7	61.0
	46.7	64.2	70.3	6.1	60.1	57.3	55.9	58.8
Et-(b)(*)	1	69.0	67.1	-1.9	61.9	58.5	58.4	60.7
	46.7	67.6	65.2	-2.4	60.7	56.9	56.6	59.0
	exptl.	66.5	65.7	-0.8	59.3	56.6	56.0	57.6
CN-(a)	1	66.4	71.1	4.7	64.2	60.2	56.7	60.9
	46.7	68.3	69.9	1.6	61.5	58.9	54.5	59.4
CN-(b)(*)	1	74.1	66.4	-7.7	63.5	59.3	57.4	60.6
	46.7	72.9	64.8	-8.1	62.4	58.2	55.2	58.9
	exptl.	71.1	63.5	-7.6	61.1	57.2	55.0	57.2
CHO-(a)(*)	1	60.4	68.2	5.8	61.5	60.6	55.9	61.2
	46.7	60.4	67.7	7.3	60.1	58.6	54.2	59.4
CHO-(b-1)	1	69.7	64.3	-5.4	62.1	59.4	57.0	61.0
	46.7	68.1	62.9	-5.2	60.5	57.9	54.7	59.2
CHO-(b-2)	1	69.6	65.2	-4.4	61.2	59.4	56.9	61.1
	46.7	68.5	63.0	-5.5	60.1	58.5	54.5	59.4
	exptl.	61.3	66.7	5.4	59.0	57.9	54.5	58.3
F-(a)(*)	1	76.5	72.4	-4.1	77.5	59.7	58.7	61.2
	46.7	76.9	70.9	-6.0	76.1	58.2	56.7	59.7
F-(b)	1	82.4	67.3	-15.1	76.6	59.5	58.9	60.9
	46.7	81.2	65.8	-15.4	75.8	57.9	57.0	59.3
	exptl.	76.7	68.2	-8.5	73.6	57.5	56.7	58.4
Cl-(a)	1	72.3	71.4	-0.9	71.4	59.0	57.9	61.7
	46.7	72.7	70.0	-2.7	69.9	57.4	56.0	60.2
Cl-(b)(*)	1	79.2	66.7	-12.5	70.8	58.5	58.3	61.3
	46.7	78.3	64.9	-13.4	70.1	56.9	56.5	59.7
	exptl.	72.5	67.3	-5.2	67.4	56.5	56.0	58.4
Br-(a)	1	70.4	70.9	0.5	68.9	58.4	57.9	61.9
	46.7	70.7	69.6	-1.1	67.3	56.7	56.0	60.3
Br-(b)(*)	1	77.9	66.3	-11.6	68.4	57.8	58.5	61.4
	46.7	76.6	64.4	-12.2	67.5	56.2	56.5	59.9
	exptl.	71.5	67.0	-4.5	65.8	56.0	56.0	58.5
t-Bu	exptl.	67.3	64.7	-2.6	59.7	56.6	55.6	57.5

^a An asterisk denotes the preferential conformation (see Table 2). ^b $\Delta = ^1J_{C_1C_6} - ^1J_{C_1C_2}$.

compound **2**, and the corresponding LP₁(O) negative hyperconjugative interactions do not show such a difference. Therefore, it is thought that the above-mentioned differences originate mainly in proximate interactions between the LP₁(O) and the methyl group; a similar assertion holds for compound **3**. The dielectric solvent effect on such differences is worth noting. For the four conformations of compound **2** shown in Figure 5, the negative hyperconjugative interactions, LP₁(O) $\rightarrow \sigma^*_{CC}$, and the conjugative interactions, LP₂(O) $\rightarrow \pi^*$, are displayed in Table 6 as calculated for $\epsilon = 1$ and $\epsilon = 46.7$. In the same Table 6, the corresponding interactions in **1** are also shown for reference purposes. To rationalize adequately the data displayed in Table 6, it is important to recall that in a previous paper³⁵ it was studied

how electrostatic interactions can inhibit or enhance electron delocalization interactions.²⁶ Keeping these trends in mind, it is expected that in Table 6 there is not a general trend for the influence of the dielectric solvent on the displayed interactions. While a polar solvent enhances the LP₁(O) $\rightarrow \sigma^*_{CC}$ interaction for the **a** conformation, an important inhibition for the **b-1** conformation is observed. This suggests that, for the **a** conformation, interactions of type O-H \cdots H-C, which are mainly electrostatic, slightly inhibit the LP₁(O) $\rightarrow \sigma^*_{C_1C_2}$ interactions. A polar solvent shields the O-H \cdots H-C interaction (Figure 6), and therefore the negative hyperconjugative interaction LP₁(O) $\rightarrow \sigma^*_{C_1C_2}$ tends to recover its original strength, that is, for $\epsilon = 1$. For the **b** conformations, an important increase in LP₁(O) $\rightarrow \sigma^*_{CC}$ is

Table 5. $C_1=C_2$ and $C_1=C_6$ Bond Lengths, $d_{C_1C_2}$ and $d_{C_1C_6}$, Respectively (in Å); Negative Hyperconjugative Interactions, of Types $HI_{1,2} = LP_1(O) \rightarrow \sigma^*_{C_1C_2}$ and $HI_{1,6} = LP_1(O) \rightarrow \sigma^*_{C_1C_6}$; and Conjugative Interactions, CI, of Types $CI_{1,2} = LP_2(O) \rightarrow \pi^*_{C_1C_2}$ and $CI_{1,6} = LP_2(O) \rightarrow \pi^*_{C_1C_6}$ ^a

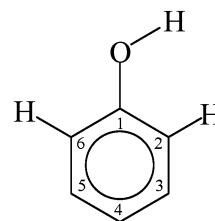
X	$d_{C_1C_2}$	$d_{C_1C_6}$	$HI_{1,2}$	$HI_{1,6}$	$CI_{1,2}$	$CI_{1,6}$
H	1.400	1.399	<0.5	7.4	28.9	<0.5
CH ₃ -(a-1)	1.405	1.401	7.6	<0.5	27.2	<0.5
CH ₃ -(b-1)	1.406	1.400	<0.5	10.2	<0.5	26.9
CH ₃ -(a-2)	1.406	1.397	7.4	<0.5	26.9	<0.5
CH ₃ -(b-2)	1.408	1.397	<0.5	10.3	<0.5	28.2
Et-(a)	1.407	1.397	7.6	<0.5	20.9	1.4
Et-(b)	1.408	1.397	<0.5	10.3	<0.5	28.1
CN-(a)	1.412	1.400	8.9	<0.5	34.5	<0.5
CN-(b)	1.412	1.398	0.9	7.6	34.1	<0.5
CHO-(a) ^b	1.419	1.399	7.8	<0.5	36.1	<0.5
CHO-(b-1)	1.417	1.400	0.7	7.1	<0.5	33.8
CHO-(b-2)	1.413	1.401	0.9	6.8	32.7	<0.5
F-(a)	1.396	1.397	10.1	<0.5	<0.5	27.9
F-(b)	1.400	1.397	<0.5	10.9	29.0	<0.5
Cl-(a)	1.401	1.401	8.1	<0.5	32.4	<0.5
Cl-(b)	1.402	1.400	0.7	7.6	32.2	<0.5
Br-(a)	1.401	1.402	8.5	0.6	32.2	<0.5
Br-(b)	1.402	1.401	<0.5	7.5	32.1	<0.5

^a All of these interactions are given in kilocalories per mole. ^b For the **a** conformation, the NBO parameters are not reliable since the NBO approach yields the OH proton as a separated unit.

Table 6. Dielectric Solvent Influence on the $LP_1(O) \rightarrow \sigma^*_{CC}$ and $LP_2(O) \rightarrow \pi^*_{CC}$ Interactions (in kcal/mol) for Conformations **a-1**, **b-1**, **a-2**, and **b-2** of Compound **2**

conformer	interaction	$\epsilon = 1$	$\epsilon = 46.7$
X = H	$LP_1(O) \rightarrow \sigma^*_{CC}$	7.1	7.4
	$LP_2(O) \rightarrow \pi^*_{CC}$	28.2	28.9
a-1	$LP_1(O) \rightarrow \sigma^*_{CC}$	7.1	7.6
	$LP_2(O) \rightarrow \pi^*_{CC}$	27.5	27.3
b-1	$LP_1(O) \rightarrow \sigma^*_{CC}$	12.2	10.2
	$LP_2(O) \rightarrow \pi^*_{CC}$	25.7	26.9
a-2	$LP_1(O) \rightarrow \sigma^*_{CC}$	6.9	7.4
	$LP_2(O) \rightarrow \pi^*_{CC}$	27.1	26.9
b-2	$LP_1(O) \rightarrow \sigma^*_{CC}$	10.0	10.3
	$LP_2(O) \rightarrow \pi^*_{CC}$	27.0	28.2

observed. For **b-1** and $\epsilon = 1$, it amounts to 12.2 kcal/mol, and that decreases to 10.2 kcal/mol for $\epsilon = 46.7$. This suggests that the proximity between the $LP_1(O)$ and the methyl C–H bonds, that is, interactions of type $LP_1(O) \cdots H-C$, enhances notably negative hyperconjugative interactions of type $LP_1(O) \rightarrow \sigma^*_{CC}$. A polar solvent should shield the former, and therefore the $LP_1(O) \rightarrow \sigma^*_{CC}$ interaction should recover in part its original value; for example, for **b-1**, it goes from 12.2 to 10.2 kcal/mol. It is interesting to observe the opposite behavior for the **b-2** conformation, where a methyl C–H bond points directly to $LP_1(O)$. It is stressed that, in compound **2-(b-2)**, the difference $\Delta = J_{C_1C_6} - J_{C_1C_2} = 65.5 - 67.4 \text{ Hz} = -1.9 \text{ Hz}$ is in excellent agreement with the experimental difference, -1.8 Hz . This difference is smaller, in absolute value, than for any other conformation. This trend seems to originate in the small value for the $J_{C_1C_2}$ SSCC, which is affected by the $(C_{Me}-H) \rightarrow \sigma^*_{C_1C_2} = 2.5 \text{ kcal/mol}$ σ -hyperconjugative interaction since



	$^1J_{C_6H}$	$^1J_{C_2H}$	$LP_1(O) \rightarrow \sigma^*_{C_1C_2}$
$\epsilon = 1$	165.2	158.4	7.1
$\epsilon = 46.7$	164.4	163.1	7.4

Figure 6. Interplay between electrostatic and negative hyperconjugative interactions in phenol. SSCCs are in hertz, and negative hyperconjugative interactions are in kilocalories per mole. The $C_6-H \cdots LP_1(O)$ interaction inhibits in part the negative hyperconjugative interaction $LP_1(O) \rightarrow \sigma^*_{C_1C_2}$. This effect yields an algebraic increase of the $\sigma_{C_1C_2}$ “other bond contribution” to $^1J_{C_2H}$, decreasing its total value. The proximity between the C_6-H bond and the $LP_1(O)$ lone pair causes an increase in the $^1J_{C_6H}$ SSCC. These values are taken from ref 32.

the in-plane $C_{Me}-H$ bond is placed in an anti-periplanar configuration with respect to the $C_1=C_2$ bond. This same effect seems to be present for the **2-(a-2)** conformation affecting the difference $^1J_{C_1C_6} - ^1J_{C_1C_2} = 6.8 \text{ Hz} = 70.6 - 63.8 \text{ Hz}$, which is larger than usual (compare with compound **1**, Figure 4), suggesting that the $^1J_{C_1C_2}$ SSCC is reduced by the mentioned σ -hyperconjugative interaction. It is to be noted that in both cases such a reduction in the $^1J_{C_1C_2}$ SSCC is between 2.3 and 2.6 Hz. The enhancement of the negative hyperconjugative interaction $LP_1(O) \rightarrow \sigma^*_{CC}$ due to the proximity effects between $LP_1(O)$ and the methyl group seems to play a key role in defining as more preferential the **b** conformation with respect to the **a** conformation.

Another important point to note in Table 5 is that the conjugative $LP_2(O) \rightarrow \pi^*$ interaction depends on both the methyl and the hydroxyl groups conformations; that is, this interaction is also affected by proximity interactions between the $LP_2(O)$ and the methyl C–H bonds. Quite similar effects are also observed for the **a** and **b** conformations of compound **3**, although for the former conformation, a very strong inhibition of the conjugative effect takes place, probably due to the closeness between the π -type oxygen lone pair, $LP_2(O)$, and the methyl moiety of the ethyl group. It is important to note that for compound **3** only one conformation of the CH_3 moiety was considered. Comparing the conjugative interaction for **2-(b-1)** and **2-(b-2)**, it is suggested that the larger inhibition of the conjugative effect that takes place in the former defines the latter as the preferential conformation. While in **2-(b-1)** there are two out-of-plane C–H methyl bonds close to the π -oxygen lone pair, in **2-(b-2)**, there is only one in-plane C–H methyl bond that points to the node of that LMO representing the $LP_2(O)$ nonbonding electron pair.

In compound **4-(a)**, the $LP_1(O) \rightarrow \sigma^*_{CC}$ negative hyperconjugative interaction is stronger than that in the **4-(b)**

conformation. This seems to indicate that a charge-transfer interaction into the (O–H)* antibonding orbital enhances the corresponding negative hyperconjugative interaction $LP_1(O) \rightarrow \sigma^*_{CC}$. On the other hand, in **4-(b)**, the $LP_1(O) \rightarrow \sigma^*_{CC}$ interaction is close to that in **1**. The expected enhancement of the $LP_2(O) \rightarrow \pi^*$ resonance for the electron-donating O–H group when it is in a position ortho to an electron-withdrawing X substituent is observed in compounds **4** and **5**, although in the latter, this effect is partially inhibited for the **b** conformation of the OH group. On the other hand, in **4**, the important resonance enhancement, $LP_2(O) \rightarrow \pi^*$, due to the electron-donor OH group and the electron-withdrawing CN group depends only slightly on the O–H conformation. How much does the charge-transfer interactions into the (O–H)* antibond observed for the **a** conformations of compounds **4**, **6–8** affect the negative hyperconjugative interactions of type $LP_1(O) \rightarrow \sigma^*_{CC}$? In **4-(a)**, $LP_1(O) \rightarrow \sigma^*_{C_1C_2}$ amounts to 8.9 kcal/mol, while for the **4-(b)** conformation, $LP_1(O) \rightarrow \sigma^*_{C_1C_6}$ amounts to 7.6 kcal/mol. This suggests that, when the (O–H)* antibond participates in an interaction like that in **4-(a)**, then the negative hyperconjugative interaction $LP_1(O) \rightarrow \sigma^*_{C_1C_2}$ is enhanced. It is observed that a similar effect seems to be operating for the **6-(a)** compound, where the O–H...F hydrogen bond shows also a strong electrostatic character. It is to be noted that in **6-(b)** there is also an important enhancement of the electron-donor ability of $LP_1(O)$ due to the proximity between $LP_1(O)$ and $LP_2(F)$. This interaction, as commented upon previously,²⁶ favors the $LP_1(O) \rightarrow \sigma^*_{C_1C_6}$ charge-transfer interaction. In **7-(a)** and **8-(a)**, a similar effect to that in **6-(a)** can be appreciated, although it is a weaker effect, a fact that is consistent with a weaker electrostatic hydrogen bond involving Cl or Br rather than F (see the respective H...X distances in Table 3).

In order to verify the trends commented upon above for charge-transfer interactions in **4**, the geometry of the **a** conformer 2-OH-phenol was optimized at the same level as compounds **1–8**, and the relevant charge-transfer interactions were calculated; they are displayed in Figure 7. These results are compatible with these comments. (i) The $LP_1(O_7) \rightarrow \sigma^*_{C_1C_2}$ negative hyperconjugative interaction is enhanced due to the $LP_1(O_8) \rightarrow \sigma^*_{OH}$ hydrogen bond. (ii) The charge-transfer interaction involved in that hydrogen bond is weaker in 2-OH-phenol than those reported above in compounds **4** and **6**. (iii) The $LP_1(O_8) \rightarrow \sigma^*_{C_2C_3}$ negative hyperconjugative interaction is slightly weaker than the corresponding interaction in compound **1**. This comment is in line with results discussed previously³⁵ on the slight inhibition of a negative hyperconjugation when the corresponding lone pair is involved in a standard hydrogen bond. (iv) The conjugative interactions of both $LP_2(O)$'s show a slight inhibition typical of two-electron-donor substituents placed ortho to each other. However, this effect is more important for $LP_2(O_8)$ than for $LP_2(O_7)$, indicating that the hydrogen-bond acceptor shows a larger inhibition than the hydrogen-bond donor, which actually seems to be enhanced by the hydrogen-bond interaction. (v) It is interesting to observe the different solvent trends exhibited by the different interactions displayed in Figure 7. In general, they are compatible with observations i–iv, which tend to confirm the effects of proximate

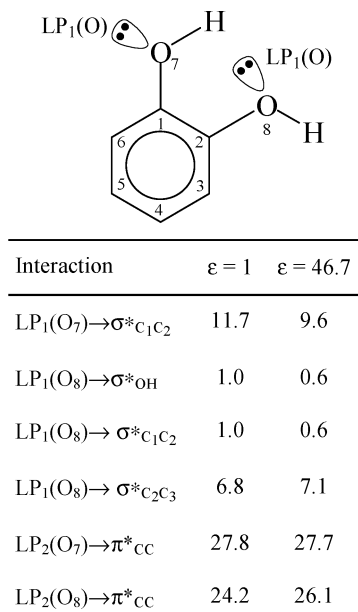


Figure 7. Negative hyperconjugative and conjugative interactions in 2-OH-phenol for the **a** conformation as calculated for $\epsilon = 1$ and $\epsilon = 46.7$.

interactions on negative hyperconjugative and conjugative interactions.

3. Experimental and Computational Details

NMR Measurements. Compounds studied in this work are commercially available, and their identities and purities were checked by taking their 1H and ^{13}C NMR spectra. Such spectra were recorded using 5 mm sample tubes from DMSO–DMSO- d_6 solutions at 30 °C in concentrations of ca. 80% w/w and were run on a Bruker AM 500 spectrometer, operating at 125.76 MHz when observing ^{13}C signals. $^1J_{^{13}C^{13}C}$ coupling constants were measured at natural abundance using the INADEQUATE technique,³⁶ which was adjusted for $^1J_{^{13}C^{13}C} = 60$ Hz. ^{13}C signals were first assigned from the 1H -coupled ^{13}C spectra and then confirmed when performing the INADEQUATE experiments. According to the acquisition parameters used, the digital resolution was in the range of 0.07–0.21 Hz per point. All spectra were recorded at the LANAIS NMR-500 facility of the Department of Physics, FCEyN, University of Buenos Aires.

Computational Details. All DFT calculations carried out in this work were performed using the hybrid B3LYP functional of Lee et al.,³⁷ where the exchange part is treated according to Becke's three-parameter approach.³⁸ In all cases, the geometries of compounds **1–8** were fully optimized at the B3LYP/6-311G** level considering both an isolated molecule, $\epsilon = 1$, and a molecule in an infinitely diluted DMSO solution, $\epsilon = 46.7$. In each case, two different conformations of the hydroxyl group were considered, namely, **a**, with the O–H bond eclipsing the aromatic $C_1=C_2$ bond, and **b**, with the O–H bond eclipsing the aromatic $C_1=C_6$ bond. All conformations were verified to correspond to true minima on the potential surface.

Besides, for X = CH₃, two different conformations of the methyl group were considered yielding, therefore, four stable rotamers **a-1**, **a-2**, **b-1**, and **b-2**, see Figure 5. In compound

5, that is, with $X = \text{CHO}$, for the **b** conformation of the hydroxyl group, two different conformations of the CHO group were considered, namely, **b-1** with $\beta = 0^\circ$ and **b-2** with $\beta = 180^\circ$ (Figure 5). It is important to recall that samples were prepared using in general high solute concentrations, and therefore these solutions depart from the “infinitely diluted DMSO” solution model. Besides, in a previous paper, it was observed that dielectric solvent effects on aromatic $^1J_{\text{CC}}$ SSCCs show a saturation effect for ϵ slightly larger than 10.²⁹ For this reason, PCM calculations considering $\epsilon = 46.7$ are thought just to take into account a highly polar solution and not only an infinitely diluted DMSO solution.

For compounds **1–6**, all four isotropic terms of $^1J_{\text{CC}}$ SSCCs, that is, the FC, SD, PSO, and DSO, were calculated using the EPR-III basis set³⁹ which is of a triple- ζ quality and includes diffuse and polarization functions. The s part of this basis set is enhanced to better reproduce the electronic density in the nuclear region; this point is particularly important when calculating the FC term. It is to be stressed that coupling constants calculated at the B3LYP/EPR-III level are close to the basis-set converged values.⁴⁰ For Cl- or Br-containing compounds, inner-shell electrons were taken into account by using the LANL2DZ effective core potential. In these compounds, for all other atoms, the EPR-III basis set was used. The CP-DFT perturbative approach was used for calculating all three of the second-order terms of SSCCs, that is, FC, SD, and PSO. SSCC calculations were performed using the respective optimized geometry. Dielectric solvent effects were taken into account using the SCRF-PCM version of the PCM of Tomasi et al.⁴¹ Calculations of SSCCs including dielectric solvent effects were performed using optimized geometries obtained within the SCRF-PCM model using the respective dielectric constant, for $\epsilon = 1$ and $\epsilon = 46.7$. All DFT calculations were carried out with the Gaussian 03 program.⁴² Conjugative and hyperconjugative interactions were calculated using the NBO approach^{22,43} as included in the Gaussian 03 suite of programs.⁴²

4. Concluding Remarks

The aim of this work was to study in detail some stereochemical effects of oxygen nonbonding electron pairs on $^1J_{\text{CC}}$ SSCCs. It is expected that the results presented in this work will contribute to supporting the use $^1J_{\text{CC}}$ SSCCs as adequate probes for certain stereochemical studies in dicoordinated oxygen-containing aromatic compounds. This study was carried out on a series of nine 2-X-phenols where, for each compound, the six $^1J_{\text{CC}}$ SSCCs were measured. DFT calculations were carried out for different side-chain conformations in most of the nine 2-X-phenols, and stereochemical oxygen lone-pair effects were nicely verified, obtaining very good agreement between measured and calculated $^1J_{\text{CC}}$ couplings. This same study allowed a detailed analysis of some interesting features of ortho interactions and their effect on side-chain conformations. One of these unexpected results is this one: the OH mesomeric effect can be much affected by the ortho interaction with C–H bonds. The stereospecific properties of $^1J_{\text{CC}}$ SSCCs depend more strongly than expected on ortho interactions, and this calls for some caution when

using these parameters as probes to study some stereochemical aspects.

It is important to recall that ortho interactions calculated in this work do not depend on any molecular orbital model employed in this analysis. However, since it is expected that within the readership of this journal there are many chemists familiar with concepts such as conjugative and hyperconjugative interactions, it is considered important to describe results obtained in this work in terms of a pictorial representation resorting to that type of interaction. In this way, a qualitative or semiquantitative model with interesting predictive character in terms of NBOs is used to discuss the main results obtained in this work.

Results presented above suggest that negative hyperconjugative and conjugative interactions involving a two-coordinated oxygen atom acting as the α atom of a substituent in an aromatic ring strongly depend on ortho interactions. Such effects present several subtleties that are worth mentioning in detail, indicating at the same time possible applications. They are as follows:

(1) $\text{LP}_1(\text{O}) \rightarrow \sigma^*_{\text{CC}}$ negative hyperconjugative interactions are enhanced up to a few kilocalories per mole (a) when $\text{LP}_1(\text{O})$ is involved in a weak hydrogen bond of type $\text{LP}_1(\text{O}) \cdots \text{H}-\text{C}$ and (b) when O belongs to an OH group that is involved as a donor in a hydrogen-bond interaction of type $\text{O}-\text{H} \cdots \text{X}$, where X is an electronegative element.

(2) $\text{LP}_1(\text{O}) \rightarrow \sigma^*_{\text{CC}}$ negative hyperconjugative interactions are slightly inhibited when $\text{LP}_1(\text{O})$ is involved in conventional hydrogen bonds of type $\text{LP}_1(\text{O}) \cdots \text{H}-\text{X}$, where X is an electronegative element. This point is worth highlighting: a conventional hydrogen bond shows an opposite effect of that of a weak hydrogen bond.

(3) $\text{LP}_2(\text{O}) \rightarrow \pi^*_{\text{CC}}$ conjugative interactions are easily inhibited by weak hydrogen-bond interactions of type $\text{LP}_2(\text{O}) \cdots \text{H}-\text{C}$.

All of these effects depend upon the dielectric solvent. Apparently, in most cases, the dielectric solvent effects operate both by shielding proximate electrostatics interactions and by a slight inhibition of negative hyperconjugative interactions. In general, the latter are less important than the former.

One of the consequences of point 1 is this: the preferential conformation of an O–H group can be strongly defined by an $\text{O}-\text{H} \cdots \text{X}$ interaction where there is only a modest charge-transfer interaction into the $(\text{O}-\text{H})^*$ antibonding orbital, but it is reinforced by a notably enhanced $\text{LP}_1(\text{O}) \rightarrow \sigma^*_{\text{CC}}$ interaction (up to a few kilocalories per mole).

The lone-pair orientation effect on $^1J_{\text{CC}}$ coupling appears to be an adequate probe to study the interplay of the 1–3 effects commented upon above. It is also important to stress, as observed in previous papers, that in aromatic compounds $^1J_{\text{CC}}$ SSCCs can be adequately reproduced at the level of theory used in this work, B3LYP/6-311G**//B3LYP-EPRIII. However, it should be mentioned that this assertion could fail when there are two electron-rich atoms bonded directly to the C–C bond containing the coupling nuclei.

Acknowledgment. O.E.T. is grateful to SeCyT, National University of Río Cuarto, for financial support; R.H.C.

gratefully acknowledges financial support from UBACYT (X-222) and CONICET (PIP 5119/05), and C.F.T. acknowledges the financial support from FAPESP (grant 06/03980-2).

References

- (1) Wray, V. *Prog. NMR Spectrosc.* **1979**, *13*, 177.
- (2) Krivdin, L. B.; Kalabin, G. A.; Nesterenko, R. N.; Trofimov, B. A. *Tetrahedron Lett.* **1984**, *25*, 4817.
- (3) Krivdin, L. B.; Shcherbakov, U. V. *J. Org. Chem. USSR (Engl. Transl.)* **1986**, *22*, 300.
- (4) Gil, V. M. S.; von Philipsborn, W. *Magn. Reson. Chem.* **1989**, *27*, 409.
- (5) Krivdin, L. B.; Kalabin, G. A. *Prog. NMR Spectrosc.* **1989**, *21*, 293.
- (6) Krivdin, L. B.; Della, E. W. *Prog. NMR Spectrosc.* **1991**, *23*, 301.
- (7) Kamienska-Trela, K. *Annu. Rep. NMR Spectrosc.* **1995**, *30*, 131.
- (8) Krivdin, L. B.; Zinchenko, S. V. *Curr. Org. Chem.* **1998**, *2*, 173.
- (9) Barone, V.; Peralta, J. E.; Contreras, R. H.; Sosnin, A. V.; Krivdin, L. B. *Magn. Reson. Chem.* **2001**, *39*, 600.
- (10) Provasi, P. F.; Aucar, G. A.; Sauer, S. P. A. *Int. J. Mol. Sci.* **2003**, *4*, 231.
- (11) Krivdin, L. B.; Scherbina, N. A.; Istomina, N. V. *Magn. Reson. Chem.* **2005**, *43*, 435.
- (12) Krivdin, L. B.; Larina, L. I.; Chernyshev, K. A.; Rozentsveig, I. B. *Magn. Reson. Chem.* **2005**, *43*, 937.
- (13) Krivdin, L. B.; Nedolya, N. A. *Tetrahedron Lett.* **2005**, *46*, 7367.
- (14) Krivdin, L. B.; Larina, L. I.; Chernyshev, K. A.; Yu Rulev, A. *Magn. Reson. Chem.* **2006**, *44*, 178.
- (15) Krivdin, L. B.; Larina, L. I.; Chernyshev, K. A.; Keiko, N. A. *Aust. J. Chem.* **2006**, *59*, 211.
- (16) Peralta, J. E.; Contreras, R. H.; Snyder, J. P. *J. Chem. Soc., Chem. Commun.* **2000**, 2025.
- (17) Krivdin, L. B.; Contreras, R. H. *Annu. Rep. NMR Spectrosc.* In press.
- (18) Wu, A.; Gräfenstein, J.; Cremer, D. *J. Phys. Chem. A* **2003**, *107*, 7043.
- (19) Krivdin, L. B.; Zinchenko, S. V.; Kalabin, G. A.; Facelli, J. C.; Tufro, M. F.; Contreras, R. H.; Yu Denisov, A.; Gavriilyuk, O. A.; Mamatyuk, V. I. *J. Chem. Soc., Faraday Trans. II* **1992**, *88*, 2459.
- (20) Afonin, A. V.; Ushakov, I. A.; Zinchenko, S. V.; Tarasova, O. A.; Trofimov, B. A. *Magn. Reson. Chem.* **2000**, *38*, 994.
- (21) de Kowalewski, D. G.; Contreras, R. H.; Díez, E.; Esteban, A. L. *Mol. Phys.* **2004**, *102*, 2607.
- (22) Reed, E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899. Weinhold, F. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: New York, 1998; Vol. 3, pp 1792.
- (23) Reed, A. E.; Schleyer, P. v. R. *J. Am. Chem. Soc.* **1990**, *112*, 1434.
- (24) Wray, V.; Ernst, L.; Lund, T.; Jakobsen, H. J. *J. Magn. Reson.* **1980**, *40*, 55.
- (25) Sandor, P.; Radic, L. *Magn. Reson. Chem.* **1986**, *24*, 607. Krivdin, L. B.; Kalabin, G. A. *J. Org. Chem. USSR (Engl. Transl.)* **1989**, *25*, 618. Krivdin, L. B.; Shcherbakov, V. V.; Aliev, I. A.; Kalabin, G. A. *J. Org. Chem. USSR (Engl. Transl.)* **1987**, *23*, 514. Trofimov, B. A.; Krivdin, L. B.; Shcherbakov, V. V.; Aliev, I. A. *Bull. Acad. Sci. USSR, Div. Chem. Sci. (Engl. Transl.)* **1989**, *38*, 54. Krivdin, L. B.; Kalabin, G. A.; Mirskov, R. G.; Solov'eva, S. P. *Bull. Acad. Sci. USSR, Div. Chem. Sci. (Engl. Transl.)* **1982**, *31*, 1799. Kamienska-Trela, K.; Dąbrowski, A.; Januszeewski, H. *Spectrochim. Acta, Part A* **1993**, *49*, 1613. Kamienska-Trela, K.; Dąbrowski, A.; Januszeewski, H. *J. Mol. Struct.* **1993**, *293*, 167.
- (26) Contreras, R. H.; Esteban, A. L.; Díez, E.; Della, E. W.; Lochert, I. J.; dos Santos, F. P.; Tormena, C. F. *J. Phys. Chem. A* **2006**, *110*, 4266.
- (27) Oddershede, J. In *Advances in Quantum Chemistry*; Löwdin, P.-O., Ed.; Academic Press: New York, 1978; Vol. 11, pp 275. Diz, A. C.; Giribet, C. G.; Ruiz de Azúa, M. C.; Contreras, R. H. *Int. J. Quantum Chem.* **1990**, *37*, 663. Contreras, R. H.; Ruiz de Azúa, M. C.; Giribet, C. G.; Aucar, G. A.; Lobayan de Bonczok, R. *THEOCHEM* **1993**, *284*, 249. Giribet, C. G.; Ruiz de Azúa, M. C.; Contreras, R. H.; Lobayan de Bonczok, R.; Aucar, G. A.; Gomez, S. *THEOCHEM* **1993**, *300*, 467.
- (28) Dewar, M. J. S.; Dougherty, R. C. *The PMO Theory of Organic Chemistry*; Plenum Press: New York, 1975.
- (29) Taurian, O. E.; de Kowalewski, D. G.; Pérez, J. E.; Contreras, R. H. *J. Mol. Struct.* **2005**, *754*, 1.
- (30) de Kowalewski, D. G.; Díez, E.; Esteban, A. L.; Barone, V.; Peralta, J. E.; Contreras, R. H. *Magn. Reson. Chem.* **2004**, *42*, 938.
- (31) Eliel, E. L.; Giza, G. A. *J. Org. Chem.* **1968**, *33*, 3754. Lemieux, R. U.; Pavia, A.; Marti, J. C.; Watanabe, K. A. *Can. J. Chem.* **1969**, *47*, 4427.
- (32) Taurian, O. E.; Contreras, R. H.; de Kowalewski, D. G. *J. Argent. Chem. Soc.* In press.
- (33) Desiraju, G. R.; Steiner, T. *The Weak Hydrogen Bond in Structural Chemistry and Biology*; Oxford University Press: New York, 1999; pp 202.
- (34) Lithoxidou, T.; Bakalbassis, E. G. *J. Phys. Chem. A* **2005**, *109*, 366. Bakalbassis, E. G.; Lithoxidou, A. T.; Vafiadis, A. P. *J. Phys. Chem. A* **2006**, *110*, 11151. Han, J.; Lee, H.; Tao, F.-M. *J. Phys. Chem. A* **2005**, *109*, 5186.
- (35) Contreras, R. H.; Peralta, J. E. *Prog. NMR Spectrosc.* **2000**, *37*, 321.
- (36) Bax, A.; Freeman, R.; Frenkiel, T. A. *J. Am. Chem. Soc.* **1981**, *110*, 2102.
- (37) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.
- (38) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098. Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (39) Barone, V. *J. Chem. Phys.* **1994**, *101*, 6834.
- (40) Peralta, J. E.; Scuseria, G. E.; Cheeseman, J. R.; Frisch, M. J. *J. Chem. Phys. Lett.* **2003**, *375*, 452.
- (41) Cancs, M. T.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032. Cossi, M.; Barone, V.; Mennucci, B.; Tomasi, J. *J. Chem. Phys. Lett.* **1998**, *286*, 253. Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *106*, 5151.

- (42) 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 B 05; Gaussian, Inc.: Pittsburgh, PA, 2003.
- (43) Glendening, E. D.; Reed, A. E.; Carpenter, J. E.; Weinhold, F. *NBO*, version 3.1 (included in the Gaussian 03 package of programs).
CT7000396

Extension of the Core-Valence-Rydberg B3LYP Functional to Core-Excited-State Calculations of Third-Row Atoms

Ayako Nakata, Yutaka Imamura, and Hiromi Nakai*

*Department of Chemistry, School of Science and Engineering, Waseda University,
Tokyo 169-8555, Japan*

Received December 19, 2006

Abstract: A modified core-valence-Rydberg Becke's three-parameter exchange (B3) + Lee–Yang–Parr (LYP) correlation (CVR-B3LYP) functional is proposed in order to calculate core-excitation energies of third-row atoms with reasonable accuracy. The assessment of conventional exchange-correlation functionals shows that the appropriate portions of Hartree–Fock (HF) exchange for core-excited-state calculations depend on shells: 70% and 50% for K-shell and L-shell excitations, respectively. Therefore, the modified CVR-B3LYP functional is designed to use the appropriate portions of HF exchange, 70%, 50%, and 20%, for K-shell, L-shell, and valence regions separately. Time-dependent density functional theory calculations with the modified CVR-B3LYP functional yield both K-shell and L-shell excitation energies with reasonable accuracy. The modified CVR-B3LYP also provides valence-excitation energies and standard enthalpies of formation accurately. Thus, the modified CVR-B3LYP describes all of the K-shell, L-shell, and valence electrons appropriately.

1. Introduction

Time-dependent density functional theory (TDDFT)^{1–8} has been one of the most widely used tools for excited-state calculations. TDDFT provides quantitative results for low-lying valence-excited states with low computational costs because electron correlations are included through exchange-correlation functionals. However, the disadvantages of TD-DFT have been reported: TDDFT calculations with conventional exchange-correlation functionals tend to largely underestimate the core- and Rydberg-excitation energies.^{9–14} The underestimation is considered to occur because of the inappropriate behavior of exchange-correlation functionals in core and Rydberg regions.^{11,15} Several methods for improving TDDFT accuracy for core and Rydberg excitations have been advocated.^{9,12,16–20}

For core excitations, core-valence Becke's three-parameter exchange (B3)²¹ + Lee–Yang–Parr (LYP)²² correlation (CV-B3LYP)⁹ hybrid functional has been proposed. CV-

B3LYP yields both core- and valence-excitation energies with high accuracy by using appropriate portions of Hartree–Fock (HF) exchange for core and valence regions separately. Density functional theory (DFT) with the transition state approach and the delta-Kohn–Sham (ΔE_{KS}) method,^{23,24} which are based on DFT but not on TDDFT, also reproduce core excitation energies with high accuracy. For Rydberg excitations, the van Leeuwen–Baerends' 94 (LB94) functional,¹⁶ the statistical average of different orbital model potential (SAOP) functional,¹⁷ the asymptotically corrected (AC) Kohn–Sham (KS) equation of Tozer and Handy,¹⁸ and the long-range correction (LC) scheme for the exchange functional¹⁹ have been proposed. The modified LB94 and the Becke's 1988 exchange (B88)²⁵ + LYP correlation (BLYP) pure functional, BmLBLYP,¹² and the core-valence-Rydberg B3LYP (CVR-B3LYP)²⁰ hybrid functional have been reported as the methods for improving the descriptions of both core and Rydberg excitations. By taking advantage of the appropriate portions of HF exchange not only for core and occupied-valence orbitals but also for the unoccupied-valence and Rydberg orbitals separately, TDDFT calculations

* Corresponding author phone: +81-3-5286-3452; fax: +81-3-3205-2504; e-mail: nakai@waseda.jp.

Table 1. 1s Core-Excitation Energies of SiH₄, PH₃, H₂S, SO₂, HCl, and Cl₂ Molecules by TDHF and TDDFT with the BLYP, B3LYP, and BHLLYP Functionals with cc-pCVTZ Plus Rydberg Basis Functions (in eV)^a

molecule	assignment	BLYP			B3LYP			BHLLYP			TDHF			
		NR ^b	R ^c	Δ _{R-NR} ^d	NR ^b	R ^c	Δ _{R-NR} ^d	NR ^b	R ^c	Δ _{R-NR} ^d	NR ^b	R ^c	Δ _{R-NR} ^d	
SiH₄	Si 1s → σ*	1781.0 (-61.5)	1784.7 (-57.8)	3.7	1797.7 (-44.8)	1801.1 (-41.4)	3.4	1823.7 (-18.8)	1827.6 (-14.9)	3.9	1864.9 (+22.4)	1869.1 (+26.6)	4.2	1842.5 ^e
PH₃	P 1s → σ*(e)	2077.5 (-68.4)	2082.4 (-63.4)	4.9	2095.1 (-50.8)	2100.1 (-45.7)	5.1	2122.8 (-23.0)	2128.1 (-17.8)	5.2	2166.9 (+21.1)	2172.5 (+26.6)	5.6	2145.8 ^f
H₂S	S 1s → 3b ₂ (σ*)	2397.2 (-75.9)	2403.7 (-69.4)	6.5	2415.9 (-57.2)	2422.6 (-50.5)	6.7	2445.6 (-27.5)	2452.5 (-20.6)	6.9	2492.9 (+19.8)	2500.1 (+27.0)	7.3	2473.1 ^g
SO₂	S 1s → 3b ₁ (σ*)	2397.7 (-76.1)	2404.3 (-69.5)	6.5	2416.3 (-57.5)	2423.0 (-50.8)	6.7	2445.8 (-28.0)	2452.7 (-21.1)	6.9	2492.3 (+18.5)	2499.6 (+25.8)	7.3	2473.8 ^g
HCl	Cl 1s → 3pσ*	2739.9 (-84.0)	2748.4 (-75.5)	8.5	2760.4 (-63.5)	2769.0 (-54.9)	8.7	2792.4 (-31.5)	2801.3 (-22.6)	8.9	2843.3 (+19.4)	2852.6 (+28.7)	9.4	2823.9 ^h
Cl₂	Cl 1s → 3pσ _u *	2738.0 (-83.3)	2746.5 (-74.8)	8.5	2757.9 (-63.4)	2766.5 (-54.8)	8.7	2789.4 (-31.9)	2798.3 (-23.0)	8.9	2839.7 (+18.4)	2849.1 (+27.8)	9.4	2821.3 ^h
MEⁱ		-64.2	-58.6		-48.2	-42.6		-23.0	-17.1		17.1	23.2		

^a Differences from experimental data are shown in parentheses. ^b Excitation energies by relativistic calculations. ^c Excitation energies by nonrelativistic calculations. ^d Differences between NR and R. ^e Reference 36. ^f Reference 37. ^g Reference 38. ^h Reference 39. ⁱ Mean errors from experimental data.

with CVR-B3LYP have succeeded in describing Rydberg excitations with reasonable accuracy.

In the previous CV- and CVR-B3LYP studies, the calculations have been performed on the small molecules containing second-row atoms. In this study, we extend the CVR-B3LYP functional to core-excited-state calculations of third-row atoms. The assessment of time-dependent HF (TDHF) and TDDFT calculations with conventional exchange-correlation functionals on the molecules containing third-row atoms are shown in the next section. Based on the assessment, the CVR-B3LYP functional is modified in the third section in order to improve the descriptions of core excitations from third-row atoms. The last section gives the conclusions of the present study.

2. Assessment of Conventional Exchange-Correlation Functionals for Core-Excited-State Calculations on Third-Row Atoms

In this section, the appropriate portions of HF exchange for describing K-shell and L-shell core excitations have been investigated by performing TDDFT calculations with conventional exchange-correlation functionals: BLYP, B3LYP, and Becke's half-and-half exchange + LYP correlation (BHLLYP).²⁶ TDHF calculations were carried out for comparison. The correlation consistent polarized core-valence triple- ζ (cc-pCVTZ) basis set²⁷⁻³⁰ was used. Single (s, p) Rydberg basis functions were added for describing (3s, 3p) orbitals of second-row atoms and (4s, 4p) orbitals of third-row atoms.³¹⁻³³ All molecular structures were optimized at the B3LYP/cc-pVTZ²⁷ level. The scalar relativistic effect is included by using the relativistic scheme by eliminating small-components (RESC) method.^{34,35} Spin-orbit interactions are not included in the present calculations.

The 1s and 2p core-excitation energies of SiH₄, PH₃, H₂S, SO₂, HCl, and Cl₂ molecules calculated with TD-BLYP, TD-B3LYP, TD-BHLLYP, and TDHF are shown in Tables 1 and 2. Si, P, S, and Cl in boldface correspond to the atoms whose 1s or 2p electrons are excited. The results of relativistic (R) and nonrelativistic (NR) calculations and their differences are shown in the tables. The differences from the experimental values are shown in parentheses. For the 2p excitation energies of SiH₄, PH₃, and Cl₂, the weighted averaged values between P_{1/2} and P_{3/2} states, which are obtained by the procedure mentioned in ref 44, are adopted as the experimental data. As for the 1s core-excitation energies in Table 1, the relativistic effect becomes larger as the atomic number increases: The differences between the results with and without relativistic corrections are 3.4–4.2, 4.9–5.6, 6.5–7.3, and 8.5–9.4 eV for Si, P, S, and Cl, respectively. It is also shown that the relativistic correction becomes larger in the order BLYP < B3LYP < BHLLYP < TDHF, which is consistent with the portions of HF exchange in the functionals. The mean errors (MEs) of TD-BLYP, TD-B3LYP, TD-BHLLYP, and TDHF with relativistic corrections for third-row atoms are -58.6, -42.6, -17.1, and 23.2 eV, respectively, which are significantly larger than -17.9, -12.0, -2.8, and 11.1 eV for the second-row atoms obtained in the previous study.²⁰ 1s core-excitation

Table 2. 2p Core-Excitation Energies of SiH₄, PH₃, H₂S, SO₂, HCl, and Cl₂ Molecules by TDHF and TDDFT with the BLYP, B3LYP, and BHHLYP Functionals with cc-pCVTZ Plus Rydberg Basis Functions (in eV)^a

molecule	assignment	BLYP			B3LYP			BHHLYP			TDHF			exptl
		NR ^b	R ^c	Δ _{R-NR} ^d	NR ^b	R ^c	Δ _{R-NR} ^d	NR ^b	R ^c	Δ _{R-NR} ^d	NR ^b	R ^c	Δ _{R-NR} ^d	
SiH ₄	Si 2p → σ*	94.4	94.4	0.0	97.8	97.8	0.0	102.4	102.4	0.0	109.4	109.4	0.0	102.8 ^{e,i}
		(-8.4)	(-8.4)		(-5.0)	(-5.0)		(-0.4)	(-0.4)		(+6.6)	(+6.6)		
PH ₃	P 2p → σ*(a ₁)	122.4	122.4	0.0	126.5	126.5	0.0	132.2	131.4	-0.8	139.0	140.9	1.9	132.3 ^{e,i}
		(-9.9)	(-9.9)		(-5.8)	(-5.9)		(-0.1)	(-0.9)		(+6.7)	(+8.6)		
H ₂ S	S 2p → 3b ₂ (σ*)	154.8	154.8	0.0	158.6	158.5	0.0	164.4	164.4	0.0	172.9	172.9	0.0	164.5 ^e
		(-9.7)	(-9.7)		(-5.9)	(-6.0)		(-0.1)	(-0.1)		(+8.4)	(+8.4)		
SO ₂	S 2p → 3b ₁ (π*)	154.7	154.7	0.0	158.4	158.4	0.0	163.6	163.6	0.0	171.6	171.7	0.0	164.4 ^f
		(-9.6)	(-9.7)		(-6.0)	(-6.0)		(-0.7)	(-0.7)		(+7.3)	(+7.3)		
HCl	Cl 2pπ → 3pσ*	189.5	189.5	0.0	194.0	194.0	0.0	200.4	200.4	0.0	210.1	210.1	0.0	201.0 ^g
		(-11.5)	(-11.5)		(-7.0)	(-7.0)		(-0.6)	(-0.6)		(+9.1)	(+9.1)		
Cl ₂	Cl 2pπ → 3pσ _v *	187.3	187.3	0.0	191.4	191.4	0.0	197.4	197.4	0.0	206.8	206.8	0.0	198.7 ^{h,i}
		(-11.5)	(-11.5)		(-7.3)	(-7.3)		(-1.4)	(-1.3)		(+8.1)	(+8.1)		
ME/		-8.7	-8.7		-5.3	-5.3		-0.5	-0.6		6.6	6.9		

^a Differences from experimental data are shown in parentheses. ^b Excitation energies by relativistic calculations. ^c Excitation energies by nonrelativistic calculations. ^d Differences between NR and R. ^e Reference 40. ^f Reference 41. ^g Reference 42. ^h Reference 43. ⁱ Weighted average value calculated with the method in ref 44. ^j Mean errors from experimental data.

energies are underestimated by pure TDDFT, while those are overestimated by TDHF. Using hybrid functionals such as B3LYP and BHHLYP improves the descriptions of 1s core-excitations in comparison with a pure functional, which shows that HF exchange reduces the underestimation of the pure TDDFT method. The behavior of 1s core-excitation energies of third-row atoms discussed above is analogous to that of second-row atoms in ref 20.

With regards to the 2p core-excitation energies in Table 2, the relativistic corrections are at most 1.9 eV, which are smaller than those for the 1s excitation energies. The MEs of TD-BLYP, TD-B3LYP, TD-BHHLYP, and TDHF with relativistic corrections are -8.5, -5.2, -0.5, and 7.0 eV, respectively. The underestimation by TDDFT, the overestimation by TDHF, and the improvement of the results by using hybrid functionals instead of pure functionals are also observed for 2p core-excitations. BHHLYP, which have the smallest mean absolute error (MAE) of 0.5 eV, gives the excitation energies accurately enough to discuss 2p core-excitations.

In order to investigate the effect of HF exchange more precisely, we performed the additional TDDFT calculations with the following exchange-correlation functionals

$$E_{xc} = a \sum_{ij} (-K_{ij}) + (1 - a)E_x^{\text{B88}}[\rho] + E_c^{\text{LYP}}[\rho] \quad (1)$$

where K_{ij} , E_x^{B88} , and E_c^{LYP} represent HF exchange, B88 exchange, and LYP correlation energies. Suffixes i and j denote the indexes of occupied orbitals. ρ is the total electron density. The portion of HF exchange is changed to 60%, 70%, 80%, 90%, or 100% by setting the coefficient a to 0.6, 0.7, 0.8, 0.9, or 1.0, respectively. We denote the functional in eq 1 with $X\%$ portions of HF exchange "HF+B88+LYP ($X\%$)" in the present study. The scalar relativistic effects were considered by the RESC method.

Tables 3 and 4 show the 1s and 2p core-excitation energies calculated with nine kinds of methods: BLYP, B3LYP, BHHLYP, HF+B88+LYP ($X\%$) ($X = 60, 70, 80, 90,$ and

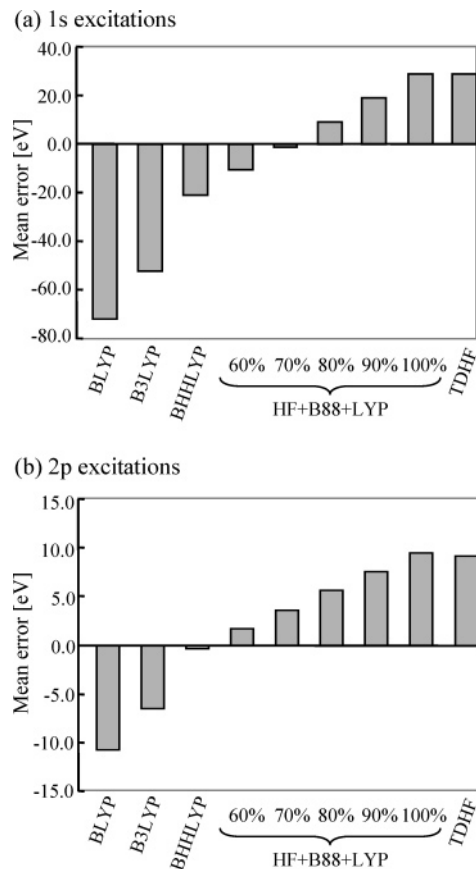


Figure 1. Mean errors of 1s and 2p core-excitation energies by TDHF and TDDFT with the BLYP, B3LYP, BHHLYP, and HF+B88+LYP functionals with cc-pCVTZ plus Rydberg basis functions.

100), and TDHF. In Table 3, the MAEs of BLYP with no HF exchange and HF+B88+LYP (100%) and TDHF only with HF exchange are significantly large: 70.8, 28.8, and 28.4 eV, respectively. The difference of the MAEs between HF+B88+LYP (100%) and TDHF is 0.4 eV, which indicates that the effect of the correlation functional on calculated

Table 3. 1s Core-Excitation Energies of SiH₄, PH₃, H₂S, SO₂, HCl, and Cl₂ Molecules by TDHF and TDDFT with the BLYP, B3LYP, BHHLYP, and HF+B88+LYP Functionals with cc-pCVTZ Plus Rydberg Basis Functions (in eV)^a

molecule	assignment	BLYP 0%	B3LYP 20%	BHHLYP 50%	HF+B88+LYP					TDHF 100%	exptl
					60%	70%	80%	90%	100%		
SiH ₄	Si 1s → σ*	1784.7	1801.1	1827.6	1836.0	1844.4	1852.8	1861.1	1869.3	1869.1	1842.5 ^b
		(−57.8)	(−41.4)	(−14.9)	(−6.5)	(+1.9)	(+10.3)	(+18.6)	(+26.8)	(+26.6)	
PH ₃	P 1s → σ*(e)	2082.4	2100.1	2128.1	2137.1	2146.0	2155.0	2163.8	2172.7	2172.5	2145.8 ^c
		(−63.4)	(−45.7)	(−17.8)	(−8.8)	(+0.2)	(+9.1)	(+18.0)	(+26.8)	(+26.6)	
H ₂ S	S 1s → 3b ₂ (σ*)	2403.7	2422.6	2452.5	2462.2	2471.8	2481.3	2490.9	2500.3	2500.1	2473.1 ^d
		(−69.4)	(−50.5)	(−20.6)	(−10.9)	(−1.3)	(+8.2)	(+17.8)	(+27.2)	(+27.0)	
SO ₂	S 1s → 3b ₁ (π*)	2404.3	2423.0	2452.7	2462.2	2471.7	2481.1	2490.4	2499.7	2499.6	2473.8 ^d
		(−69.5)	(−50.8)	(−21.1)	(−11.6)	(−2.1)	(+7.3)	(+16.6)	(+25.9)	(+25.8)	
HCl	Cl 1s → 3pσ*	2748.4	2769.0	2801.3	2811.7	2822.1	2832.4	2842.6	2852.8	2852.6	2823.9 ^e
		(−75.5)	(−54.9)	(−22.6)	(−12.2)	(−1.8)	(+8.5)	(+18.7)	(+28.9)	(+28.7)	
Cl ₂	Cl 1s → 3pσ _u *	2746.5	2766.5	2798.3	2808.6	2818.8	2829.0	2839.1	2849.2	2849.1	2821.3 ^e
		(−74.8)	(−54.8)	(−23.0)	(−12.7)	(−2.5)	(+7.7)	(+17.8)	(+27.9)	(+27.8)	
MAE ^f		70.8	51.3	20.5	10.6	1.2	9.2	19.0	28.8	28.4	

^a The portion of HF exchange in the exchange-correlation functional is shown for each method. Differences from experimental data are shown in parentheses. ^b Reference 36. ^c Reference 37. ^d Reference 38. ^e Reference 39. ^f Mean absolute errors from experimental data.

Table 4. 2p Core-Excitation Energies of SiH₄, PH₃, H₂S, SO₂, HCl, and Cl₂ Molecules by TDHF and TDDFT with the BLYP, B3LYP, BHHLYP, and HF+B88+LYP Functionals with cc-pCVTZ Plus Rydberg Basis Functions (in eV)^a

molecule	assignment	BLYP 0%	B3LYP 20%	BHHLYP 50%	HF+B88+LYP					TDHF 100%	exptl
					60%	70%	80%	90%	100%		
SiH ₄	Si 2p → σ*	94.4	97.8	102.4	103.9	105.4	106.8	108.3	109.6	109.4	102.8 ^{b,f}
		(−8.4)	(−5.0)	(−0.4)	(+1.1)	(+2.6)	(+4.0)	(+5.5)	(+6.8)	(+6.6)	
PH ₃	P 2p → σ*	122.4	126.5	131.4	134.1	134.6	137.8	139.5	141.3	140.9	132.3 ^{b,f}
		(−9.9)	(−5.9)	(−0.9)	(+1.8)	(+2.3)	(+5.5)	(+7.2)	(+8.9)	(+8.6)	
H ₂ S	S 2p → σ*	154.8	158.5	164.4	166.2	167.9	169.7	171.4	173.2	172.9	164.5 ^b
		(−9.7)	(−6.0)	(−0.1)	(+1.7)	(+3.4)	(+5.2)	(+6.9)	(+8.7)	(+8.4)	
SO ₂	S 2p → 3b ₁ (π*)	154.7	158.4	163.6	165.3	167.0	168.7	170.3	171.9	171.7	164.4 ^c
		(−9.7)	(−6.0)	(−0.7)	(+1.0)	(+2.6)	(+4.3)	(+5.9)	(+7.5)	(+7.3)	
HCl	Cl 2pπ → 3pσ*	189.5	194.0	200.4	202.4	204.5	206.5	208.4	210.4	210.1	201.0 ^d
		(−11.5)	(−7.0)	(−0.6)	(+1.4)	(+3.5)	(+5.5)	(+7.4)	(+9.4)	(+9.1)	
Cl ₂	Cl 2pπ → 3pσ _u *	187.3	191.4	197.4	199.4	201.3	203.2	205.1	207.0	206.8	198.7 ^{e,f}
		(−11.5)	(−7.3)	(−1.3)	(+0.6)	(+2.6)	(+4.5)	(+6.4)	(+8.3)	(+8.1)	
MAE ^g		10.8	6.6	0.7	1.6	3.4	5.5	7.5	9.4	9.0	

^a The portion of HF exchange in the exchange-correlation functional is shown for each method. Differences from experimental data are shown in parentheses. ^b Reference 40. ^c Reference 41. ^d Reference 42. ^e Reference 43. ^f Weighted average value calculated with the method in ref 44. ^g Mean absolute errors from experimental data.

1s core-excitation energies is small. HF+B88+LYP (70%) with a MAE of 1.2 eV shows the best performance among the nine methods. In Table 4, BHHLYP gives the smallest MAE, 0.6 eV. HF+B88+LYP (60%) with a MAE of 1.7 eV shows high performance. The effect of the correlation

functional is small on 2p core excitations: The difference of the MAEs between HF+B88+LYP (100%) and TDHF is 0.4 eV.

The MEs of 1s and 2p excitation energies calculated with the nine methods are illustrated in Figure 1. Figure 1

clearly demonstrates that both calculated 1s and 2p core-excitation energies increase as the portion of HF exchange in the employed functional increases. The appropriate portion of HF exchange for 1s core excitations is different from that for 2p core excitations: about 70% and 50% for 1s and 2p excitations, respectively. This fact is thought to be due to the large self-interaction errors of 1s electrons because K-shell electrons in a third-row atom are attracted to the nucleus more strongly than L-shell electrons in a third-row atom and K-shell electrons in a second-row atom are.^{11,15}

3. Extension of CVR-B3LYP to Core-Excited-State Calculations of Third-Row Atoms

3.1. Modified CVR-B3LYP Equations for Core Excitations from Third-Row Atoms. As mentioned in section 2, 70% and 50% portions of HF exchange are appropriate for describing K-shell and L-shell electrons, while B3LYP with 20% portion of HF exchange is well-known to perform well for valence properties. Therefore, CVR-B3LYP is modified to use appropriate portions of HF exchange for K-shell, L-shell, and valence regions separately. In the previous CVR-B3LYP, the occupied orbitals are distinguished into core (C) and occupied-valence (OV) orbitals. In the present modified CVR-B3LYP, the occupied orbitals are distinguished into three groups, namely, K-shell (C1), L-shell (C2), and occupied-valence (OV) orbitals. Thus, the electronic energy is decomposed into C1–C1, C1–C2, C1–OV, C2–C2, C2–OV, and OV–OV interactions

$$\begin{aligned}
 E = & 2 \sum_k^{C1} H_k + 2 \sum_m^{C2} H_m + 2 \sum_p^{OV} H_p + \sum_{kl}^{C1} 2J_{kl} + \\
 & \sum_k^{C1} \sum_m^{C2} 2J_{km} + \sum_k^{C1} \sum_p^{OV} 2J_{kp} + \sum_m^{C2} \sum_k^{C1} 2J_{mk} + \sum_{mn}^{C2} 2J_{mn} + \\
 & \sum_m^{C2} \sum_p^{OV} 2J_{mp} + \sum_p^{OV} \sum_k^{C1} 2J_{pk} + \sum_p^{OV} \sum_m^{C2} 2J_{pm} + \sum_{pq}^{OV} 2J_{pq} + \\
 & a_{C1C1} \sum_{kl}^{C1} (-K_{kl}) + a_{C1C2} \sum_k^{C1} \sum_m^{C2} (-K_{km}) + \\
 & a_{C1OV} \sum_k^{C1} \sum_p^{OV} (-K_{kp}) + a_{C1C2} \sum_m^{C2} \sum_k^{C1} (-K_{mk}) + \\
 & a_{C2C2} \sum_{mn}^{C2} (-K_{mn}) + a_{C2OV} \sum_m^{C2} \sum_p^{OV} (-K_{mp}) + \\
 & a_{C1OV} \sum_p^{OV} \sum_k^{C1} (-K_{pk}) + a_{C2OV} \sum_p^{OV} \sum_m^{C2} (-K_{pm}) + \\
 & a_{OV OV} \sum_{pq}^{OV} (-K_{pq}) + b_{C1C1} E_{xc}[\rho_{C1}] + b_{C2C2} E_{xc}[\rho_{C2}] + \\
 & b_{OV OV} E_{xc}[\rho_{OV}] + b_{C1C2} (E_{xc}[\rho_{C1+C2}] - E_{xc}[\rho_{C1}] - \\
 & E_{xc}[\rho_{C2}]) + b_{C1OV} (E_{xc}[\rho_{C1+OV}] - E_{xc}[\rho_{C1}] - E_{xc}[\rho_{OV}]) + \\
 & b_{C2OV} (E_{xc}[\rho_{C2+OV}] - E_{xc}[\rho_{C2}] - E_{xc}[\rho_{OV}]) \quad (2)
 \end{aligned}$$

Table 5. Coefficients of Exchange-Correlation Functionals in the Modified CVR-B3LYP Functional

	C1C1	C1C2	C1OV	C2C2	C2OV	OVOV
<i>a</i> (HF exchange)	0.7	0.6	0.45	0.5	0.35	0.2
<i>b</i> (Slater exchange)	0	0	0.04	0	0.04	0.08
(B88 exchange)	0.3	0.4	0.51	0.5	0.61	0.72
(VWN5 correlation)	0	0	0.095	0	0.095	0.19
(LYP correlation)	1	1	0.905	1	0.905	0.81

where *H* and *J* are 1-electron and Coulomb integrals, and *a* and *b* are the coefficients of HF exchange and DFT exchange-correlation functionals. The “C1”, “C2”, and “OV” on the Σ mean that the summation runs over the K-shell, L-shell, and occupied-valence orbitals, respectively; therefore, suffixes (*k*, *l*), (*m*, *n*), and (*p*, *q*) correspond to K-shell, L-shell, and occupied-valence orbitals. The definitions of the electron densities are as follows

$$\begin{aligned}
 \rho_{C1} &= \sum_k^{C1} |\varphi_k|^2, \rho_{C2} = \sum_m^{C2} |\varphi_m|^2, \rho_{OV} = \sum_p^{OV} |\varphi_p|^2 \\
 \rho_{C1+C2} &= \sum_i^{\neq OV} |\varphi_i|^2, \rho_{C1+OV} = \sum_i^{\neq C2} |\varphi_i|^2, \\
 \rho_{C2+OV} &= \sum_i^{\neq C1} |\varphi_i|^2 \quad (3)
 \end{aligned}$$

where φ is the KS orbital, and the “ $\neq C1$ ”, “ $\neq C2$ ”, and “ $\neq OV$ ” on the Σ mean that the summation runs over all occupied orbitals without the K-shell, L-shell, and occupied-valence orbitals, respectively. The C1–C2 interaction is represented as the subtraction of $E_{xc}[\rho_{C1}]$ and $E_{xc}[\rho_{C2}]$ from $E_{xc}[\rho_{C1+C2}]$, and the same applies to C1–OV and C2–OV interactions. In eq 2, the three- and higher-body interactions in DFT exchange-correlation energies are neglected. However, our preliminary calculations have shown that the energy differences due to the truncation are small enough to be negligible. For more details, see ref 45. The exchange-correlation functional in CVR-B3LYP consists of Slater exchange,⁴⁶ B88 exchange,²⁵ Vosko–Wilk–Nusair (VWN5) correlation,⁴⁷ and LYP correlation²² functionals. The coefficients a_Y and b_Y ($Y = C1C1, C1C2, C1OV, C2C2, C2OV,$ and $OVOV$) used in the present calculations are listed in Table 5. The coefficients of C1C1, C2C2, and OVOV are set to those of HF+B88+LYP (70%), BHHLYP, and B3LYP. The coefficients of C1C2, C1OV, and C2OV are set to the mean values of {C1C1 and C2C2}, {C1C1 and OVOV}, and {C2C2 and OVOV}, respectively. The sum of the coefficients in each group *Y* becomes one.

Using the variational principle to eq 2 leads to three kinds of Fock operators

$$\begin{aligned}
 F_{C1} = & h + 2J - (a_{C1C1} K_{C1} + a_{C1C2} K_{C2} + a_{C1OV} K_{OV}) + \\
 & (b_{C1C1} - b_{C1C2} - b_{C1OV}) V_{xc}[\rho_{C1}] + b_{C1C2} V_{xc}[\rho_{C1+C2}] + \\
 & b_{C1OV} V_{xc}[\rho_{C1+OV}] \quad (4)
 \end{aligned}$$

$$F_{C2} = h + 2J - (a_{C1C2}K_{C1} + a_{C2C2}K_{C2} + a_{C2OV}K_{OV}) + \\ (b_{C2C2} - b_{C1C2} - b_{C2OV})V_{xc}[\rho_{C2}] + b_{C1C2}V_{xc}[\rho_{C1+C2}] + \\ b_{C2OV}V_{xc}[\rho_{C2+OV}] \quad (5)$$

$$F_{OV} = h + 2J - (a_{C1OV}K_{C1} + a_{C2OV}K_{C2} + a_{OV OV}K_{OV}) + \\ (b_{OV OV} - b_{C1OV} - b_{C2OV})V_{xc}[\rho_{OV}] + b_{C1OV}V_{xc}[\rho_{C1+OV}] + \\ b_{C2OV}V_{xc}[\rho_{C2+OV}] \quad (6)$$

where h is 1-electron operator, and J and K in and after eq 4 are Coulomb and HF-exchange operators. HF-exchange operators and the first derivatives of E_{xc} are as follows:

$$K_{C1} = \sum_k^{C1} K_k, \quad K_{C2} = \sum_m^{C2} K_m, \quad K_{OV} = \sum_p^{OV} K_p, \\ V_{xc}[\rho_{C1}] = \frac{\delta E_{xc}[\rho_{C1}]}{\delta \rho_{C1}}, \quad V_{xc}[\rho_{C2}] = \frac{\delta E_{xc}[\rho_{C2}]}{\delta \rho_{C2}}, \\ V_{xc}[\rho_{OV}] = \frac{\delta E_{xc}[\rho_{OV}]}{\delta \rho_{OV}}, \quad V_{xc}[\rho_{C1+C2}] = \frac{\delta E_{xc}[\rho_{C1+C2}]}{\delta \rho_{C1+C2}}, \\ V_{xc}[\rho_{C1+OV}] = \frac{\delta E_{xc}[\rho_{C1+OV}]}{\delta \rho_{C1+OV}}, \\ V_{xc}[\rho_{C2+OV}] = \frac{\delta E_{xc}[\rho_{C2+OV}]}{\delta \rho_{C2+OV}} \quad (7)$$

In order to guarantee the invariance under the unitary transformation, the coupling-operator technique of Rootaan^{48–50} is adopted. Introducing the operators R

$$R_{C1} = -\sum_m^{C2} \{ |\varphi_m\rangle\langle\varphi_m| \Theta_{C1C2} \} + \Theta_{C1C2} |\varphi_m\rangle\langle\varphi_m| - \\ \sum_p^{OV} \{ |\varphi_p\rangle\langle\varphi_p| \Theta_{C1OV} \} + \Theta_{C1OV} |\varphi_p\rangle\langle\varphi_p| \quad (8)$$

$$R_{C2} = -\sum_k^{C1} \{ |\varphi_k\rangle\langle\varphi_k| \Theta_{C2C1} \} + \Theta_{C2C1} |\varphi_k\rangle\langle\varphi_k| - \\ \sum_p^{OV} \{ |\varphi_p\rangle\langle\varphi_p| \Theta_{C2OV} \} + \Theta_{C2OV} |\varphi_p\rangle\langle\varphi_p| \quad (9)$$

$$R_{OV} = -\sum_k^{C1} \{ |\varphi_k\rangle\langle\varphi_k| \Theta_{OV C1} \} + \Theta_{OV C1} |\varphi_k\rangle\langle\varphi_k| - \\ \sum_m^{C2} \{ |\varphi_m\rangle\langle\varphi_m| \Theta_{OV C2} \} + \Theta_{OV C2} |\varphi_m\rangle\langle\varphi_m| \quad (10)$$

we obtain the coupling operators as

$$F_{C1}' = F_{C1} + R_{C1} \quad (11)$$

$$F_{C2}' = F_{C2} + R_{C2} \quad (12)$$

$$F_{OV}' = F_{OV} + R_{OV} \quad (13)$$

where Θ s are

$$\Theta_{C1C2} = (1 - \lambda)F_{C1} + \lambda F_{C2} \quad (14)$$

$$\Theta_{C2C1} = -\lambda F_{C1} + (1 + \lambda)F_{C2} \quad (15)$$

$$\Theta_{C1OV} = (1 - \mu)F_{C1} + \mu F_{OV} \quad (16)$$

$$\Theta_{OV C1} = -\mu F_{C1} + (1 + \mu)F_{OV} \quad (17)$$

$$\Theta_{C2OV} = (1 - \sigma)F_{C2} + \sigma F_{OV} \quad (18)$$

$$\Theta_{OV C2} = -\sigma F_{C2} + (1 + \sigma)F_{OV} \quad (19)$$

and λ , μ , and σ are arbitrary nonzero numbers and are set to 0.1 in the present study. Thus, the Fock operator for occupied orbitals is rewritten as follows:

$$F = \sum_k^{C1} F_{C1}' |\varphi_k\rangle\langle\varphi_k| + \sum_m^{C2} F_{C2}' |\varphi_m\rangle\langle\varphi_m| + \\ \sum_p^{OV} F_{OV}' |\varphi_p\rangle\langle\varphi_p| \quad (20)$$

The virtual orbitals are treated in a similar way as the previous CVR-B3LYP,²⁰ in which the Rydberg orbitals are distinguished by using second moments of the orbitals. F_{OV} and the Fock operator form in the HF method were adopted as the Fock operator forms of unoccupied-valence and Rydberg orbitals, respectively. In the TDDFT calculations, we adopted an approximation similar to that for the previous study,²⁰ in which we used the **A** and **B** matrix forms of B3LYP,^{1–8} while using the orbital energies and coefficients of CVR-B3LYP.

3.2. Assessment of Modified CVR-B3LYP Functional.

The descriptions of K-shell, L-shell, and valence electrons by the modified CVR-B3LYP functional are assessed in this section by calculating core- and valence-excitation energies and standard enthalpies of formations. In the CVR-B3LYP calculations, the portions of HF exchange for K-shell, L-shell, and occupied-valence orbitals were determined to be 70%, 50%, and 20% by using the coefficients given in Table 5. The scalar relativistic effects were included by using the RESC method. The basis sets and geometries of molecules used in CVR-B3LYP calculations are the same as those used in section 2.

Table 6 shows the core excitation energies and oscillator strengths of the HCl molecule calculated by TDDFT with B3LYP, BHHLYP, the modified CVR-B3LYP, and TDHF. The errors from experimental values are shown in parentheses. The $1s \rightarrow 4p\pi$ and $1s \rightarrow 4p\sigma$ excitations are assigned to the same peak experimentally. As for the $1s$ core-excitation energies, the modified CVR-B3LYP shows a significantly higher performance than conventional functionals: the errors of the modified CVR-B3LYP are about 1 eV, while those of B3LYP, BHHLYP, and TDHF are about 55, 22, and 30 eV, respectively. TD-B3LYP and TD-BHHLYP fail to reproduce the order of $1s$ excitations because $4s$ and $4p\sigma$ excitations are calculated to be strongly mixed with each other. Only the modified CVR-B3LYP represents the correct order of the four $1s$ -excited states. With regards to $2p$ excitations, the accuracy of the modified CVR-B3LYP is comparable to that of BHHLYP. TDHF overestimates $2p$ excitation energies by about 10 eV, while B3LYP underestimates those by about 10 eV. The MAE of the modified CVR-B3LYP, 0.8 eV, is significantly smaller than those of

Table 6. 1s and 2p Core-Excitation Energies and Oscillator Strengths of HCl by TDHF and TDDFT with B3LYP, BHHLYP, and the Modified CVR-B3LYP Functionals with cc-pCVTZ Plus Rydberg Basis Functions (in eV)^a

	excitation energy					oscillator strength			
	B3LYP	BHHLYP	TDHF	CVR-B3LYP	exptl	B3LYP	BHHLYP	TDHF	CVR-B3LYP
1s Excitation									
Cl 1s → 3pσ*	2769.0 (-54.9)	2801.3 (-22.6)	2852.6 (+28.7)	2824.8 (+0.9)	2823.9 ^b	0.0023	0.0045	0.0085	0.0015
Cl 1s → 4s	2771.3 (-55.7)	2805.2 (-21.8)	2859.5 (+32.5)	2827.8 (+0.8)	2827.0 ^b	0.0010	0.0004	0.0000	0.0003
Cl 1s → 4pπ	2771.4 (-56.4)	2805.2 (-22.6)	2859.3 (+31.5)	2827.9 (+0.1)	2827.8 ^b	0.0000	0.0003	0.0023	0.0001
Cl 1s → 4pσ	2771.0 (-56.8)	2804.4 (-23.4)	2858.2 (+30.4)	2829.4 (+1.6)	2827.8 ^b	0.0010	0.0007	0.0004	0.0025
2p Excitation									
Cl 2pπ → 3pσ*	194.0 (-7.0)	200.4 (-0.6)	210.1 (+9.1)	202.0 (+1.0)	201.0 ^c	0.0080	0.0112	0.0162	0.0071
Cl 2pπ → 4s	196.3 (-7.6)	204.0 (+0.1)	215.9 (+12.0)	204.9 (+1.0)	203.9 ^c	0.0014	0.0015	0.0025	0.0009
Cl 2pπ → 4pπ	196.4 (-8.2)	204.0 (-0.6)	215.9 (+11.3)	205.0 (+0.4)	204.6 ^c	0.0000	0.0000	0.0000	0.0000
Cl 2pπ → 4pσ	196.0 (-8.7)	203.2 (-1.5)	214.7 (+10.0)	206.6 (+1.9)	204.7 ^c	0.0002	0.0000	0.0009	0.0016
MAE ^d	24.8	8.7	16.9	0.8					

^a Differences from experimental data are shown in parentheses. ^b Reference 39. ^c Reference 42. ^d Mean absolute errors from experimental data.

Table 7. C-1s, F-1s, Cl-1s, and 2p Excitation Energies of CF₃Cl by TDDFT with B3LYP, BHHLYP, and the Modified CVR-B3LYP Functionals with cc-pCVTZ (in eV)^a

assignment	B3LYP	BHHLYP	CVR-B3LYP	exptl ^b
C 1s → σ* (C-Cl)	283.0 (-11.2)	291.2 (-2.9)	293.9 (-0.3)	294.2
→ σ* (C-F)	286.0 (-10.7)	293.8 (-2.9)	296.9 (+0.2)	296.7
F 1s → σ* (C-Cl)	672.0 (-18.5)	687.5 (-3.0)	688.8 (-1.7)	690.5
→ σ* (C-F)	674.9 (-17.7)	691.0 (-1.6)	691.7 (-0.9)	692.6
Cl 1s → σ* (C-Cl)	2769.1 (-54.4)	2801.0 (-22.5)	2824.9 (+1.4)	2823.5
→ σ* (C-F)	2773.5 (-53.9)	2807.1 (-20.3)	2829.3 (+1.9)	2827.4
Cl 2p → σ* (C-Cl)	194.9 (-6.8)	200.9 (-0.8)	202.7 (+1.0)	201.7 ^c
→ σ* (C-F)	198.8 (-6.0)	206.1 (+1.4)	206.7 (+1.9)	204.8 ^c
MAE ^d	22.4	6.9	1.2	

^a Differences from experimental data are shown in parentheses. ^b Reference 51. ^c Weighted average value calculated with the method in ref 44. ^d Mean absolute errors from experimental data.

B3LYP, BHHLYP, and TDHF, which are calculated to be 24.8, 8.7, and 16.9 eV, respectively. It indicates that the modified CVR-B3LYP provides quite well-balanced results for 1s and 2p excitations. The values of the oscillator strengths of CVR-B3LYP are close to those of the other three methods. In particular, the oscillator strengths of core → unoccupied-valence excitations calculated by CVR-B3LYP are close to those by B3LYP. The oscillator strengths of 4pσ excitations are slightly overestimated by CVR-B3LYP.

In order to investigate the accuracy of core-excited-state calculations on the molecules containing both second- and third-row atoms, TDDFT calculations on CF₃Cl were performed with B3LYP, BHHLYP, and the modified CVR-B3LYP functionals. No Rydberg-basis functions were used in the calculations of CF₃Cl. In order to set HF portions of core and occupied-valence orbitals of second-row atoms to 50% and 20%, which are the same values used in the previous CVR-B3LYP study on core excitations of second-row atoms,²⁰ 1s orbitals of the second-row atoms are treated as C2 orbitals in eq 2. The calculated 1s and 2p core-excitation energies of CF₃Cl are shown in Table 7. As for the core-excitation energies from 1s orbitals of the second-row atoms (C and F), CVR-B3LYP shows higher performance than B3LYP and BHHLYP do. The errors of Cl-1s-excitation energies calculated with the modified CVR-B3LYP are less than 0.5 eV, while those of B3LYP and BHHLYP are about 11 and 3 eV, respectively. The errors

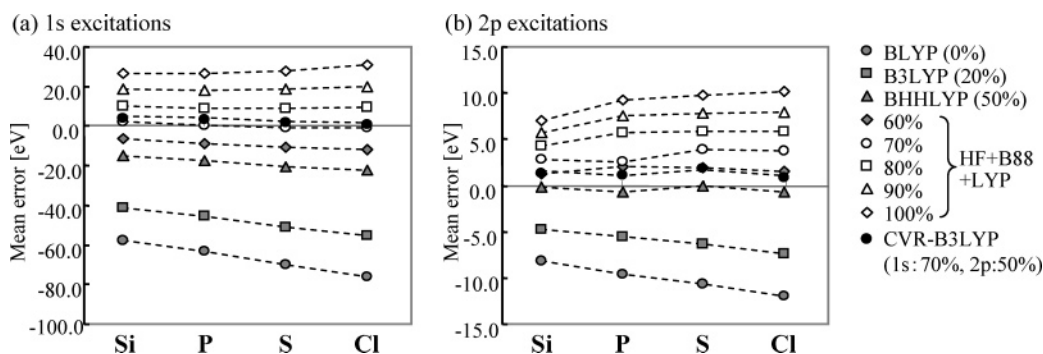
of F1s-excitation energies are underestimated more largely than Cl1s ones. As for the third-row atom (Cl), CVR-B3LYP gives both 1s and 2p core-excitation energies within the errors of 2 eV. The errors of 1s-excitation energies of B3LYP and BHHLYP, which are about 55 and 20 eV, are significantly larger than those of CVR-B3LYP. The accuracy of BHHLYP for Cl2p excitations are comparable to that of CVR-B3LYP. B3LYP has large errors of about 6 eV. Only CVR-B3LYP reproduces core excitation energies of both second- and third-row atoms with reasonable accuracy.

Table 8 shows the 1s and 2p core-excitation energies of SiH₄, PH₃, H₂S, HCl, and Cl₂ molecules calculated with the modified CVR-B3LYP functional. The comparison of the MAEs of CVR-B3LYP in Table 8, (1.5, 1.1) eV for (1s, 2p) core-excitation energies, with those of HF+B88+LYP (70%) and BHHLYP in Table 3 and 4 (1.2, 3.5) and (20.5, 0.6) eV, clarifies that the modified CVR-B3LYP provides well-balanced results for any third-row atoms. As for 1s excitations, the accuracy of the modified CVR-B3LYP is comparable to that of HF+B88+LYP (70%) and significantly higher than that of BHHLYP. On the other hand, the MAE of CVR-B3LYP for 2p core excitations is closer to that of BHHLYP than that of HF+B88+LYP (70%). The modified CVR-B3LYP gives more accurate results than HF+B88+LYP (70%) does for 2p excitations. Thus, it is demonstrated that the modified CVR-B3LYP shows high performance both for K-shell and L-shell core excitations, whereas the conven-

Table 8. 1s and 2p Core-Excitation Energies of SiH₄, PH₃, H₂S, HCl, and Cl₂ Molecules by TDDFT with the Modified CVR-B3LYP Functional with cc-pCVTZ Plus Rydberg Basis Functions (in eV)^a

molecule	1s excitation			2p excitation		
	assignment	CVR-B3LYP	exptl	assignment	CVR-B3LYP	exptl
SiH ₄	Si 1s → σ*	1846.6 (+4.1)	1842.5 ^b	Si 2p → σ*	103.7 (+0.9)	102.8 ^{g,i}
PH ₃	P 1s → σ*(e)	2148.9 (+3.1)	2145.8 ^c	P 2p → σ*	133.1 (+0.8)	132.3 ^{g,i}
H ₂ S	S 1s → 3b ₂ (σ*)	2474.7 (+1.6)	2473.1 ^d	S 2p → σ*	166.1 (+1.6)	164.5 ^f
	S 1s → 4p _{b2}	2477.4 (+1.1)	2476.3 ^d	S 2p → 4s	168.3 (+1.8)	166.5 ^f
HCl	Cl 1s → 3pσ*	2824.8 (+0.9)	2823.9 ^e	Cl 2pπ → 3pσ*	202.0 (+1.0)	201.0 ^g
	Cl 1s → 4pπ	2827.9 (+0.1)	2827.8 ^e	Cl 2pπ → 4pπ	205.0 (+0.4)	204.6 ^g
Cl ₂	Cl 1s → 3pσ _u *	2822.1 (+0.8)	2821.3 ^e	Cl 2pπ → 3pσ _u *	199.1 (+0.4)	198.7 ^{h,i}
	Cl 1s → 4p	2829.2 (+0.7)	2828.5 ^e	Cl 2pπ → 4s	205.8 (+1.0)	204.8 ^{h,i}
MAE ^j		1.5			1.0	

^a Differences from experimental data are shown in parentheses. ^b Reference 36. ^c Reference 37. ^d Reference 38. ^e Reference 39. ^f Reference 40. ^g Reference 41. ^h Reference 43. ⁱ Weighted average value calculated with the method in ref 44. ^j Mean absolute errors from experimental data.

**Figure 2.** Atom-dependent mean errors of 1s and 2p excitation energies by TDHF and TDDFT with the BLYP, B3LYP, BHHLYP, HF+B88+LYP, and CVR-B3LYP functionals with cc-pCVTZ plus Rydberg basis functions.

tional functionals can describe either K-shell or L-shell excitation with high accuracy. Figure 2 shows the atom-dependent MEs of Z1s and Z2p core-excitation energies (Z = Si, P, S, and Cl) calculated with BLYP, B3LYP, BHHLYP, HF+B88+LYP (X%) (X = 60, 70, 80, 90, and 100), and the modified CVR-B3LYP, which are tabulated in Tables 3, 4, and 8. Both the underestimation with a small portion of HF exchange in the functional and the overestimation with a large portion of HF exchange becomes larger in the order, Si < P < S < Cl, which means that the errors become larger for heavier atom species, i.e., deeper K-shell and L-shell orbitals. In Figure 2(a) for 1s core-excitation energies, the errors of HF+B88+LYP(80%) and HF+B88+LYP(90%) are less atom-dependent, while those of BLYP, B3LYP, and BHHLYP largely depend on the kind of atom: The range of errors for Si, P, S, and Cl atoms are 18.5, 14.1, 7.6, 5.4, 3.1, 1.6, 2.0, 3.7, and 3.5 eV for BLYP, B3LYP, BHHLYP, HF+B88+LYP (X%) (X = 60, 70, 80, 90, and 100), and CVR-B3LYP, respectively. The atom-dependency of CVR-B3LYP is comparable to that of HF+B88+LYP(70%). For 2p core-excitation energies in Figure 2(b), the range of errors for Si, P, S, and Cl atoms are 8.3, 2.5, 0.6, 0.7, 1.3, 1.6, 2.3, 3.1, 2.9, and 0.9 eV for BLYP, B3LYP, BHHLYP, HF+B88+LYP (X%) (X = 60, 70, 80, 90, and 100), and CVR-B3LYP. CVR-B3LYP, BHHLYP, and HF+B88+LYP(60%) have significantly less atom-dependency.

In order to assess the accuracy of the description of occupied-valence electrons, excitation energies from oc-

cupied-valence orbitals of SiH₄, PH₃, H₂S, HCl, and Cl₂ molecules were calculated by TDHF and TDDFT with B3LYP, BHHLYP, and the modified CVR-B3LYP. Table 9 lists the calculated excitation energies. In Table 9, BHHLYP shows high performance, and the accuracy of BLYP, B3LYP, and TDHF are slightly worse than BHHLYP: The MAEs of BLYP, B3LYP, BHHLYP, and TDHF are 0.8, 0.5, 0.3, and 0.7 eV, respectively. The excitation energies of CVR-B3LYP are close to and higher than those of B3LYP for occupied-valence → unoccupied-valence and occupied-valence → Rydberg excitations, respectively. This is because the valence and Rydberg orbitals of CVR-B3LYP are designed to be similar to those of B3LYP and HF. The MAE of CVR-B3LYP is 0.6 eV, which is comparable to that of B3LYP. Therefore, CVR-B3LYP describes valence-excitation energies with reasonable accuracy as like conventional DFT methods.

The standard enthalpies of formation of SiH₄, PH₃, H₂S, HCl, and Cl₂ molecules, which is one of the valence-electron properties in the ground states, were calculated by the procedure mentioned in ref 55. The results of HF and DFT calculations with the BLYP, B3LYP, BHHLYP, HF+B88+LYP (X%) (X = 60, 70, 80, 90, and 100), and CVR-B3LYP functionals are shown in Table 10. The DFT method gives more accurate results than the HF method does: The MAE of the HF method is 52.0 kcal/mol, while all of the MAEs of the DFT methods are less than 10 kcal/mol. The accuracy of BLYP and B3LYP are significantly high among the DFT methods, whose MAEs are 2.0 and 1.5 kcal/mol. The MAE

Table 9. Valence- and Rydberg-Excitation Energies of SiH₄, PH₃, H₂S, HCl, and Cl₂ Molecules by TDHF and TDDFT with BLYP, B3LYP, BHHLYP, and the Modified CVR-B3LYP Functional with cc-pCVTZ Plus Rydberg Basis Functions (in eV)^a

molecule	assignment	BLYP	B3LYP	BHHLYP	TDHF	CVR-B3LYP	exptl
SiH ₄	t ₂ → 4s	8.0 (−0.8)	8.5 (−0.3)	9.2 (+0.4)	9.9 (+1.1)	9.4 (+0.6)	8.8 ^b
PH ₃	n → 4p	6.8 (−1.0)	7.2 (−0.6)	8.0 (+0.2)	8.4 (+0.6)	8.8 (+1.0)	7.8 ^c
H ₂ S	2b ₁ → σ*	5.8 (+0.4)	6.0 (+0.5)	6.1 (+0.6)	6.2 (+0.8)	6.0 (+0.6)	5.5 ^c
HCl	3pπ → 4s	8.3 (−1.3)	8.9 (−0.7)	9.8 (+0.2)	10.5 (+0.9)	9.8 (+0.2)	9.6 ^d
Cl ₂	π _g → σ _u	3.2 (−0.6)	3.3 (−0.4)	3.6 (−0.2)	4.0 (+0.2)	3.3 (−0.5)	3.8 ^d
MAE ^e		0.8	0.5	0.3	0.7	0.6	

^a Differences from experimental data are shown in parentheses. ^b Reference 52. ^c Reference 53. ^d Reference 54. ^e Mean absolute errors from experimental data.

Table 10. Standard Enthalpies of Formation of SiH₄, PH₃, H₂S, HCl, and Cl₂ Molecules by HF and DFT with the BLYP, B3LYP, BHHLYP, HF+B88+LYP, and Modified CVR-B3LYP Functionals with cc-pCVTZ Plus Rydberg Basis Functions (in eV)^a

molecule	BLYP	B3LYP	BHHLYP	HF+B88+LYP					HF	CVR-B3LYP	exptl ^b
				60%	70%	80%	90%	100%			
SiH ₄	13.3 (+5.1)	7.9 (−0.3)	7.9 (−0.3)	6.5 (−1.7)	5.1 (−3.1)	3.7 (−4.5)	2.1 (−6.1)	0.5 (−7.7)	75.0 (+66.8)	5.9 (−2.3)	8.2
PH ₃	1.2 (−0.1)	−0.4 (−1.7)	3.4 (+2.1)	3.5 (+2.2)	3.6 (+2.3)	3.7 (+2.4)	3.6 (+2.3)	3.5 (+2.2)	71.7 (+70.4)	−2.5 (−3.8)	1.3
H ₂ S	−2.8 (+2.1)	−3.7 (+1.2)	0.5 (+5.4)	1.0 (+5.9)	1.5 (+6.4)	2.0 (+6.9)	2.4 (+7.3)	2.8 (+7.7)	48.7 (+53.6)	−5.4 (−0.5)	−4.9
HCl	−19.9 (+2.2)	−20.3 (+1.8)	−17.5 (+4.6)	−17.0 (+5.1)	−16.6 (+5.5)	−16.1 (+6.0)	−15.7 (+6.4)	−15.3 (+6.8)	7.7 (+29.8)	−21.5 (+0.6)	−22.1
Cl ₂	−0.5 (−0.5)	2.7 (+2.7)	10.3 (+10.3)	12.2 (+12.2)	14.1 (+14.1)	15.9 (+15.9)	17.6 (+17.6)	19.3 (+19.3)	39.4 (+39.4)	2.1 (+2.1)	0.0
MAE ^c	2.0	1.5	4.5	5.4	6.3	7.1	7.9	8.7	52.0	1.9	

^a Differences from experimental data are shown in parentheses. ^b Reference 55. ^c Mean absolute errors from experimental data.

becomes larger as the portion of HF exchange increases. Therefore, the appropriate portion of HF exchange for describing valence electrons is suggested to be 0%–20%. The accuracy of CVR-B3LYP with a MAE of 1.9 kcal/mol is comparable to BLYP and B3LYP. Thus, we confirm that CVR-B3LYP is capable of describing the behaviors of not only K-shell and L-shell electrons but also valence ones with reasonable accuracy, while HF+B88+LYP (70%) and BHHLYP are appropriate only for K-shell and L-shell excitations, respectively.

4. Conclusions

The CVR-B3LYP functional is extended to core-excited-state calculations of the molecules containing third-row atoms. The assessment of TDDFT calculations with conventional exchange-correlation functionals demonstrates that 70% and 50% portions of HF exchange are appropriate for calculating K-shell and L-shell core-excitation energies, respectively. Therefore, the CVR-B3LYP functional is modified to possess the appropriate portions of HF exchange for K-shell, L-shell, and occupied-valence regions separately. TDDFT calculations on HCl, CF₃Cl, and several molecules containing third-row atoms show that the modified CVR-B3LYP functional reproduces the K-shell and L-shell core-excitation energies with reasonable accuracy. For valence properties, the calculations of valence-excitation energies and standard enthalpies of formation confirm that CVR-B3LYP describes valence electrons accurately as well as B3LYP does. The numerical assessments have revealed the

high accuracy of CVR-B3LYP for the descriptions of all of the K-shell, L-shell, and valence electrons.

Appendix

Size-Consistency and Size-Extensivity. CVR-B3LYP does not satisfy size-consistency rigorously: In the case that electrons transfer between two categories in the dissociation process, CVR-B3LYP is size-inconsistent. However, the orbitals are categorized into three groups, K-shell-, L-shell- and valence-orbital groups in the present study, and electrons hardly transfer between two different categories in the process of the dissociation in most realistic cases. Figure 3 shows the dissociation curves of HCl, which are obtained by performing unrestricted DFT calculations with BHHLYP, B3LYP, and CVR-B3LYP. The energies at the largest H–Cl distance in the calculations are set to zero in Figure 3. The electrons transferred in the dissociation are valence ones in most cases. CVR-B3LYP well reproduces the curve of B3LYP, because the valence orbitals of CVR-B3LYP are designed to reproduce those of B3LYP. Thus, CVR-B3LYP is size-consistent in practical cases.

On the other hand, CVR-B3LYP satisfies size-extensivity. We have numerically examined size-extensive nature by performing CVR-B3LYP calculations of a Cl₂ monomer and (Cl₂)₂ dimer separated at 50 Å. The energy difference between two Cl₂ monomers and the largely separated (Cl₂)₂ dimer is only 6.3*10^{−08} hartree, which indicates that CVR-B3LYP is size-extensive.

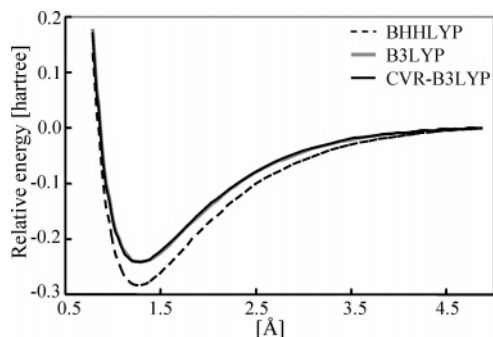


Figure 3. Potential energy curves for the dissociation of HCl calculated by DFT with the B3LYP, BHHLYP, and CVR-B3LYP functionals with cc-pCVTZ plus Rydberg basis functions.

Acknowledgment. The calculations were performed in part at the Research Center for Computational Science (RCCS) of the Okazaki National Research Institutes. This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas “Molecular Theory for Real Systems” ‘KAKENHI 18066016’ from Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT), a 21st century Center of Excellence (21COE) “Practical Nano-Chemistry” from MEXT, the Next Generation Super Computing Project, Nanoscience Program, of MEXT, and a project research grant “Development of high-performance computational environment for quantum chemical calculation and its assessment” from the Advanced Research Institute for Science and Engineering (RISE) of Waseda University. Two of the authors (A.N. and Y.I.) are indebted to the Japanese Society for the Promotion of Science (JSPS) Research Fellowship for Young Scientists.

References

- Casida, M. E. In *Recent Advances in Density Functional Methods*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; Part I, Chapter 5, pp 155–192.
- Bauernschmitt, R.; Ahlrichs, R. *Chem. Phys. Lett.* **1996**, *256*, 454–464.
- Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218–8224.
- Hirata, S.; Head-Gordon, M. *Chem. Phys. Lett.* **1999**, *314*, 291–299.
- Hirata, S.; Head-Gordon, M.; Bartlett, R. J. *J. Chem. Phys.* **1999**, *111*, 10774–10786.
- Jamorski, C.; Casida, M. E.; Salahub, D. R. *J. Chem. Phys.* **1996**, *104*, 5134–5147.
- Casida, M. E.; Jamorski, C.; Casida, K. C.; Salahub, D. R. *J. Chem. Phys.* **1998**, *108*, 4439–4449.
- van Gisbergen, S. J. A.; Snijders, J. G.; Baerends, E. J. *J. Chem. Phys.* **1995**, *103*, 9347–9354.
- Nakata, A.; Imamura, Y.; Ostuka, T.; Nakai, H. *J. Chem. Phys.* **2006**, *124*, 094105.
- Imamura, Y.; Otsuka, T.; Nakai, H. *J. Comput. Chem.* In press.
- Imamura, Y.; Nakai, H. *Int. J. Quantum Chem.* **2007**, *107*, 23–29.
- Imamura, Y.; Nakai, H. *Chem. Phys. Lett.* **2006**, *419*, 297–303.
- Casida, M. E.; Salahub, D. R. *J. Chem. Phys.* **2000**, *113*, 8918–8935.
- Appel, H.; Gross, E. K. U.; Burke, K. *Phys. Rev. Lett.* **2003**, *90*, 043005.
- Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, *23*, 5048–5079.
- van Leeuwen, R.; Baerends, E. J. *Phys. Rev. A* **1994**, *49*, 2421–2431.
- Schipper, P. R. T.; Gritsenko, O. V.; van Gisbergen, S. J. A.; Baerends, E. J. *J. Chem. Phys.* **2000**, *112*, 1344–1352.
- Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 10180–10189.
- Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425–8433.
- Nakata, A.; Imamura, Y.; Nakai, H. *J. Chem. Phys.* **2006**, *125*, 064109.
- Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- Hu, C.-H.; Chong, D. P. *Chem. Phys. Lett.* **1996**, *262*, 729–732.
- Chong, D. P. *J. Electron Spectrosc. Relat. Phenom.* **2005**, *148*, 115–121.
- Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.
- Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358–1371.
- Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1995**, *103*, 4572–4585.
- Peterson, K. A.; Dunning, T. H., Jr. *J. Chem. Phys.* **2002**, *117*, 10548–10560.
- Dunning, T. H., Jr.; Hay, P. J. In *Methods of Electronic Structure Theory*; Schaefer, H. F., III, Ed.; Plenum Press: New York, 1977; Vol. 3.
- Magnusson, E.; Schaefer, H. F., III *J. Chem. Phys.* **1985**, *83*, 5721–5726.
- Dunning, T. H., Jr.; Harrison, P. J. In *Modern Theoretical Chemistry*; Schaefer, H. F., III, Ed.; Plenum Press: New York, 1977; Vol. 2.
- Nakajima, T.; Hirao, K. *Chem. Phys. Lett.* **1999**, *302*, 383–391.
- Fedorov, D. G.; Nakajima, T.; Hirao, K. *Chem. Phys. Lett.* **2001**, *335*, 183–187.
- Bodeur, S.; Millié, P.; Nenner, I. *Phys. Rev. A* **1990**, *41*, 252–263.
- Cavell, R. G.; Jürgensen, A. *J. Electron Spectrosc. Relat. Phenom.* **1999**, *101–103*, 125–129.
- Bodeur, S.; Esteva, J. M. *Chem. Phys.* **1985**, *100*, 415–427.
- Bodeur, S.; Maréchal, J. L.; Reynaud, C.; Bazin, D.; Nenner, I. *Z. Phys. D-Atoms, Molecules Clusters* **1990**, *17*, 291–298.
- Robin, M. B. *Chem. Phys. Lett.* **1975**, *31*, 140–144.

- (41) Gedat, E.; Püttner, R.; Domke, M.; Kaindl, G. *J. Chem. Phys.* **1998**, *109*, 4471–4477.
- (42) Fronzoni, G.; Stener, M.; Decleva, P.; De Alti, G. *Chem. Phys.* **1998**, *232*, 9–23.
- (43) Nayandin, O.; Kukk, E.; Wills, A. A.; Langer, B.; Bozek, J. D.; Canton-Rogan, S.; Wiedenhoeft, M.; Cubaynes, D.; Berrah, N. *Phys. Rev. A* **2001**, *63*, 062719.
- (44) Segala, M.; Takahata, Y.; Chong, D. P. *J. Electron Spectrosc. Relat. Phenom.* **2006**, *151*, 9–13.
- (45) The three- and higher-body interactions are neglected in eq 2. The energy differences due to the truncation are about 0.02% of the total energies for HCl, Cl₂, H₂S, PH₃, and SiH₄ molecules. Furthermore, the excitation energies and the standard enthalpies of formation, which correspond to the energy differences between two or more states, have been calculated accurately under the truncation. Therefore, the effect of the truncation seems negligible in the present study.
- (46) Slater, J. C. *Phys. Rev.* **1951**, *81*, 385–390.
- (47) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (48) Roothaan, C. C. J. *Rev. Mod. Phys.* **1960**, *32*, 179–185.
- (49) Huzinaga, S. In *Bunshikidouhou*; Iwanami Shoten: Tokyo, 1980; pp 131–147 (in Japanese).
- (50) Hirao, K.; Nakatsuji, H. *J. Chem. Phys.* **1973**, *59*, 1457–1462.
- (51) Zhang, W.; Ibuki, T.; Brion, C. E. *Chem. Phys.* **1992**, *160*, 435–450.
- (52) Itoh, U.; Toyoshima, Y.; Onuki, H. *J. Chem. Phys.* **1986**, *85*, 4867–4872.
- (53) Robin, M. B. In *Higher Excited States of Polyatomic Molecules*; Academic Press: New York and London, 1974; Vol. I, Chapter III.
- (54) Huber, K. P.; Herzberg, G. In *Molecular Spectra and Molecular Structure IV. Constants of Diatomic Molecules*; Van Nostrand Reinhold: New York, 1979.
- (55) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

CT600368F

CASCI Reference Wave Functions for Multireference Perturbation Theory Built from Hartree–Fock or Kohn–Sham Orbitals

David Robinson and Joseph J. W. McDouall*

*School of Chemistry, The University of Manchester, Oxford Road,
Manchester M13 9PL, U.K.*

Received February 13, 2007

Abstract: The MRMP2 method and many similar variants of multireference perturbation theory have a potentially wide range of applicability. However they typically require a CASSCF calculation to define the reference wave function. It is worthwhile to investigate whether ‘simpler’ orbitals than those obtained from the full CASSCF procedure can provide useful accuracy. In this study we investigate six reactions taken from the Zhao–González-García–Truhlar database and investigate the MRMP2 procedure when used with a variety of different orbital sets in order to assess the reliability of such procedures. The results are encouraging and suggest that multireference perturbation theory may be used, for some systems, with the simplified procedures presented here.

1. Introduction

In studying the electronic structure of molecules there are many well-known situations that require a many determinant approach in order to obtain a description that is even qualitatively correct. Examples of such situations include the following: the computation of potential energy curves far from equilibrium; certain types of excited states; the location of transition structures containing diradical character; and the mapping of complete reaction paths. The most commonly used multideterminant method is the complete active space self-consistent field (CASSCF) method,¹ that deals with the nondynamic (structure-dependent) correlation, but does not account for the dynamic electron correlation to any significant degree. The dynamic electron correlation must be dealt with by the multireference analogues of perturbation,² configuration interaction,³ and coupled-cluster⁴ theories. In particular among these methods are a number of multireference perturbation theories^{5–11} that have been developed and applied with considerable success. The popularity of these methods stems from their relative computational efficiency. This gives a manageable cost/accuracy ratio for dealing with multiconfigurational problems.

The use of CASSCF based methods presents additional levels of complexity for the user when compared with single reference methods. The most obvious conceptual challenge is to choose a meaningful active space for describing a given chemical problem.¹² Assuming this can be done reliably, the next challenge is to converge the CASSCF wave function. Each cycle of the CASSCF orbital optimization involves a partial integral transformation from the atomic orbital to the molecular orbital basis. When large active spaces are used, a substantial CI eigenvalue problem must also be solved in each cycle. These factors make the CASSCF procedure relatively demanding in terms of computational resources. A number of groups^{13–16} have investigated the possibility of avoiding the CASSCF step, by using orbitals obtained from simpler methods to define the active spaces for use in multireference treatments. In a recent paper¹⁷ we have also studied this matter for the case of the $X^1\Sigma_g^+$, $B^1\Delta_g$, and $B^1\Sigma_g^+$ state potential energy curves of the C_2 molecule. We used complete active space configuration interaction (CASCI) reference wave functions in a multireference perturbation theory scheme. The CASCI wave functions were built from Hartree–Fock or Kohn–Sham orbitals with no further refinement of the orbital sets. The C_2 potential energy curves provide demanding multiconfigurational test cases for which full CI results have been published. In comparing our

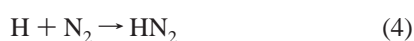
* Corresponding author phone: +44 (0)161-275-4720; fax: +44 (0)161-275-4598; e-mail: joe.mcdouall@manchester.ac.uk.

Table 1. Reference Values for the Forward (V_r^{f}) and Reverse (V_r^{r}) Reaction Barriers (kcal mol⁻¹) for Reactions 1–6 Taken from the Minnesota Database Collection^{18,19}

	reaction	barrier	database value
1	H + FH → HF + H	V_{fr}^{f}	42.18
2	H + ClH → HCl + H	V_{fr}^{f}	18.00
3	H + F ₂ → HF + F	V_f^{f}	2.27
4	H + N ₂ → HN ₂	V_r^{f}	106.18
		V_f^{f}	14.69
5	H + CO → HCO	V_r^{f}	10.72
		V_f^{f}	3.17
6	HCN → HNC	V_r^{f}	22.68
		V_f^{f}	48.16
		V_r^{f}	33.11

calculations with the benchmark results we found that reliable results were obtained with a variety of orbitals provided that the region of interest was not too far from equilibrium. For accurate reproduction of full potential energy curves only the curves built from fully optimized CASSCF orbitals were adequate. However, since many studies are primarily concerned with regions of the potential energy surface not too far from equilibrium, one would like to encourage the use of orbitals generated by simpler techniques, provided that the results remain consistent in a wide range of applications. This will be important in enabling multireference perturbation theory methods to be applied to a wider variety of systems, given that the CASSCF step in the calculations is avoided.

In this work we wish to investigate whether the encouraging results we obtained in the study of diatomic potential energy curves carry over to the study of chemical reactions. In particular, transition states often possess electronic structures that are multideterminantal in nature, as they are typically intermediate between two different bonding situations. Zhao, González-García, and Truhlar^{18,19} have produced a database of barrier heights for heavy atom transfer, nucleophilic substitution, and unimolecular and association reactions. We have chosen 6 examples, (1)–(6) below, from this database with which to test our procedure.



Calculating forward and backward reactions gives 10 barrier heights for comparison with our calculations. Reactions 1–5 possess an overall spin, $s = 1/2$, and so refer to doublet surfaces. Reaction 6, with $s = 0$, refers to a singlet surface. Table 1 collects the database reference values for the barrier heights.

These reference values have been obtained by the WI theory as described in ref 18 and include corrections for a variety of factors including relativistics, core correlation, and

spin-orbit effects. For the reactions we have studied, the net effect of these additional corrections is on average <0.07 kcal mol⁻¹ with a maximum correction of 0.43 kcal mol⁻¹ for reaction 3. Hence it is acceptable to compare our computed results directly with these reference values.

2. Computational Details

Our multireference perturbation theory program follows the multireference second-order Møller–Plesset perturbation theory (MRMP2) formalism of Hirao.^{7–9} In this approach, the first-order density matrix, γ , obtained from the full CI expansion in the chosen active space is used to construct the matrix representation of the generalized Fock operator, F :

$$F_{pq} = h_{pq} + \sum_{ij}^{\text{occupied}} \gamma_{ij} \left[(pq | ij) - \frac{1}{2} (pj | iq) \right] \quad (7)$$

For a CASCI expansion the energy is invariant to rotations within the inactive, active, and virtual orbital subspaces. F is canonicalized within each subspace, and the resulting diagonal elements are used to define the eigenvalues, $E^{(0)}$, of the model Hamiltonian, H_0 . The second-order correction to the energy is given by

$$E_p^{(2)} = - \sum_Q \frac{|\langle Q | H | P \rangle|^2}{E_Q^{(0)} - E_p^{(0)}} \quad (8)$$

The CASSCF or CASCI state is labeled by P , and Q refers to one of the set of all allowed double excitations between the orbital subspaces. Calculations based on (7) and (8) with different types of orbital are denoted as the MRMP2(*method*), where *method* refers to the level of theory used to define the orbitals. We have investigated orbitals obtained from Hartree–Fock (HF) and Kohn–Sham (KS) calculations, the latter using the B3LYP and BLYP exchange–correlation functionals. Hence we report results at the MRMP2-(CASSCF), MRMP2(HF), MRMP2(BLYP), and MRMP2-(B3LYP) levels.

The issue of intruder states is important in multireference perturbation theory. A simple scheme has been developed for intruder state avoidance (ISA) within the formalism of Hirao's MRMP2.^{20,21} In this approach, eq 8 is modified by introducing a shift in the denominator

$$E_p^{(2)-\text{ISA}} = - \sum_Q \frac{|\langle Q | H | P \rangle|^2}{E_Q^{(0)} - E_p^{(0)} + \Delta_Q}, \quad \text{where } \Delta_Q = \frac{b}{E_Q^{(0)} - E_p^{(0)}} \quad (9)$$

We have also recalculated all results using this technique. The value of the parameter, b , which is used to define the energy denominator shifts, is that recommended in ref 21 ($b = 0.02$). These results are denoted as the MRMP2-ISA-(*method*).

All geometries were taken from ref 19 and refer to the QCISD/MG3 level. The MG3 basis consists of the 6-311++G(3d2f,2df,2p)²² basis set for the atoms H–Si, with an extended basis for atoms P–Ar.^{23,24} The calculations we report use the MG3S basis, which is equivalent to the MG3

Table 2. Barrier Heights for Reactions 1–6 at the CASSCF, MRMP2(CASSCF), and MRMP2-ISA(CASSCF) Levels^a

reaction	barrier	CASSCF	€	MRMP2	€	MRMP2-ISA	€
1	$V_{f,r}^{\ddagger}$	55.08	12.90	41.62	-0.56	41.77	-0.41
2	$V_{f,r}^{\ddagger}$	29.08	11.08	17.23	-0.77	17.35	-0.65
3	V_f^{\ddagger}	7.66	5.39	2.04	-0.23	2.09	-0.18
	V_r^{\ddagger}	105.48	-0.70	104.19	-1.99	104.27	-1.91
4	V_f^{\ddagger}	26.34	11.65	14.56	-0.13	14.69	0.00
	V_r^{\ddagger}	-0.56	11.28	11.74	1.02	11.74	1.02
5	V_f^{\ddagger}	10.02	6.85	3.36	0.19	3.42	0.25
	V_r^{\ddagger}	11.16	-11.52	22.86	0.18	22.85	0.17
6	V_f^{\ddagger}	53.89	5.73	48.38	0.22	48.42	0.26
	V_r^{\ddagger}	37.49	4.38	33.55	0.44	33.57	0.46
	$ \bar{\epsilon} $		8.15		0.57		0.53
	max $ \epsilon $		12.90		1.99		1.91

^a Absolute values and errors (€) are given in kcal mol⁻¹.

basis except for the case of H atom, for which the diffuse functions are excluded. We also include a wider study of basis set influence, which is given in the Supporting Information and discussed at the end of the next section. In calculating the energies of reactants and products, the systems were treated as supermolecules with a separation between moieties of 100 Å. This avoids any issues related to size-consistency of the MRMP2 approach, and a detailed discussion can be found in refs 25 and 26.

In all cases, a HF or KS calculation was run, followed by a wave function stability analysis and, if necessary, reoptimization of the wave function. For stable wave functions, spin-restricted calculations (for closed- and open-shell systems, respectively) were used to generate the initial orbitals for the CASSCF calculation. In the cases where wave function instabilities were found, the spin-unrestricted natural orbitals were used as the initial orbitals for the CASSCF calculation. The same initial orbitals were also used, without further optimization, to perform a CASCI calculation to define the target state used in MRMP2 calculations. The active space in all calculations consists of the full valence shell orbitals of all atoms. The CASSCF/CASCI and MRMP2/MRMP2-ISA calculations were performed using our in-house codes which we have interfaced with the Gaussian 03 suite of programs.²⁷ All atomic orbital integrals were obtained using standard procedures in Gaussian 03, as were the HF, BLYP, B3LYP, and stability calculations.

3. Results

We begin with the CASSCF and MRMP2(CASSCF) and MRMP2-ISA(CASSCF) results, given in Table 2. The MRMP2 and MRMP2-ISA calculations based on CASSCF orbitals are our best estimates of the barriers for reactions 1–6.

As is to be expected the CASSCF results show significant errors, since no appreciable account of the effects of dynamic electron correlation is included at this level. The MRMP2-(CASSCF) results show good agreement with the database values. With a full valence shell active space and a large basis set, the MRMP2(CASSCF) method should provide

Table 3. Barrier Heights for Reactions 1–6 Obtained from CASCI Wave Functions Built from HF, B3LYP, and BLYP Orbitals^a

reaction	barrier	CASCI-(HF)	€	CASCI-(B3LYP)	€	CASCI-(BLYP)	€
1	$V_{f,r}^{\ddagger}$	41.96	-0.22	57.42	15.24	56.14	13.96
2	$V_{f,r}^{\ddagger}$	40.49	22.49	19.66	1.66	21.00	3.00
3	V_f^{\ddagger}	44.41	42.14	4.45	2.18	11.57	9.30
	V_r^{\ddagger}	127.93	21.75	94.53	-11.65	107.12	0.94
4	V_f^{\ddagger}	-33.17	-47.86	31.46	16.77	30.36	15.67
	V_r^{\ddagger}	-5.17	-15.89	-0.86	-11.58	-3.95	-14.67
5	V_f^{\ddagger}	10.61	7.44	15.66	12.49	13.81	10.64
	V_r^{\ddagger}	60.52	37.84	8.84	-13.84	4.52	-18.16
6	V_f^{\ddagger}	43.01	-5.15	40.95	-7.21	43.38	-4.78
	V_r^{\ddagger}	33.16	0.05	30.22	-2.89	30.54	-2.57
	$ \bar{\epsilon} $		20.08		9.55		9.37
	max $ \epsilon $		47.86		16.77		18.16

^a Absolute values and errors (€) are given in kcal mol⁻¹.

good accuracy and the mean absolute error of 0.57 kcal mol⁻¹ is very acceptable. The MRMP2-ISA(CASSCF) results show a slight improvement over the MRMP2(CASSCF) giving a mean absolute error of 0.53 kcal mol⁻¹. Considering the shift, Δ_Q , in eq 9, its effect is to essentially remove the contribution of a double excitation from the perturbation expansion if the energy of that determinant approaches the energy of the reference state. If the contribution of such an intruder state is significant, then eq 9 will not correct the situation, and a substantial error in the perturbation energy may be expected. In such circumstances one must either expand the reference space to include the intruder state or use a multistate method. Given the relatively small effect on the barriers dealt with here, we may conclude that there are no major intruder state problems associated with the systems studied.

We next consider the barrier heights calculated using a CASCI reference wave function, in which the CAS expansion is built from orbitals obtained by standard HF or KS methods. As with the CASSCF results, the CASCI errors are generally quite large and can be attributed to the lack of sufficient dynamical correlation within the CASCI wave function. It is noteworthy that the average error obtained with HF orbitals is significantly larger than when KS orbitals are employed. Typically the KS orbitals reduce the average error by a factor of 2, and the maximum error is reduced by a factor of approximately 3. Table 3 gives the relevant results.

Now adding the dynamic electron correlation we find that the MRMP2 correction to the CASCI reference shows quite good agreement with the database values. The mean absolute errors are 1.43, 1.40, and 1.87 kcal mol⁻¹ for MRMP2(HF), MRMP2(B3LYP), and MRMP2(BLYP), respectively. The results are shown in Table 4. The HF and B3LYP orbitals behave quite similarly, whereas the BLYP orbitals show significantly increased mean and maximum errors.

In the case of the HF and B3LYP orbitals, the maximum error occurs for reaction 6. Comparing the CI vectors for the CASCI wave functions with that of the CASSCF reveals a need for CASSCF orbitals in this case. In HCN, the contribution of the determinants describing the $\pi \rightarrow \pi^*$ (doubly degenerate) excitation is underestimated via the CASCI methods. In the CASSCF calculation the HF

Table 4. Barrier Heights for Reactions 1–6 Obtained from MRMP2 Calculations in Which the Reference CASCI Wave Function Is Built from HF, B3LYP, and BLYP Orbitals^a

reaction	barrier	MRMP2- (HF)	ϵ	MRMP2- (B3LYP)	ϵ	MRMP2- (BLYP)	ϵ
1	V_{fr}^{\ddagger}	42.42	0.24	40.98	-1.20	40.92	-1.26
2	V_{fr}^{\ddagger}	16.30	-1.70	17.26	-0.74	17.87	-0.13
3	V_f^{\ddagger}	2.31	0.04	1.42	-0.85	0.57	-1.70
4	V_r^{\ddagger}	105.22	-0.96	103.65	-2.53	100.35	-5.83
	V_f^{\ddagger}	15.11	0.42	13.04	-1.65	13.17	-1.52
5	V_r^{\ddagger}	10.07	-0.65	10.92	0.20	12.04	1.32
	V_f^{\ddagger}	5.30	2.13	1.97	-1.20	2.27	-0.90
6	V_r^{\ddagger}	23.89	1.21	23.01	0.33	23.91	1.23
	V_f^{\ddagger}	52.95	4.79	51.55	3.39	51.06	2.90
	V_r^{\ddagger}	35.31	2.20	34.97	1.86	35.00	1.89
	$ \bar{\epsilon} $		1.43		1.40		1.87
	$\max \epsilon $		4.79		3.39		5.83

^a Absolute values and errors (ϵ) are given in kcal mol⁻¹.

determinant has a coefficient of 0.961, and the $\pi \rightarrow \pi^*$ excitations have a coefficient of -0.111. When HF orbitals are used the coefficient of the HF determinant in the CASCI wave function is 0.998, with negligible contribution from the $\pi \rightarrow \pi^*$ excitations. The situation is somewhat improved when KS orbitals are used, and the coefficients then become for B3LYP orbitals 0.976 (HF determinant) and -0.103 ($\pi \rightarrow \pi^*$ excitations). BLYP orbitals produce coefficients of 0.970 (HF determinant) and -0.114 ($\pi \rightarrow \pi^*$ excitations). In the transition structure, a similar situation is found. In Figure 1, one of the symmetry unique π^* orbitals of HCN and the corresponding orbital in the transition structure are shown for the different levels of theory considered. These are the orbitals obtained following canonicalization and are shown on an equal scale and orientation and so may be directly compared. On the left-hand side is shown the reactant and on the right-hand side the transition structure. Looking down either column of Figure 1 we immediately note that the most compact π^* orbitals are obtained by the full CASSCF optimization, Figure 1(a). Conversely the most diffuse orbitals are obtained by the HF procedure, Figure 1(b). The KS orbitals (Figure 1(c),(d)) are intermediate between those of the CASSCF and HF orbitals. The B3LYP orbitals are slightly more compact for HCN than the BLYP orbitals; however, both sets of KS orbitals show an exaggerated polarization of the lobes away from each other. These are subtle effects but clearly have an effect on the CI coefficients and consequently the predicted barrier heights.

Clearly the poor description of these virtual orbitals by the non-CASSCF methods leads to the underestimation of the above determinants in the CASCI wave functions.

Finally the MRMP2-ISA results are shown in Table 5. Small improvements are observed for the KS orbitals, while the HF results change by only 0.01 kcal mol⁻¹ on average. There is no significant difference between the results with and without the ISA corrections, implying that the set of reactions chosen is not plagued by intruder state problems.

3.1. Influence of Basis Sets. To assess the influence of a basis set on the MRMP2 procedures we also carried out calculations on the reaction set using the cc-pVDZ and cc-

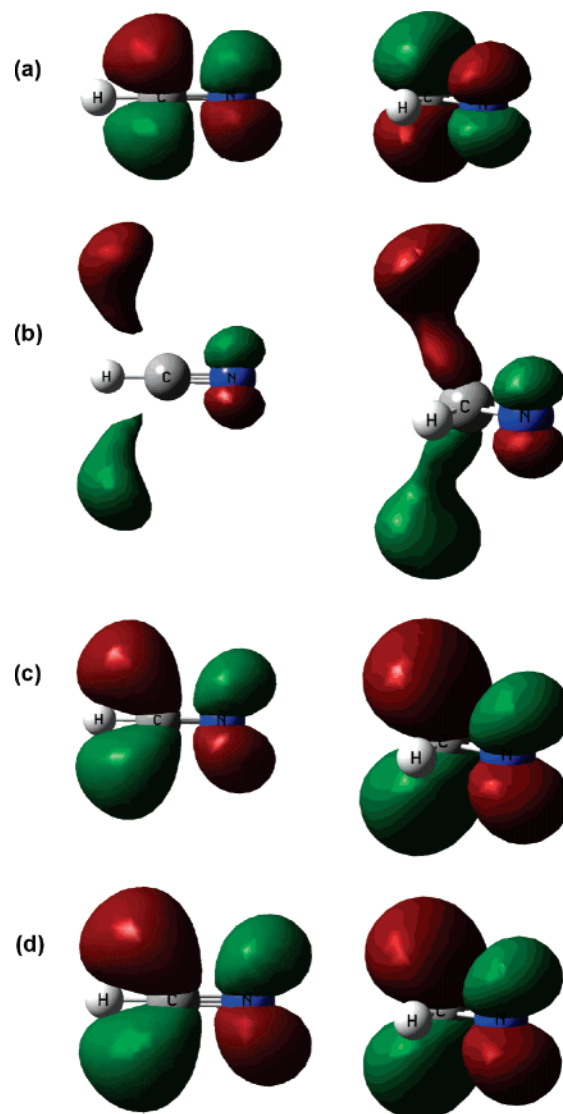


Figure 1. Canonical π^* orbitals of HCN (left-hand side) and the isomerization transition state (right-hand side) shown at an isosurface value of 0.05 au obtained by different methods: (a) CASSCF, (b) HF, (c) B3LYP, and (d) BLYP.

pVTZ basis sets of Dunning.^{28,29} The detailed results can be found in the Supporting Information.

For the CASSCF results, the mean absolute errors are 8.47 kcal mol⁻¹, 8.41 kcal mol⁻¹, and 8.15 kcal mol⁻¹ for the cc-pVDZ, cc-pVTZ, and MG3S bases, respectively. At the CASCI(HF) level, the mean absolute errors are 10.39 kcal mol⁻¹ (cc-pVDZ), 10.37 kcal mol⁻¹ (cc-pVTZ), and 20.08 kcal mol⁻¹ (MG3S). The large discrepancy between the correlation consistent bases and the MG3S must be attributed to the presence of diffuse functions in the latter, since the MG3S and cc-pVTZ bases are quite similar in other respects. Additionally, the MG3S basis includes *3d2f* polarization functions for chlorine (reaction 2), whereas the cc-pVTZ includes only *2d1f* polarization functions. This interpretation is also borne out by the HF orbitals shown in Figure 1(b). The KS orbitals do not show the same dependence on the presence of diffuse functions, and we find that for the CASCI(B3LYP) level the mean absolute errors are 10.09 kcal mol⁻¹ (cc-pVDZ), 10.25 kcal mol⁻¹ (cc-pVTZ), and 9.55

Table 5. Barrier Heights for Reactions 1–6 Obtained from MRMP2-ISA Calculations in Which the Reference CASCI Wave Function Is Built from HF, B3LYP, and BLYP Orbitals^a

reaction barrier		MRMP2- (HF)	€	MRMP2- (B3LYP)	€	MRMP2- (BLYP)	€
1	$V_{f,r}^{\text{cc}}$	42.47	0.29	41.11	-1.07	41.12	-1.06
2	$V_{f,r}^{\text{cc}}$	16.64	-1.36	17.36	-0.64	17.98	-0.02
3	V_f^{cc}	2.47	0.20	1.45	-0.82	0.57	-1.70
	V_r^{cc}	105.36	-0.82	103.76	-2.42	100.51	-5.67
4	V_f^{cc}	14.92	0.23	13.43	-1.26	13.57	-1.12
	V_r^{cc}	10.10	-0.62	10.98	0.26	12.02	1.30
5	V_f^{cc}	5.37	2.20	2.22	-0.95	2.51	-0.66
	V_r^{cc}	24.24	1.56	23.13	0.45	24.04	1.36
6	V_f^{cc}	52.99	4.83	51.51	3.35	51.03	2.87
	V_r^{cc}	35.38	2.27	34.99	1.88	35.02	1.91
$ \bar{\epsilon} $			1.44		1.31		1.77
max $ \epsilon $			4.83		3.35		5.67

^a Absolute values and errors (ϵ) are given in kcal mol⁻¹.

Table 6. Mean and Maximum Absolute Errors (kcal Mol⁻¹) in Barrier Heights for Reactions 1–6 Obtained with Different Types of Orbitals

method		orbitals used to build CASCI reference space			
		CASSCF	HF	B3LYP	BLYP
CASCI	$ \bar{\epsilon} $	8.15	20.08	9.55	9.37
	max $ \epsilon $	12.90	47.86	16.77	18.16
MRMP2	$ \bar{\epsilon} $	0.59	1.43	1.40	1.87
	max $ \epsilon $	1.99	4.79	3.39	5.83
MRMP2-ISA	$ \bar{\epsilon} $	0.53	1.44	1.31	1.77
	max $ \epsilon $	1.91	4.83	3.35	5.67

kcal mol⁻¹ (MG3S). The situation changes slightly at the CASCI(BLYP) level, see the Supporting Information.

For the MRMP2(CASSCF) level, the mean absolute errors are 2.56 kcal mol⁻¹ (cc-pVDZ), 1.05 kcal mol⁻¹ (cc-pVTZ), and 0.57 kcal mol⁻¹ (MG3S). The influence of the diffuse functions is much reduced at the MRMP2(HF) level, which gives mean absolute errors of 2.13 kcal mol⁻¹ (cc-pVDZ), 1.13 kcal mol⁻¹ (cc-pVTZ), and 1.43 kcal mol⁻¹ (MG3S). Finally, at the MRMP2(B3LYP) level we find errors of 2.00 kcal mol⁻¹ (cc-pVDZ), 1.24 kcal mol⁻¹ (cc-pVTZ), and 1.40 kcal mol⁻¹ (MG3S). We may conclude that for quantitative accuracy a large basis set (better than cc-pVDZ) is required.

4. Conclusions

This study has looked at the feasibility of using orbitals obtained from simpler methods than CASSCF optimization for building reference wave functions for multireference perturbation theory. However one must be conscious of the errors that can be introduced by adopting such a strategy, for example, the generalized Brillouin conditions of CASSCF theory are not satisfied, and some multireference perturbation theories (including MRMP2) are built on the assumption that these conditions are met. Table 6 summarizes our findings. We believe these results suggest that our strategy of avoiding the CASSCF step and using HF or B3LYP orbitals can be justified for many situations, though not for all. We are investigating other simplifications that may be applied to

multireference perturbation theory calculations. Ultimately we hope to see the methodology adopted for a much larger class of problems than has traditionally been the case. The elimination of the CASSCF step in the process should go a ways in achieving this aim.

Acknowledgment. We thank the University of Manchester for the provision of computer time on the Bezier high-performance computing facility and also the Engineering and Physical Sciences Research Council for the provision of computing equipment and the award of a studentship during which this work was carried out.

Supporting Information Available: Detailed results with the cc-pVDZ and cc-pVTZ basis sets. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Roos, B. O. *Adv. Chem. Phys.* **1987**, *69*, 399–446.
- (2) See e.g., Andersson, K.; Roos, B. O. In *Modern Electronic Structure Theory*; Yarkony, D. R., Ed.; World Scientific: 1995; pp 55–109.
- (3) See e.g., Werner, H.-J. *Adv. Chem. Phys.* **1987**, *69*, 1–62.
- (4) See e.g., Laidig, W. D.; Saxe, P.; Bartlett, R. J. *J. Chem. Phys.* **1987**, *86*, 887–907.
- (5) Andersson, K.; Malmqvist, P.-A.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. *J. Phys. Chem.* **1990**, *94*, 5483–5488.
- (6) Murphy, R. B.; Messmer, R. P. *Chem. Phys. Lett.* **1991**, *183*, 443–448.
- (7) Hirao, K. *Chem. Phys. Lett.* **1992**, *190*, 374–380.
- (8) Hirao, K. *Chem. Phys. Lett.* **1992**, *196*, 397–403.
- (9) Hirao, K. *Chem. Phys. Lett.* **1993**, *201*, 59–66.
- (10) Kozłowski, P. M.; Davidson, E. R. *J. Chem. Phys.* **1994**, *100*, 3672–3682.
- (11) Kozłowski, P. M.; Davidson, E. R. *Chem. Phys. Lett.* **1994**, *222*, 615–620.
- (12) Rogers, D. M.; Wells, C.; Joseph, M.; Boddington, V. J.; McDouall, J. J. W. *J. Mol. Struct. (Theochem)* **1998**, *434*, 239–245.
- (13) Choe, Y. -K.; Nakao, Y.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 621–629.
- (14) Potts, D. M.; Taylor, C. M.; Chaudhuri, R. K.; Freed, K. F. *J. Chem. Phys.* **2001**, *114*, 2592–2600.
- (15) Nakao, Y.; Choe, Y. -K.; Nakayama, K.; Hirao, K. *Mol. Phys.* **2002**, *100*, 729–745.
- (16) Hupp, T.; Engels, B.; Görling, J. A. *J. Chem. Phys.* **2003**, *119*, 11591–11601.
- (17) Robinson, D.; McDouall, J. J. W. *Mol. Phys.* **2006**, *104*, 681–690.
- (18) Zhao, Y.; González-García, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012–2018.
- (19) Lynch, B. J.; Zhao Y.; Truhlar, D. G. <http://comp.chem.umn.edu/database> (accessed April 11, 2007).
- (20) Choe, Y. -K.; Witek, H. A.; Finley, J. P.; Hirao, K. *J. Chem. Phys.* **2001**, *114*, 3913–3918.
- (21) Choe, Y. -K.; Witek, H. A.; Finley, J. P.; Hirao, K. *J. Comput. Chem.* **2002**, *23*, 957–965.

- (22) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265–3269.
- (23) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764–7776.
- (24) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1999**, *110*, 4703–4709.
- (25) Rintelman, J. M.; Adamovic, I.; Varganov, S.; Gordon, M. S. *J. Chem. Phys.* **2005**, *122*, 44105:1–44105:7.
- (26) Szabados, Á.; Rolik, Z.; Tóth, G.; Surján, P. *J. Chem. Phys.* **2005**, *122*, 114104:1–114104:12.
- (27) *Gaussian 03, Revision C.02*; Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, M. G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian, Inc.: Wallingford, CT, 2004*.
- (28) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (29) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358–1371.

CT700037Z

A Novel Approach to Parallel Coupled Cluster Calculations: Combining Distributed and Shared Memory Techniques for Modern Cluster Based Systems

Ryan M. Olson,[†] Jonathan L. Bentz,[‡] Ricky A. Kendall,^{‡,§} Michael W. Schmidt,[†] and Mark S. Gordon^{*,†}

Department of Chemistry, Iowa State University, Ames, Iowa, Department of Computer Science, Iowa State University, Ames, Iowa 50011, and Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831

Received January 17, 2007

Abstract: A parallel coupled cluster algorithm that combines distributed and shared memory techniques for the CCSD(T) method (singles + doubles with perturbative triples) is described. The implementation of the massively parallel CCSD(T) algorithm uses a hybrid molecular and “direct” atomic integral driven approach. Shared memory is used to minimize redundant replicated storage per compute process. The algorithm is targeted at modern cluster based architectures that are comprised of multiprocessor nodes connected by a dedicated communication network. Parallelism is achieved on two levels: parallelism within a compute node via shared memory parallel techniques and parallelism between nodes using distributed memory techniques. The new parallel implementation is designed to allow for the routine evaluation of mid- (500–750 basis function) to large-scale (750–1000 basis function) CCSD(T) energies. Sample calculations are performed on five low-lying isomers of water hexamer using the aug-cc-pVTZ basis set.

I. Introduction

Coupled-cluster (CC) methods^{1–3} are now widely accepted as the premier single-reference electronic structure methods for small chemical systems at or near equilibrium geometries. One of the most popular CC methods is CCSD(T), which is based on an iterative solution of the single and double (SD) cluster amplitude equations⁴ with a noniterative perturbative correction for the triples (T).⁵ The CCSD(T) approach has been shown⁶ to be a good compromise between the chemical accuracy of the higher-order CCSDT (full triples) method⁷ and the computational efficiency of low order many-body perturbation theory (MBPT). Equation of motion (EOM) CC methods^{8–12} have been developed for excited-state calculations. Spin flip^{13,14} and method of moments CC methods,¹⁵ including the popular renormalized (R),¹⁵ completely renormalized (CR),¹⁵ and CR-CCSD(T)_L (CCL) methods,¹⁶ have

extended formally single-reference CC methods into the regime of bond making and bond breaking, an area where traditional CC methods break down.

The biggest drawback of CC methods is the large computational demands required to perform such calculations. However, due to the popularity of methods like CCSD(T), considerable research has been carried out to generate highly efficient algorithms^{4,17–21} and their implementations. A variety of CC methods can be found in all of the major electronic structure programs available today, including GAMESS,²² MOLPRO,²³ ACES II,²⁴ Q-CHEM,²⁵ PSI3,²⁶ NWCHEM,^{27,28} DALTON,²⁹ and GAUSSIAN03.³⁰ Most CC programs are highly optimized to run sequentially. This usually means the calculation is performed on a single processor. The speed of the processor and the size of the associated memory and disk are limiting factors for sequential algorithms. CCSD(T) calculations, especially those run in C_1 symmetry, reach the limit of most single processor workstations at around 400–500 basis functions (BF); even then, calculations of these sizes may require weeks of time on a dedicated workstation.³¹

* Corresponding author e-mail: mark@si.fi.ameslab.gov.

[†] Department of Chemistry, Iowa State University.

[‡] Department of Computer Science, Iowa State University.

[§] Oak Ridge National Laboratory.

One means of evaluating computationally demanding problems such as large basis set (>500 BF) CCSD(T) calculations is to make use of parallel computing. Parallel computing involves simultaneously evaluating multiple portions of a larger computational problem on multiple processors, in order to achieve an overall reduction in the real-time evaluation of the problem. Equally important, parallel computing can extend computationally demanding methods like CCSD(T) to larger problems because of increased computational resources and also storage (memory/disk) resources. There is a wide range of parallel computing environments and methodologies, two examples of which are addressed herein. These are as follows: (1) parallelism that is achieved by using multiple computers or *nodes* which are connected by a dedicated communication network and (2) parallelism that is achieved by multiple processors within a single *node* that share “local” system resources including memory and I/O channels.

The tools and methodologies for these two traditional types of parallel computing environments are very different. Multinode parallelism focuses on combining replicated and/or distributed memory techniques using parallel communication libraries such as TCGMSG,³² SHMEM,³³ MPI,³⁴ Global Arrays (GA),³⁵ and the Distributed Data Interface (DDI).^{36,37} One advantage of multinode models is that the aggregate system resources increase as the number of nodes increases, thereby facilitating more resource demanding calculations. However, since the nodes are distinct and internode communication must travel over a high-speed network, there are three factors that will strongly affect the performance for these types of calculations: (1) the performance (bandwidth and latency) of the network, (2) the total amount of internode communication required, and (3) the degree to which the necessary communication can be overlapped with computation. Single node multiprocessor parallel schemes have traditionally focused on a relatively small number of compute processes (or threads), usually between 2 and 16, using shared resources as a means to reduce (1) message passing communication and (2) replicated storage overhead, i.e., using the shared resources of the system to store certain data arrays only one time, rather than stored multiple identical copies for each process (or thread). A major focus of these techniques involves sharing portions of the system memory among all parallel processes (or threads) and providing tools to control access to this shared data. Examples of shared-memory based programming models include the POSIX Pthreads model, the OpenMP model,³⁸ and the System V interprocess communication model.

In general, the multinode and single node parallel strategies were developed separately based on two different types of parallel architectures. However, it is the evolution of the *node*, specifically the use of multiprocessor “shared-memory” nodes as the building blocks for multinode cluster based systems, which is bringing about a convergence of these methodologies. That is, it is possible to embed the use of shared-memory programming techniques within each node of a cluster based system yet retain the advantages of increased aggregate system resources from a multinode platform. This becomes especially important when examining

the roadmap for future generations of computers. The next generation(s) of processors is(are) not expected to dramatically increase in frequency, which traditionally has accounted for 80% of the performance improvements. Rather, the current trend is to add multiple processing “cores” on each processor. This use of multicore processors in multiprocessor nodes further increases the computational density per node and further emphasizes the need to address different parallel strategies for intra- and internode computing and data management within current and future cluster based systems.

The focus of this work is to describe an algorithm for the CCSD(T) method that can utilize both intranode and internode forms of parallelism. Algorithms for parallel CC methods^{39–43} have been developed by other groups. These methodologies for the parallelization of CC methods and other correlation methods were divided into two categories: those aimed at shared memory machines (SMPs) and those aimed at distributed memory machines. Early work by Komornicki, Lee, and Rendell³⁹ described a highly vectorized shared memory algorithm for evaluating the connected triples excitations (T) on the CRAY Y-MP. Vectorized shared-memory CCSD and CCSD(T) algorithms based on AO integrals stored on disk were later implemented by Koch and co-workers.^{44,45} These early shared-memory vectorized algorithms primarily used optimized library calls to gain computational speedup (the libraries, not the programs themselves, were multiprocess or multithread based), although some directives to parallelize the loops were employed. Rendell, Lee, and Lindh⁴⁰ implemented the first distributed memory CCSD algorithm on an Intel i860 hypercube. In that work, asymptotic speedups were quickly reached due to I/O bottlenecks based on retrieval of the molecular integrals. The authors proposed the use of a “semidirect” method in which atomic integrals evaluated “on demand” could be used to alleviate the I/O bottleneck. Rendell, Guest, and Kendall⁴¹ improved the previous MO-based distributed memory CCSD approach and extended the program to include CCSD(T). Later, Kobayashi and Rendell⁴² implemented a “direct” AO-driven CCSD(T) algorithm which avoided the I/O bottlenecks of earlier MO-based distributed memory methods; this development formed the basis for the parallel CCSD(T) module within the NWChem package.²⁷ As another means of avoiding potential I/O bottlenecks, Rendell and Lee proposed⁴⁶ that some two-electron integrals can be approximated using the resolution of the identity (RI) technique. RI-based approaches can dramatically reduce the storage requirements needed for CCSD and CCSD(T) calculations, while maintaining $O(N^6)$ and $O(N^7)$ scaling for the computational effort where N is a description of the size of the system being calculated; the number of atomic basis functions is an upper limit to N . MOLPRO²³ also offers a parallel implementation of its coupled cluster methods. Most recently, Janowski and co-workers⁴³ have presented a parallel algorithm for the CCSD method using the Array Files toolkit.⁴⁷

Another exciting advance in the development of parallel computer codes for high level ab initio quantum chemistry methods is the tensor contraction engine (TCE),⁴⁸ a program used for automatic code generation for a general set of high

level ab initio methods, including coupled cluster methods. Hirata⁴⁹ has shown the utility of the TCE for deriving and implementing many common second-quantized many-electron theories including a variety of coupled cluster methods. The TCE also has the ability to automatically generate *parallel* computer codes. A recent study by Piecuch and co-workers⁵⁰ used the TCE to generate a parallel code for the completely renormalized CCSD(T) method¹⁵ which showed that a ten times execution speedup could be achieved using 64 processors. As illustrated by this example, parallel codes generated by the TCE are generally not as efficient as hand-tuned computer codes; however, the major benefits of using the TCE are its ease of use, the avoidance of errors in generating very complex codes, and its general applicability to higher order ab initio methods in which detailed hand tuning and parallelization can be very difficult. The contributions from a number of researchers⁵¹ to the improvement of the generation of highly efficient parallel codes via the TCE program has extraordinary potential and could someday result in automatically generated code that is as good as or better than hand-tuned programs.

The major purpose of this paper is to describe a new parallel CCSD(T) implementation that seeks to find the best balance between the $O(N^7)$ computational cost and the $O(N^4)$ data storage requirements of CCSD(T). The algorithm described here is targeted toward today's basic computer building block: a node with several processor cores and also an appreciable total memory within the node (e.g., $p = 4$ processors and 8 GB of RAM or more). The algorithm also eliminates disk usage while seeking to minimize communications costs. The unique feature of the algorithm presented here is combining the use of both distributed memory (internode) and shared memory (intranode) techniques in a massively parallel (MP) program. Clearly, any MP-CCSD(T) algorithm requires tradeoffs be made in each of these areas, so the proof of the algorithm's viability necessarily must be to demonstrate its ability to do large CCSD(T) computations on realistic hardware, in a reasonable amount of time. The MP-CCSD(T) algorithm described here is an adaptation of the sequential algorithm, previously implemented in GAMESS²² by Piecuch et al.⁵² Because the MP-CCSD(T) method described here is based on the same spin-free equations used by the EOM and renormalized CC methods in GAMESS, the approach to closed shell CCSD(T) parallelism described here can be extended to the other types of coupled cluster methods in a straightforward manner.

To provide an example of the viability of the MP-CCSD(T) algorithm on modern cluster based architectures, CCSD(T) calculations on geometric isomers of water hexamer using the aug-cc-pVTZ basis set⁵³ are presented. These calculations are important, since there are five low lying isomers of water hexamer (Figure 1), some of which have three-dimensional structures, whose relative energies are very close to each other. Since these are the smallest 3-D water clusters, it is very important to be able to predict the correct energy order for the low-energy isomers with high accuracy. This means that one needs both large basis sets that approach the complete basis set limit, in order to avoid basis set superposition error (BSSE), and a high theoretical level, such

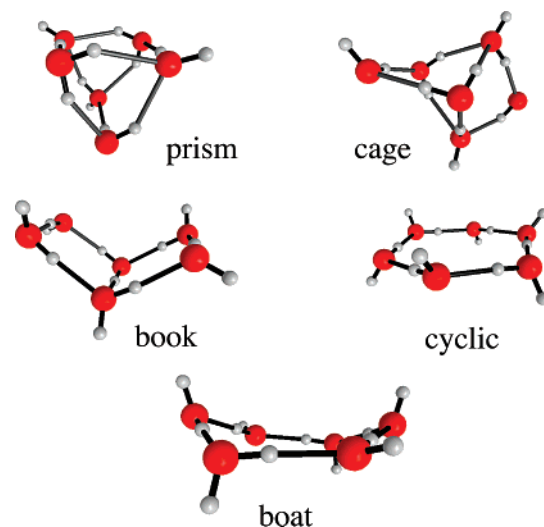


Figure 1. Images of the five geometric isomers used in this study. The geometries correspond to MP2 optimized structures using the DH(d,p) basis set obtained by Day et al.⁴⁴

as CCSD(T). A number of high-level ab initio studies^{40–44} have been performed on the water hexamer. In a very thorough and systematic study of the potential energy surface for small water clusters using second-order Moller–Plesset perturbation theory⁵⁴ (MP2) and a series of augmented correlation consistent basis sets^{53,55} that are systematically improved (aug-cc-pVXZ ranging from X=D,T,Q,5) Xanthreas and co-workers⁵⁶ have predicted that the prism structure is the global minimum. However, the predicted energy differences among the water hexamer isomers are very small (a range of less than 1.2 kcal/mol for the four isomers studied). Given the known tendency of MP2 to overbind clusters, it is important to employ a more sophisticated level of theory, e.g., CCSD(T), with a sufficiently large basis set such that BSSE approaches zero.⁵⁷ The calculations performed herein represent, to the authors' knowledge, the most accurate CCSD(T) calculations on water hexamer to date.

This paper highlights the key features of the MP-CCSD(T) program and demonstrates that the algorithm is viable on modern cluster based MP platforms. The goal of the MP-CCSD(T) algorithm is to enable high-level CC calculations to provide accurate energies and potential energy surfaces for systems, like water hexamer, that are currently very difficult to achieve. As an illustration of the new method, the CCSD(T)/aug-cc-pVTZ energies for the five low-lying water hexamer isomers are calculated and the performance of the MP-CCSD(T) method is examined. Since the primary focus of the present work is on the MP-CCSD(T) algorithm, the issues of extrapolation to the complete basis set limit and basis set superposition error are deferred to a later paper.

II. CCSD/CCSD(T) Theory

The MP-CCSD(T) method described in this work is an adaptation of the sequential CCSD(T) program previously implemented by Piecuch et al.,⁵² therefore, the same notation used in ref 52 is followed here. The letters i, j, k, l, \dots will be used to denote occupied spatial molecular orbitals, a, b, c, d, \dots will be used to represent unoccupied (virtual) orbitals, $\mu, \nu, \lambda, \sigma, \dots$ are used to represent atomic orbital indices or

atomic shell indices, and p, q, r, s, \dots are a general set of indices. Details of the CCSD and perturbative (T) correction have been discussed in several reviews,^{58,59} so only a brief outline is given here.

The CCSD method derives from CC theory in the following manner. Let the exact CC wave function ($|\Psi_{CC}\rangle$) be defined as

$$|\Psi_{CC}\rangle = e^T |\Phi\rangle \quad (1)$$

where $|\Phi\rangle$ is the reference wave function (for this work, $|\Phi\rangle$ is the restricted closed-shell Hatree-Fock reference wave function), and T is the complete cluster operator containing all possible single (T_1), double (T_2), triple (T_3), etc. excitation operators

$$T = T_1 + T_2 + T_3 + \dots \quad (2)$$

The CCSD method results from the truncation of T such that only single and double excitation operations are included

$$T \approx T_1 + T_2 \quad (3)$$

Projecting the connected-cluster form of the CCSD equation

$$(H_N e^{T_1+T_2})_C |\Phi\rangle = E_{CCSD} |\Phi\rangle \quad (4)$$

where H_N is the normal product electronic Hamiltonian ($H - \langle \Phi | H | \Phi \rangle$), onto the set of excited determinants defined by the truncated excitation operator (eq 3) gives rise to a set of coupled nonlinear equations,

$$\langle \Phi_i^a | (H_N e^{T_1+T_2})_C |\Phi\rangle = 0 \quad (5)$$

$$\langle \Phi_{ij}^{ab} | (H_N e^{T_1+T_2})_C |\Phi\rangle = 0 \quad (6)$$

eqs 5 and 6, which are solved iteratively for the single and double excitations, respectively. In terms of amplitudes (t_i^a , t_{ij}^{ab}), Fock matrix elements (f_p^q), and two-electron molecular integrals ($v_{rs}^{pq} = \langle pq | (1/r_{12}) | rs \rangle$), respectively, the CCSD amplitude equations (eqs 5 and 6) are given (using the Einstein summation convention). [The Einstein summation convention implies a summation over all possible values of repeated indexes found in the lower or upper positions of a single term. For example, $t_{ei}^e = \sum_e t_{ei}^e$.]

$$D_{it_i^a}^a = f_i^a + t_{ei}^e + t_{im}^m + t_{im}^m (2t_{mi}^{ea} - t_{im}^{ea}) + (2v_{ei}^{ma} - v_{ei}^{am})t_m^e - v_{ei}^{mn} (2t_{mn}^{ea} - t_{mn}^{ae}) + v_{ef}^{ma} (2t_{mi}^{ef} - t_{im}^{ef}) \quad (7)$$

$$D_{ij}^{ab} t_{ij}^{ab} = v_{ij}^{ab} + P(ij/ab) \left[t_{ij}^e t_{ij}^e - t_{im}^m t_{ij}^m + \frac{1}{2} v_{ef}^{ab} c_{ij}^{ef} + \frac{1}{2} c_{mn}^{ab} t_{ij}^{mn} - t_{mj}^{ae} - t_{ie}^{ma} t_{mj}^{eb} + (2t_{mi}^{ea} - t_{im}^{ea}) t_{ej}^{mb} + t_i^e t_{ej}^{ab} - t_m^a t_{ij}^{mb} \right] \quad (8)$$

In eqs 7 and 8, c_{ij}^{ab} is defined as

$$c_{ij}^{ab} = t_{ij}^{ab} + t_i^a t_j^b \quad (9)$$

The permutation operator $P(ij/ab)$ acting on an arbitrary term (X) has the following properties

$$P(ij/ab) X_{ij}^{ab} = X_{ij}^{ab} + X_{ji}^{ba} \quad (10)$$

and the general MBPT denominators are used to define D_i^a , D_{ij}^{ab} , and D_{ijk}^{abc} such that

$$D_{ij\dots}^{ab\dots} = \epsilon_i - \epsilon_a + \epsilon_j - \epsilon_b \dots \quad (11)$$

where

$$\epsilon_p = f_p^p \quad (12)$$

are the diagonal elements of the Fock matrix. The intermediates (I and I') of eqs 7 and 8 are

$$I_a^i = f_i^a + 2v_{ae}^{im} t_m^e - v_{ea}^{im} t_m^e \quad (13)$$

$$I_b^a = (1 - \delta_b^a) f_b^a + 2v_{be}^{am} t_m^e - v_{be}^{ma} t_m^e - 2v_{mn}^{ea} c_{mn}^{ea} + v_{be}^{mn} c_{mn}^{ea} - t_m^a t_m^b \quad (14)$$

$$I_j^i = (1 - \delta_j^i) f_j^i + 2v_{je}^{im} t_m^e - v_{ej}^{im} t_m^e + v_{ef}^{mi} t_{mj}^{ef} - v_{ef}^{im} t_{mj}^{ef} \quad (15)$$

$$I_j^i = I_j^i + I_{e_j}^e \quad (16)$$

$$I_{kl}^{ij} = v_{kl}^{ij} + v_{ef}^{ij} t_{kl}^{ef} + P(ik/jl) t_e^k v_{el}^{ij} \quad (17)$$

$$I_{jb}^{ia} = v_{jb}^{ia} - \frac{1}{2} v_{eb}^{jm} t_{jm}^{ea} - v_{jb}^{im} t_m^a + v_{eb}^{ja} t_m^e \quad (18)$$

$$I_{bj}^{ia} = v_{bj}^{ia} + v_{be}^{im} t_{mj}^{ea} - \frac{1}{2} v_{mb}^{je} t_{jm}^{ae} - \frac{1}{2} v_{be}^{im} t_{mj}^{ae} + v_{be}^{ja} t_m^e - v_{bj}^{im} t_m^a \quad (19)$$

$$I_{ci}^{ab} = v_{ci}^{ab} - v_{ci}^{am} t_m^b - t_m^a v_{ci}^{mb} \quad (20)$$

$$I_{jk}^{ia} = v_{jk}^{ia} + v_{ef}^{ia} t_{jk}^{ef} + t_j^e v_{be}^{ia} + t_j^e v_{ef}^{ia} t_k^f \quad (21)$$

where δ_p^q represents the standard Kronecker delta.

The CCSD correlation energy from the CCSD method is calculated after eqs 7 and 8 are solved iteratively for t_i^a and t_{ij}^{ab} and is given by the following formula

$$\Delta E^{CCSD} = 2f_{at_i^a}^i + (2v_{ab}^{ij} - v_{ba}^{ij}) c_{ij}^{ab} \quad (22)$$

Noniterative solutions to the full CCSDT problem using only lower order excitation operators (T_1 and/or T_2) were first developed by Urban and co-workers.⁶⁰ These methods eventually led to the CCSD(T) method derived by Raghavachari and co-workers.⁵ The (T) of CCSD(T) is an *a posteriori* noniterative correction to the CCSD energy. In a study analyzing a variety of different approximations to the full CCSDT treatment, Scuseria and Lee⁶ found the CCSD(T) method to be the most accurate and the most computationally efficient of all the approximate methods examined. In terms of molecular integrals and amplitudes,⁵² the correction to the CCSD energy is given by

$$E^{(T)} = \bar{t}_{abc}^{ijk} (2) t_{ijk}^{abc} (2) D_{ijk}^{abc} + \bar{z}_{abc}^{ijk} (2) D_{ijk}^{abc} \quad (23)$$

where an arbitrary \bar{X}_{abc}^{ijk} term is expanded such that

$$\bar{X}_{abc}^{ijk} = \frac{4}{3} X_{abc}^{ijk} - 2X_{acb}^{ijk} + \frac{2}{3} X_{bca}^{ijk} \quad (24)$$

The $t_{ijk}^{abc} (2)$ coefficients are defined in terms of t_i^a and t_{ij}^{ab}

$$D_{ijk}^{abc} t_{ijk}^{abc} (2) = P(ia/jb/kc) [t_{ij}^{ae} v_{ek}^{bc} - t_{im}^{ab} v_{jk}^{mc}] \quad (25)$$

where the permutation operator $P(ia/jb/kc)$ expands a quantity containing the (ia) , (jb) , and/or (kc) pair into a summation of up to six quantities:

$$P(ia/jb/kc)[\dots]_{ijk}^{abc} = [\dots]_{ijk}^{abc} + [\dots]_{ikj}^{acb} + [\dots]_{kij}^{cab} + [\dots]_{kji}^{cba} + [\dots]_{jki}^{bca} + [\dots]_{jik}^{bac} \quad (26)$$

The second term of eq 23 is the disconnected triples correction to $E^{(T)}$ where

$$z_{abc}^{ijk} \equiv (z_{ijk}^{abc})^* = (t_a^i v_{bc}^{jk} + t_c^k v_{ab}^{ij} + t_b^j v_{ac}^{ik})/D_{ijk}^{abc} \quad (27)$$

and the z_{abc}^{ijk} and z_{ijk}^{abc} are complex conjugates. The final CCSD(T) energy is given by

$$E^{\text{CCSD(T)}} = \langle \Phi | H | \Phi \rangle + \Delta E^{\text{CCSD}} + E^{(T)} \quad (28)$$

A detailed discussion of the individual terms in the CCSD and (T) equations is presented in section 3 of ref 52. The summary presented in eqs 1–28 provides a sufficient background for the following discussion of the implementation of the MP-CCSD(T) method.

III. Parallel Design

There are two primary issues that must be considered in order to perform large-scale CCSD(T) calculations in a massively parallel environment: How can the computational workload be divided among the available parallel processes? How can the large data sets associated with such demanding calculations be stored and utilized efficiently by the available parallel processes?

The amount of computational effort associated with the CCSD and CCSD(T) algorithms scales asymptotically as $O(N^6)$ and $O(N^7)$, respectively. N is a measure of system size and can be broken down more specifically in terms of the number of occupied molecular orbitals (N_o) and the number of unoccupied (virtual) molecular orbitals (N_v). More generally (and more conservatively), one can use the number of one-electron atomic basis functions (N_{bf}). In terms of molecular orbitals, the CCSD and CCSD(T) algorithms scale on the order of their most expensive terms, $O(N_o^2 N_v^4)$ and $O(N_o^3 N_v^4)$, respectively. Each of the terms in the sequential code⁵² was parallelized, with specific attention paid to the terms which comprise the *computational bottlenecks*. However, the distribution of the computational work is very closely related to the distribution of the large data sets required by the CCSD(T) method. Therefore, before detailed examples of the manner in which the terms of the CCSD(T) method were parallelized, an examination of data distribution is required.

The second major consideration addresses the storage requirements for large CCSD(T) calculations in a massively parallel environment. As mentioned in section II, the CCSD(T) equations are written in terms of cluster amplitudes and molecular (or atomic) integrals. The manner in which the integrals and amplitudes are stored on a large parallel computer has a direct effect on how the computational workload can be distributed. Equally important, the choice of how the amplitudes and integrals are stored will directly affect the *storage bottlenecks* of the algorithm.

The MP-CCSD(T) algorithm was designed to address these bottlenecks by first examining the data storage problem and then addressing the parallel work division based on a defined data distribution. In the following discussion, the storage bottlenecks are examined in the scope of the programming model and the available types of storage. Based on these ideas and an outlined storage model, section IV describes how the computational work is divided into internode and intranode components.

A. Parallel Programming Model. The MP-CCSD(T) algorithm introduces and utilizes the third generation of the Distributed Data Interface (DDI) for communication and data storage in a massively parallel environment. The DDI model is a high-level abstraction of the virtual shared-memory model for use in the GAMESS quantum chemistry suite of programs. DDI was designed as a means to provide a consistent set of parallel programming tools for the quantum chemistry code, while maintaining enough generality to be implemented using a variety of existing parallel libraries that offer one-sided message passing, including the following: SHMEM, Global Arrays (GA),³⁵ MPI,³⁴ and a native implementation based on point-to-point libraries such as MPI³⁴ and/or TCP/IP sockets. The DDI model was strongly influenced by the structure and functionality of the Global Arrays (GA) Toolkit; however, to maintain a high degree of portability only a subset of the GA functionality is used within the DDI model.

The first generation of DDI,^{37,61} DDI/1, provided a process-based implementation of the distributed-memory programming model in which large arrays could be evenly divided over all available nodes yet remain globally accessible via one-sided message operations. DDI/1 was modeled on the design of the Cray T3E in which the system image of each *node* contained a single processor and some associated system memory. The nodes formed the building blocks of the parallel computer and were connected to other nodes by a high-speed network. DDI/1 is a process-based model, because the data and the computational workload are divided over the parallel processes.

The second generation of DDI,³⁶ DDI/2, introduced a greater awareness of the memory topology by recognizing that multiple parallel processes could coexist within the same node, i.e., multiple processors in a single node sharing the same local system memory. This shared-memory awareness increases the amount of data that can be considered “local” and can significantly reduce the number of remote communication operations for calculations run using multiprocessor nodes; this was recognized for point-to-point communication in many MPI implementations and also in the one-sided communications for both GA³⁵ and DDI.³⁶

The third generation of DDI, DDI/3, further enhances the shared-memory capabilities of DDI by providing the tools needed for multiprocessor nodes to utilize shared-memory outside of the distributed-memory model. Specifically, DDI/3 provides the ability to create and control access to shared-memory segments as well as the ability to perform point-to-point and collective operations within the node. The shared-memory model in DDI/3 is based on multiple processes using SystemV shared-memory and semaphores

for interprocess communication rather than a thread-based model. This maintains the integrity of the former DDI models, whereas a shift to a thread-based model for intranode parallelism would require a radical change to the DDI programming model. DDI/3 provides all the necessary tools for process-based *and* node-based parallelism.

Node-based parallelism differs from process-based parallelism in that the data and the computational work are first divided by node (internode), and then the work assigned to each node is further decomposed and parallelized over the “local” processes within each node. Node-based parallel schemes have the advantage of being able to handle larger replicated data sets when compared to process-based schemes, because shared-memory can be used to store particular quantities *once per node*, rather than *once per process*. The MP-CCSD(T) algorithm described here utilizes both process-based and node-based parallel techniques.

B. Memory. DDI/3 supports three types of memory storage to be used in the MP-CCSD(T) algorithm: replicated, shared, and distributed. Replicated memory is process-based, and the amount of memory needed for data stored in replicated memory scales linearly with the number of processes. Typically, arrays that scale as $O(N^2)$ and some that scale as $O(N^3)$ can be stored in replicated-memory. Shared memory is node-based, and the amount of memory needed for data stored in shared memory scales linearly with the number of nodes. Shared-memory allows for the storage of larger arrays than does replicated-memory, since the arrays are only stored *once per node*. In a shared-memory environment, every process within the node can access and modify the data in shared-memory segments. This feature provides a convenient means of parallelizing the computational work over a shared data set, since each process has direct access to the data in that memory (by physical address). However, allowing multiple processes to have access to shared resources means that special care must be taken to prevent possible race conditions, i.e., situations that occur during parallel execution in which one process seeks to modify data that are concurrently being used by another process. To handle these race conditions, DDI/3 uses SystemV semaphores and collective synchronizations over all intranode processes to control access to shared resources and guarantee data integrity.

Distributed memory is the aggregate of portions of “local” system memory reserved by each process for the storage of distributed data. In the DDI framework, the number of columns of a distributed two-dimensional matrix is divided evenly over the total number of parallel processes; the disjoint sets of columns are mapped in a one-to-one manner onto the set of parallel processes, and the data associated with each set of columns are stored in the memory reserved by each process for distributed memory storage. In contrast to shared memory, access to distributed memory requires calling subroutines from the DDI library. The amount of distributed-memory needed for a given calculation is defined solely by the parameters of the calculation and has *no* dependence on the number of parallel processes used for the calculation. The requirements for distributed memory can in some cases be very large; in those cases, the number of

nodes must be chosen to accommodate the required distributed memory.

There are two types of distributed-memory: local and remote. All parallel processes are allowed to modify any element of an array stored in distributed-memory (regardless of physical location); however, due to the communication overhead of accessing remote distributed-memory, the programming strategy seeks to maximize the use of local distributed-memory and minimize the use of remote distributed-memory. In this regard, arrays stored in distributed-memory are not easily rearranged between distributed indexes. For example, when a transpose operation, i.e., the swapping of the rows and columns, is performed on a distributed matrix that is distributed evenly over the number of columns, every parallel process must communicate with all of the other parallel processes. Thus, for very large distributed matrices, this type of operation would require a large amount of communication overhead and would be an impediment to achieving good parallel speedups.

C. Molecular Integral Transformation. The MO integral classes use an “O” to denote an actively correlated occupied MO index and a “V” to denote a virtual MO index. In the present work, a modified version of the distributed-memory “direct” four-index integral transformation⁶² previously implemented by Fletcher and co-workers⁶¹ was used to calculate the MO integrals: [OO|OO], [VO|OO], [VV|OO], [VO|VO], and [VV|VO]. The original integral transformation was only able to calculate MO integrals with up to two virtual indexes and was not able to exclude frozen-core MOs from the transformation for a general set of MO integral classes. Modifications were therefore made to allow for the formation of [VV|VO] integrals and to add an option to include or exclude frozen core MOs in the transformation. These modifications maintain the integrity of the original algorithm, i.e., the identical procedures are used; however, the starting indexes and ranges of MO indexes that are transformed were modified.

The formation of the [VV|VO] integrals requires an additional distributed array to store the [NN|VO] integrals, where “N” is the total number of basis functions and the entries can be V or O. The same procedure that is used to complete the [VV|OO] integrals from the [NN|OO] set of half-transformed integrals is used to complete the [VV|VO] integrals from the [NN|VO] set of half-transformed integrals. Exclusion of the frozen-core integrals is accomplished using a straightforward modification of the starting index and the range of MO orbitals that are defined as occupied (active). If one wishes to freeze the core molecular orbitals in the coupled cluster calculation, those core molecular orbitals are not correlated, and, therefore, the MO integrals associated with the frozen-core MOs are not required. The option to exclude the frozen-core integrals can result in a significant reduction in computational effort and most importantly a reduction in the distributed-memory requirements for the integral transformation. Of course, for heavier elements such as Au, one must take care in defining those orbitals that are frozen, in order to avoid excluding orbitals that are important in the chemical process of interest.⁶³

Table 1. General List of Data Types That Describes What Type of Memory the Quantity Will Be Stored in and How the Quantity Scales as a Function of N_o and N_v

class	type	size	storage
$T_1 (t_{ij}^a)$	amplitudes	$O(N_o N_v)$	replicated
$T_2 (t_{ij}^{ab})$	amplitudes	$O(N_o^2 N_v^2)$	shared
[OO OO]	integrals	$O(N_o^4)$	distributed
[VO OO]	integrals	$O(N_o^3 N_v)$	distributed
[VV OO]	integrals	$O(N_o^2 N_v^2)$	distributed
[VV VO]	integrals	$O(N_o N_v^3)$	distributed
[VV VV]	integrals	$O(N_v^4)$	not stored

D. Memory Requirements and Bottlenecks. The scaling of the storage requirements and how the data are stored within the MP-CCSD(T) algorithm are given in Table 1 in terms of the number of actively correlated occupied (N_o) and the number of virtual (N_v) molecular orbitals (MO). In the following discussion, the memory requirements and the potential memory bottlenecks are examined over the range of $10 \leq N_o \leq 60$ and $300 \leq N_v \leq 1000$.

For midrange to high-end dedicated supercomputers, the assumption is made that 4–8 GB (GB = 2^{30} bytes) of memory per processor are available. For common 4–8 processor nodes, this means that typically 16–64 GB of “local” system memory per node is generally available. For low-end commodity clusters, these assumptions would not necessarily hold at present; however, it is assumed here that sufficient high-performance computer facilities are available.

Another working assumption is that access to quality disk storage, i.e., “local” multichanneled striped disk arrays on every node, is not generally available. This is a conservative approach to minimize the performance dependence of the algorithm on the quality of the available disk I/O, which can vary greatly from cluster to cluster. In fact, some clusters do not even have local scratch disk storage, and the only available file system may be a remote networked file server or a parallel file system such as Lustre or PVFS2. The performance of the algorithm might be improved if one could assume that quality local disk storage per node is available. In this initial implementation of the MP-CCSD(T) algorithm, only minimal system requirements are assumed.

There are two storage bottlenecks in the MP-CCSD(T) algorithm as defined by the choice of data storage (Table 1). These are the storage of the $T_2 (t_{ij}^{ab})$ in shared-memory and the storage of the [VV|VO] molecular integrals in distributed memory.

The storage of the $T_2 (t_{ij}^{ab})$ amplitudes in shared-memory is the first of two storage bottlenecks within the MP-CCSD(T) algorithm. The T_2 amplitudes require $N_o^2 N_v^2$ words of shared memory; however, two other intermediates of the same size must also be stored in shared memory. The actual size of the T_2 amplitudes measured in gigabytes (GB) is given in Table 2 (see Table 1 for a summary of all integral and amplitude types). The use of shared memory to store the T_2 amplitudes represents a compromise for the efficient use of the amplitudes, since the T_2 amplitudes are too large (in most cases > 1 GB) to be stored in replicated-memory, and these T_2 amplitudes are reordered and manipulated too frequently

Table 2. Maximum Size in Gigabytes (GB) of the Array To Hold the T_2 Amplitudes, the [VV|OO] Integrals, or the [VO|VO] MO Integrals^a

	300	400	500	600	700	800	900	1000
10	0.1	0.1	0.2	0.3	0.4	0.5	0.6	0.7
15	0.2	0.3	0.4	0.6	0.8	1.1	1.4	1.7
20	0.3	0.5	0.7	1.1	1.5	1.9	2.4	3.0
25	0.4	0.7	1.2	1.7	2.3	3.0	3.8	4.7
30	0.6	1.1	1.7	2.4	3.3	4.3	5.4	6.7
35	0.8	1.5	2.3	3.3	4.5	5.8	7.4	9.1
40	1.1	1.9	3.0	4.3	5.8	7.6	9.7	11.9
45	1.4	2.4	3.8	5.4	7.4	9.7	12.2	15.1
50	1.7	3.0	4.7	6.7	9.1	11.9	15.1	18.6
55	2.0	3.6	5.6	8.1	11.0	14.4	18.3	22.5
60	2.4	4.3	6.7	9.7	13.1	17.2	21.7	26.8

^a The rows correspond to values of N_o , and the columns correspond to values of N_v . The shaded values correspond to arrays less than or equal to 6 GB.

to be stored in distributed memory. At the limits of $N_o = 60$ and $N_v = 1000$, approximately 27 GB of shared memory per node is required for the T_2 amplitudes. In such cases, the storage of the T_2 amplitudes and the other intermediates is not possible on modern SMP clusters, which, as noted above, typically have 16–64 GB of system memory per node. The present discussion focuses on the implementation for clusters of SMPs; therefore, the practical range of N_o and N_v is defined to be those values for which the size of the T_2 amplitudes is less than 6 GB (the shaded region in Table 2). This practical range of N_o and N_v is defined to overcome the first major storage bottleneck. The same range will be used to examine the sizes for the remaining amplitude and integral classes.

The [VV|OO] and [VO|VO] integrals are similar in size (Table 2) to the T_2 amplitudes, thus over the practical range of N_o and N_v which defines the shared-memory bottleneck of less than 6 GB per array, these quantities are considered small when stored in distributed-memory. Like the T_2 amplitudes, the [VV|OO] and [VO|VO] MO integrals need to be reordered several times throughout the calculation. As mentioned earlier, the reordering of distributed arrays can be very inefficient due to the large amount of communication that is needed. However, unlike the T_2 amplitudes that get updated at the end of every CCSD iteration, the [VV|OO] and [VO|VO] MO integrals are constant for a fixed geometry and basis set. Therefore, instead of reordering the distributed arrays throughout the calculation, two copies of the [VV|OO] integrals and five copies of the [VO|VO] integrals are stored in distributed-memory in the various orders in which they are needed throughout the algorithm. This requires a one-time sorting of the [VV|OO] and [VO|VO] integrals after the integral transformation but prior to the start of the CCSD/CCSD(T) calculation.

The [VV|VO] class of MO integrals is the largest stored quantity in the MP-CCSD(T) algorithm. The distributed-memory needed to store the [VV|VO] integrals represents the second storage bottleneck in the present algorithm. The distributed-memory requirements for the [VV|VO] integrals are given in Table 3. Based on the practical limits of N_o and N_v as governed by the shared-memory bottleneck for the T_2 amplitudes, the largest [VV|VO] distributed-memory arrays can approach ~96 GB. MP-CCSD(T) calculations of this size represent a significant computational challenge. If one employed 128 or more processors for this type of calculation, the storage requirement for the [VV|VO] integrals per

Table 3. Actual Size of the [VV|VO] Integral Class as Stored in Distributed Memory^a

	300	400	500	600	700	800	900	1000
10	1	2	5	8	13	19	27	37
15	2	4	7	12	19	29	41	56
20	2	5	9	16	26	38	54	75
25	3	6	12	20	32	48	68	93
30	3	7	14	24	38	57	82	112
35	4	8	16	28	45	67	95	131
40	4	10	19	32	51	76	109	149
45	5	11	21	36	58	86	122	168
50	5	12	23	40	64	95	136	186
55	6	13	26	44	70	105	150	205
60	6	14	28	48	77	115	163	224

^a The values are in gigabytes (GB). The rows correspond to values of N_0 , and the columns correspond to values of N_v . The shaded values correspond to those values of N_0 and N_v for which the size of the T_2 amplitudes array is less than or equal to 6 GB (Table 2).

Table 4. Size of an N_v^4 Array in Gigabytes (GB)

N_v	300	400	500	600	700	800	900	1000
size (GB)	60	191	466	966	1789	3052	4888	7451

processor would be less than 1 GB. This is easily attained. To decrease the storage requirement of the [VV|VO] integrals, the permutational symmetry of the bra is exploited such that the storage requirement is $[(N_v^2 + N_v) \times N_v N_0] / 2$ words. When required by the algorithm, the lower triangular $([N_v^2 + N_v] / 2)$ rows are expanded to a square (N_v^2) set of rows after the data have been received locally. This provides a nearly 2-fold reduction in the storage and communication costs, at the cost of a slight increase in computational effort. This tradeoff is logical since the computational resources often cost much less than the memory storage or communication infrastructure.

The [VV|VV] integrals are the largest class of integrals needed for a CCSD calculation; however, due to the $O(N_v^4)$ scaling of the storage requirement for these integrals, the values cannot be practically stored in distributed-memory (Table 1). Only one term in the CCSD equations requires the use of the [VV|VV] integrals. This four-index virtual integral term scales computationally as $O(N_0^2 N_v^4)$ and will be referred to here as the *four-virtual term*. An efficient implementation of the four-virtual term is absolutely essential for a CCSD(T) program, since the computational effort required to evaluate the four-virtual term scales as $O(N_v^4)$ with respect to increasing the basis set. Consequently, this is the same rate at which the perturbative triples computation increases using the same metric. By default, most CCSD programs store the [VV|VV] integrals on disk. However, many programs provide the ability to calculate the four-virtual term directly from AO integrals that are calculated “on the fly” rather than stored, thereby avoiding the [VV|VV] integral storage requirement. Methods that avoid storage by calculating quantities “on the fly” are called “direct methods”. Table 4 shows the actual storage requirement in gigabytes for the [VV|VV] class of MO integrals. As N_v increases, the memory requirements for the [VV|VV] integrals exceed the distributed-memory capabilities on the vast majority of available computers. The inability to store the [VV|VV] integrals in distributed memory and the general lack of quality disk I/O on large supercomputers led to the implementation of a “direct” four-virtual algorithm that is calculated in parallel from AO integrals. AO driven methods, both direct and disk-based, for CC methods have been studied in

Table 5. Size of N_v^3 Arrays in Megabytes (MB)

N_v	300	400	500	600	700	800	900	1000
size (MB)	206	488	954	1648	2617	3906	5562	7629

the past.^{42,44,45} Further details about the direct AO driven four-virtual term are given in section IV.B.

Finally, arrays of size N_v^3 are required in both the CCSD and triples correction. For the majority of calculations, N_v^3 arrays are smaller than $N_0^2 N_v^2$ arrays; however, as N_v approaches or surpasses N_0^2 , these N_v^3 arrays can be similar in size (Table 5) or surpass the size of the T_2 amplitudes array (Table 2). It is for that reason that arrays of size N_v^3 are stored in shared memory and not replicated memory. All other $O(N^3)$ arrays and those of lower order are sufficiently small that they can be stored in the replicated memory of each parallel process.

IV. Parallel Implementation

A. CCSD. Once the Hartree–Fock calculation has converged and the molecular integrals have been calculated and sorted, the CCSD iterations begin. The first part of each CCSD iteration is the evaluation of the direct four-virtual term. The details of this direct calculation are described in section IV.B. As mentioned earlier, the direct evaluation of the four-virtual term eliminates the storage requirements of the [VV|VV] integral class, because the integrals are calculated “on demand” during each iteration. After the four-virtual term has been completed, the MO-based terms (eqs 7 and 8) are evaluated in essentially the same order as in the sequential algorithm. The order in which the terms are evaluated has been designed to reduce the number of floating point operations by maximizing the use of intermediate quantities. The sequential algorithm relies heavily on double-precision general matrix–matrix multiplication, DGEMM, operations for the bulk of the computational effort. [DGEMM is a level 3 BLAS (Basic Linear Algebra Subroutine) library function that performs matrix multiplications.] The node-based parallelization strategy for the DGEMM operations of the CCSD algorithm involves partitioning the DGEMM evenly by the number of nodes. Each node gets one portion of the DGEMM to work on. Then each node divides the DGEMM into equal sized work portions for each process to evaluate.

Another challenging aspect of the parallelization of the CCSD algorithm involves the location of the data, i.e., whether the data for the matrices involved in the DGEMM operation are stored in replicated, shared, or distributed memory. Since the CCSD terms involve contracting integrals and amplitudes via DGEMM operations, and since T_1 and T_2 amplitudes or temporary intermediates of the same size are stored locally on each node (T_1 sized arrays in replicated memory and T_2 sized arrays in shared-memory), the distribution of the MO integrals by node is used in the first partitioning of DGEMM operations. The subsequent intra-node partitioning divides the local work among the local processes, where “local” refers to processes within a given node.

For node-based strategies, special care must be taken to ensure the data integrity of shared quantities (both shared

and distributed memory arrays); i.e., before a shared quantity can be used, modified, or reordered, a collective synchronization of the processes that have access to the particular quantity must occur. These collective synchronizations, also known as barriers or fences, are points within the program in which all parallel processes of a collective set must enter before any are allowed to continue executing the parallel program. The MP-CCSD(T) algorithm uses the DDI_SYNC subroutine to synchronize the entire set of parallel processes, while the DDI_SMP_SYNC subroutine is used to synchronize all parallel processes that coexist on the same physical node. These collective synchronization routines help safeguard the integrity of shared resources by ensuring that all parallel processes requiring the use of a shared resource have completed a particular task before those processes are allowed to perform new tasks using the same shared resource. An example of this in terms of distributed memory arrays is found in the four-virtual term (section IV.B) where a global synchronization is used to ensure the distributed intermediate ($I_{ij}^{n\sigma}$) is complete before the second task of forming I_{ij}^{ab} from $I_{ij}^{n\sigma}$ is allowed to begin. However, the most common need for process synchronization occurs when using shared-memory segments within a node. As an example, the evaluation of two CCSD terms that use different orderings of the T_2 amplitudes requires an intranode synchronization to ensure that all local processes have completed the evaluation of the first term and another intranode synchronization to ensure that the entire set of T_2 amplitudes are in the proper order before the evaluation of the second term can begin. This kind of lock-step synchronization can reduce the parallel efficiency of an algorithm if the work between the synchronization points is not evenly balanced.

The following is an example of a node-based algorithm for evaluating the $v_{be}^{am}t_m^e$ component of the I_a^b intermediate (eq 14):

1. Divide n_o by the number of nodes so as to assign each node an equal amount of work.

2. Each node obtains a complete N_v^3 portion of the [VV|VO] integrals from a GET operation based on the index calculated in the previous step, resulting in a 4-index array with dimensions (N_v, N_v, N_v, i) for a given i th index. This array (v_{be}^{ai}) is stored in shared memory. The GET operation is performed only by the master process on each node; therefore, an intranode synchronization is needed before and after this step.

3. Each node performs the permutation of the first and third index, using a routine that allows all the processors on the node to do the permutation in parallel, without overwriting shared memory data. To ensure data integrity, an intranode synchronization is needed after this step is complete.

4. Each node executes a DGEMM (as a $N_v^2 \times N_v$ matrix times a $N_v \times 1$ matrix resulting in a $N_v^2 \times 1$ matrix [$I_a^b = v_{ae}^{bi}t_i^e$ for a fixed i]). This DGEMM is further split among the processes on the node, by dividing N_v^2 (the row dimension of the first matrix) by the number of processors. The actual DGEMM executed by each process consists of a portion of the first matrix times the entire second matrix to yield the entire resultant matrix. In this way, each process works on

a different portion of the array. The second matrix and the product matrix are stored in replicated memory on each parallel process.

5. If N_o is greater than the number of nodes, then some (possibly all) nodes will execute steps 2–4 again with a different portion of the [VV|VO] array until the entire matrix multiplication is performed.

6. Local synchronization: A local gather operation is performed to gather the disjoint set of N_v^2 rows of the product matrices into a single N_v^2 matrix on the master process of each node.

7. The term is completed by a global sum (executed by the master process on each node) over all N_v^2 partial product matrices. A global sum is a form of synchronization.

The remaining terms of the CCSD equations (eqs 7 and 8) have been parallelized using similar techniques to those illustrated in the above example.

In the development of this node-based approach, a similar process-based model was also developed.⁶⁴ Depending on the available memory and the size of the calculation, the MP-CCSD(T) parallel algorithm may be evaluated as a process-based algorithm or a node-based algorithm. The more traditional process-based algorithm, which divides the work based on individual processors, may achieve better intranode performance than the node-based model by removing many of the data synchronizations required; however, the process-based algorithm has a larger memory requirement due to the necessity of more replicated temporary memory, and this significantly limits the size of a molecular problem that can be addressed. Therefore, although both process-based and node-based algorithms have been developed and implemented,⁶⁴ only the node-based algorithm is discussed here, as a primary focus of this discussion to extend the size and complexity of molecular species that can be studied with CCSD(T) methods.

B. “Direct” AO-Driven Four-Virtual Term. The AO “direct” four-virtual term is a distributed-memory algorithm that makes use of both node-based and process-based parallel techniques. The parallel programming techniques used to implement the four-virtual term of the MP-CCSD(T) algorithm were inspired by both the implementation by Fletcher and co-workers⁶¹ of the direct four-index integral transformation⁶² and the direct CCSD algorithm of Kobayashi et al.⁴² Diagrammatic representations of these algorithms are illustrated in Figure 2 (the new MP-CCSD(T) algorithm presented here), Figure 3 (the [VV|OO] integral class of the integral transformation of Fletcher et al.⁶¹), and Figure 4 (four-virtual term of Kobayashi et al.⁴²). The four-virtual terms of each CCSD algorithm contracts AO integrals and amplitudes to calculate the contribution of the four-virtual term to each CCSD iteration.

The four-virtual term of the MP-CCSD(T) algorithm (Figure 2) “directly” calculates full sets of two virtual-indexed half-transformed MO integrals ($v_{ab}^{v\sigma}$) for the specific shell indices ν and σ . The half-transformed integrals are then contracted against the c_{ij}^{ab} amplitudes ($c_{ij}^{ab} = t_{ij}^{ab} + t_i^a t_j^b$) to form a partial contribution to the set of half-transformed intermediates ($I_{ij}^{n\sigma}$), which are complete for a given set of atomic shells ν and σ (Figure 2) corresponding

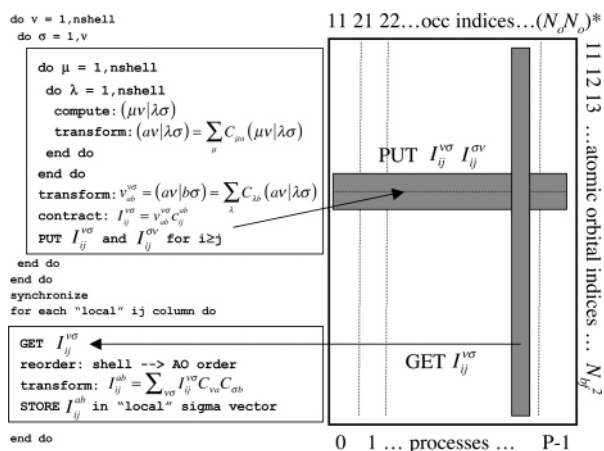


Figure 2. A diagrammatic description of the four-virtual term in the MP-CCSD(T) algorithm. The left-hand portion of the diagram is pseudocode, while the right-hand portion illustrates a distributed array. The columns of the distributed array correspond to two occupied indexes, where the total number of columns is $(N_o N_o)^*$. $(N_o N_o)^*$ refers to the lower triangular portion (including diagonal elements) of an N_o^2 matrix. The number of rows in the distributed matrix is N_b^2 . The columns are distributed evenly over the total number of parallel processes. The boxed portions of the pseudocode represent loadbalanced parallel tasks. The first half of the pseudocode forms the half-transformed intermediate ($I_{ij}^{\nu\sigma}$) in distributed memory. A global synchronization is used to ensure $I_{ij}^{\nu\sigma}$ is complete before the second parallel task begins. The second parallel task transforms $I_{ij}^{\nu\sigma}$ into I_{ij}^{ab} for all "local" i - j columns.

to the parallel task (a given ν - σ pair). For each parallel task the half-transformed intermediates $I_{ij}^{\nu\sigma}$ and $I_{ij}^{\sigma\nu}$ for $i \geq j$ are stored in distributed memory. After the first set of parallel tasks is complete, the full set of half-transformed intermediates $I_{ij}^{\nu\sigma}$ (for $i \geq j$) is stored in distributed memory. To finalize the contributions of the four-virtual CCSD term, the two remaining AO indices are transformed to the virtual MO space.

The four-virtual term gains potential performance advantages over the integral transformation on which it is modeled in two ways: an improved computation vs communication ratio and a reduction in the total number of communication calls. The first parallel task of the four-virtual term (Figure 2) evaluates a larger number of AO integrals and then forms a larger set of half-transformed integrals than the integral transformation in the Fletcher algorithm (Figure 3). In addition, the extra $O(N_o^2 N_v^2)$ contraction step makes the first parallel task of the four-virtual term of the MP-CCSD(T) algorithm significantly more computationally challenging than the first parallel task of the integral transformation. However, both methods share a similar communication profile, which places all of the communication at the end of the parallel task. In fact, the PUT operations performed for the four-virtual term in the MP-CCSD(T) algorithm communicate and store the same amount of data in distributed-memory as the integral transformation does in the formation of the [NN|OO] set of half-transformed integrals. The PUT operations in the four-virtual term of the MP-CCSD(T) algorithm actually gain a slight edge over the PUT operations

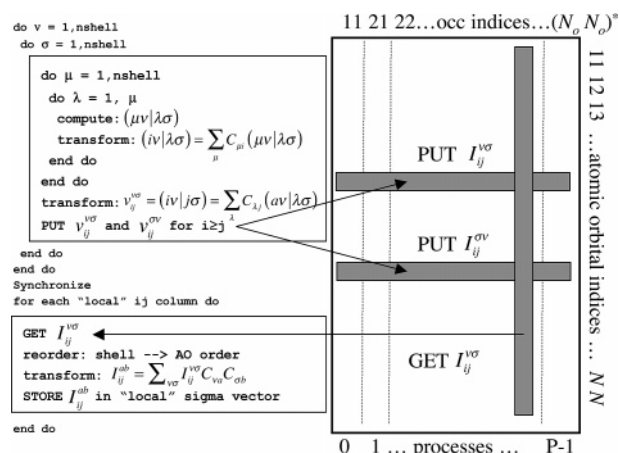


Figure 3. A diagrammatic description of the Fletcher⁵⁰ four-index integral transformation for the [VV|OO] integral class. The left-hand portion of the diagram is pseudocode, while the right-hand portion illustrates a distributed array. The columns of the distributed array correspond to two occupied indexes, where the total number of columns is $(N_o N_o)^*$. $(N_o N_o)^*$ refers to the lower triangular portion (including diagonal elements) of an N_o^2 matrix. The number of rows in the distributed matrix is N_b^2 . The columns are distributed evenly over the total number of parallel processes. The boxed portions of the pseudocode represent load-balanced parallel tasks. The first half of the pseudocode forms half-transformed integrals over two occupied indexes for a given set of two AO shell indexes. The half-transformed integrals are stored in the distributed array. A global synchronization is used to ensure the first task is complete before the second parallel task begins. The second parallel task transforms the final two AO indexes into virtual MO indexes.

of the integral transformation in that in the former, only *one* PUT operation is performed at the end of the first parallel task of the four-virtual term. In contrast, potentially two PUT operations are performed in the integral transformation, except for a single PUT operation when $\nu = \sigma$. Due to the larger *computational* profile of the four-virtual term and a *communication* profile that is similar to the integral transformation, the four-virtual term of the MP-CCSD(T) algorithm is expected to be as good or better in terms of computational efficiency when compared to the integral transformation. The latter has previously been shown to be highly efficient up to 512 processors.⁶⁵

Both distributed-memory CCSD algorithms examined herein form the half-transformed intermediate $I_{ij}^{\nu\sigma}$ of the four-virtual term in distributed memory (Figures 2 and 4). The major difference between the two algorithms is the communication profile. In the four-virtual term of Kobayashi et al.,⁴² the communication calls (GET and ACC) are performed on the inner most nested loop (Figure 4). This type of algorithm was shown to be very successful on the Cray T3E. However, the Cray T3E is very different from modern HPC platforms in that the performance of modern processors has increased by more than an order of magnitude, while the performance of the communication networks have at best doubled or tripled since the benchmarks on the T3E. Therefore the communication heavy inner loop $O(N^4)$

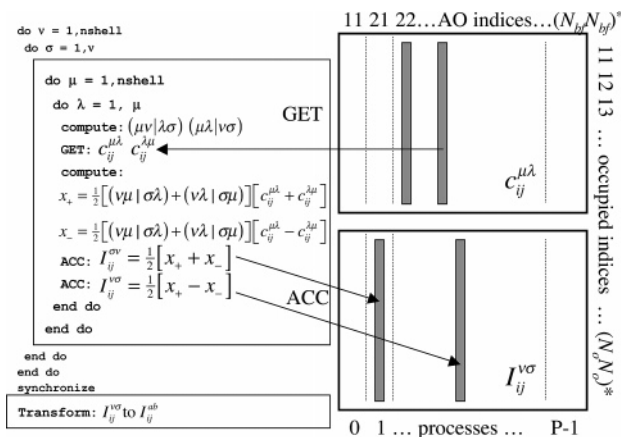


Figure 4. A diagrammatic description of the four-virtual term of Kobayashi et al.³³ The left-hand portion of the diagram is pseudocode, while the right-hand portion illustrates the distributed arrays. The columns of the distributed arrays correspond to two AO indexes where the total number of columns is $(N_{\text{bf}}N_{\text{bf}})^*$. $(N_{\text{bf}}N_{\text{bf}})^*$ refers to the lower triangular portion (including diagonal elements) of an N_{bf}^2 matrix. The number of rows in the distributed matrix is $(N_{\text{o}}N_{\text{o}})^*$ corresponding to the lower triangular portion of an N_{o}^2 matrix. The columns are distributed evenly over the total number of parallel processes. The boxed portions of the pseudocode represent load-balanced parallel tasks. The first half of the pseudocode forms the half-transformed intermediate ($I_{ij}^{\nu\sigma}$) in distributed memory. A global synchronization is used to ensure $I_{ij}^{\nu\sigma}$ is complete before the second parallel task begins. The second parallel task transforms $I_{ij}^{\nu\sigma}$ into $I_{ij}^{\mu\lambda}$ for all “local” i - j columns.

communication calls] is less favorable on modern MP platforms due to this growing discrepancy of the communication network compared to the available computational power. The main benefit of the MP-CCSD(T) routine is that the communication operations are performed at the end of the parallel task making the number of communication calls scale as $O(N^2)$ (Figure 2). The GET operation of the Kobayashi et al.⁴² method is avoided completely by the storage of the c_{ij}^{ab} amplitudes in shared-memory once on every node. The ACC operation of the Kobayashi et al.⁴² method is replaced by a less expensive PUT operation, since the set of $I_{ij}^{\nu\sigma}$ is complete for each set of ν and σ .

The diagrammatic description of the four-virtual term in the MP-CCSD(T) method (Figure 2) is a general description of the algorithm. The actual algorithm as programmed in GAMESS incorporates an extra step to further optimize the first parallel task (see Figure 5). To maximize the efficiency in the contraction step, a local buffer is used to store multiple sets of half-transformed integrals prior to the DGEMM operation. Without the use of the buffer, the size of the DGEMM operation is a function of the size of the basis set shells ν and σ . When ν and σ are s-shells, the DGEMM contraction step reduces to a less than optimal DGEMV (matrix times vector) operation. By locally buffering sets of half-transformed integrals (Figure 5), the efficiency of the DGEMM operation is increased because larger more efficient DGEMM operations are calculated rather than multiple sets of smaller less efficient DGEMV operations. The PUT

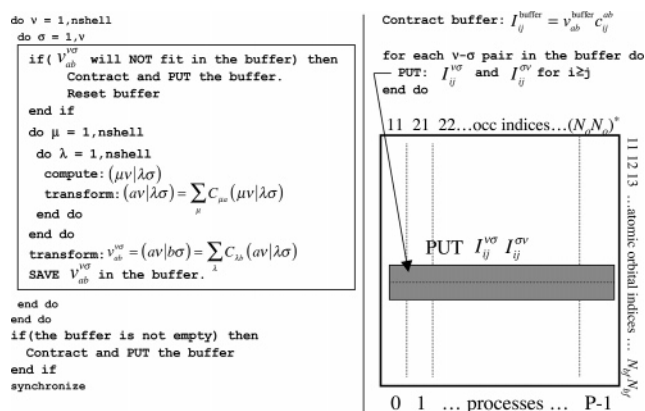


Figure 5. A more detailed overview of the first parallel task in Figure 2 describing the use of a temporary buffer to store half-transformed integrals. The left-hand portion of the figure is the pseudocode describing how the buffer is filled. The right-hand portion describes the “Contract and PUT” operation on the buffer. The description of the distributed array is the same as in Figure 2.

operation for each set of ν and σ is then performed for each ν - σ pair in the contracted buffer.

C. Triples Correction, MP-(T). The (T) portion of the MP-CCSD(T) algorithm is more straightforward to parallelize than the CCSD component. It consists of three nested loops, each of size n_i, n_j, n_k with $i \geq j \geq k$, where i, j , and k are actively occupied indexes. Within each loop, 36 DGEMM calls are made, the largest of which scales computationally as $O(N_{\text{v}}^4)$ and corresponds to DGEMM operation where a $N_{\text{v}} \times N_{\text{v}}$ matrix is multiplied by a $N_{\text{v}} \times N_{\text{v}}^2$ matrix. One feature of the (T) algorithm is that the loop iterations can be performed independently of each other, thus the algorithm can be easily partitioned into unique parallel tasks. The node-based (T) algorithm partitions these independent tasks in terms of sets based on unique values of i, j , and k (occupied indexes), where each task is evaluated on a node. Two N_{v}^3 temporary arrays are stored one time per node in shared memory. Similar to the parallelization scheme of the MO-based MP-CCSD algorithm, when a computationally intensive routine (such as a permutation or DGEMM) is encountered, the work is partitioned equally among the intranode processes, with strict control maintained to avoid overwriting shared memory array locations by multiple processors.

The intranode scaling of the MP-(T) algorithm is expected to exhibit similar trends to those of the MP-CCSD algorithm, since the lock-step synchronization needed between the intranode processes within the node-based tasks are similar. However, due to the larger amount of computational effort per parallel task, the MP-(T) algorithm is expected to perform better.

In general, the MP-(T) algorithm has a large number of independent tasks that are similar to the four-virtual algorithm; however, unlike the four-virtual algorithm, the MP-(T) algorithm does not evaluate the integrals it requires on demand. Rather, it fetches them via GET operations. This aspect of the MP-(T) algorithm increases the communication overhead of the algorithm; however, the $O(N_{\text{v}}^4)$ effort within each parallel task easily compensates to allow for a

Table 6. Total Energies (Hartree), Binding Energies (kcal/mol), and Relative Binding Energies (kcal/mol) Using the aug-cc-pVTZ Basis Set on the MP2/DH(d,p) Optimized Structures of Day et al.^a

	CCSD(T)	CCSD	MP2	MP2*
Total Energies (Hartree)				
prism	-458.13045167	-458.07282232	-458.05015535	-458.05035804
cage	-458.13001003	-458.07248668	-458.05001662	-458.05017138
book	-458.12851572	-458.07140532	-458.04884875	-458.04960143
cyclic	-458.12704114	-458.07054907	-458.04769324	-458.04785303
boat	-458.12514762	-458.06876898	-458.04579806	
Binding Energies (kcal/mol)				
prism	-48.1	-44.6	-47.9	-47.9
cage	-47.8	-44.4	-47.8	-47.8
book	-46.9	-43.7	-47.1	-47.5
cyclic	-46.0	-43.2	-46.4	-46.4
boat	-44.8	-42.1	-45.2	n/a
Relative Binding Energies with Respect to the Prism Isomer (kcal/mol)				
prism	0.0	0.0	0.0	0.0
cage	0.3	0.2	0.1	0.1
book	1.2	0.9	0.8	0.4
cyclic	2.1	1.4	1.5	1.5
boat	3.3	2.5	2.7	n/a

^a The binding energies represent the energy difference between the water hexamer isomer and six isolated water molecules. MP2* represents the MP2/aug-cc-pVTZ calculations from Xantheas et al. who did not examine the boat isomer.

favorable computational vs communication ratio. Therefore, good internode scalability is expected from the MP-(T) routine.

V. Computational Details

The starting set of geometries for the five water hexamer isomers (prism, cage, book, ring, and boat) was obtained from Day et al.⁶⁶ (Figure 1). In that work, the geometries of water hexamer were optimized with second-order Moller–Plesset perturbation theory (MP2)⁵⁴ using the double- ζ Dunning-Hay⁶⁷ [DH(d,p)] basis set. In the present work, single-point CCSD(T) energies were calculated at each previously optimized structure using the following one-electron basis sets: aug-cc-pVTZ⁵³ and aug'-cc-pVTZ, where aug'-cc-pVTZ is a mixed basis set that uses aug-cc-pVTZ on the oxygen atoms and cc-pVTZ⁵⁵ on the hydrogen atoms. The MP-CCSD(T) method in GAMESS was used for all CCSD(T) calculations. A cluster of three IBM Power4 compute nodes each containing eight 1.7 GHz Power4 processors and 32 GB of memory connected by TCP/IP over an InfiniBand network was used to perform the MP-CCSD(T) calculations.

To evaluate the performance of the MP-CCSD(T) algorithm, a series of CCSD(T)/aug'-cc-pVTZ calculations were performed on the MP2/DH(d,p) optimized prism isomer, and the parallel execution times were measured. To test intranode scalability, a set of CCSD(T)/aug'-cc-pVTZ//MP2/DH(d,p) energies were calculated using a single node; the number of parallel processes was varied from 1 to 8 in powers of 2. Internode scalability measures the changes in parallel runtime as the number of nodes is increased, while the number of parallel processes per node (1, 2, 4, or 8) is fixed. In terms of N_o , N_v and the number of Cartesian basis functions N_{bf} , the size of the aug-cc-pVTZ calculation is $N_o = 24$, $N_v = 516$, and $N_{bf} = 630$. The size of the aug'-cc-pVTZ calculation is $N_o = 24$, $N_v = 408$, and $N_{bf} = 510$.

VI. Discussion

Water Hexamer. Calculations performed on the isomers of the water hexamer were used to test the MP-CCSD(T) algorithm. In the first step of what will be a more extensive study of water clusters, CCSD(T) single point energies using the aug-cc-pVTZ and the aug'-cc-pVTZ basis sets were calculated at the MP2 optimized geometries of Day et al.⁶⁶ The absolute energies, binding energies per H₂O, and relative binding energies are given in Table 6 for calculations using the aug-cc-pVTZ basis set and for calculations using the aug'-cc-pVTZ basis set in Table 7. These calculations represent, to the authors' knowledge, the largest CCSD(T) calculations performed on water hexamers to date.

The MP2/DH(d,p) geometries of Day et al.⁶⁶ used in this study may not be as accurate as the MP2/aug-cc-pVTZ geometries of Xantheas et al.;⁵⁶ however, the differences in the binding energies for the two sets of geometries (Table 6) are very small: < 0.1 kcal/mol for the prism, cage, and cyclic isomers and 0.4 kcal/mol for the book isomer. The latter suggests that calculations reported below based on MP2/DH(d,p) geometries for the book isomer may not be as accurate as those for the prism, cage, and cyclic isomers. Xantheas and co-workers did not examine the boat structure.

A main point of interest in this study is the difference between the CCSD(T) and MP2 methods. Column 1 of Table 8 shows the difference in CCSD(T) vs MP2 relative binding energies; positive values indicate an increase in the energy difference between an isomer and the lowest energy prism isomer, i.e., the value in which the prism isomer is stabilized by the CCSD(T) method. In general, CCSD(T) and MP2 predict very similar binding energies. CCSD(T) moderately stabilizes the prism structure with respect to the cage and other higher energy isomers. The prism isomer is stabilized by 0.2 kcal/mol over the next lowest-energy cage isomer. For higher energy isomers, the difference in relative binding

Table 7. Total Energies (Hartree), Binding Energies (kcal/mol), and Relative Binding Energies (kcal/mol) Using the aug'-cc-pVTZ Basis Set on the MP2/DH(d,p) Optimized Structures of Day et al.^a

	CCSD(T)	CCSD	MP2
Total Energies (Hartree)			
prism	-458.12255430	-458.06558835	-458.04247161
cage	-458.12223107	-458.06536922	-458.04244064
book	-458.12086894	-458.06440785	-458.04140132
cyclic	-458.11967703	-458.06381033	-458.04052628
boat	-458.11779063	-458.06203960	-458.03863436
Binding Energies (kcal/mol)			
prism	-46.6	-43.2	-46.6
cage	-46.4	-43.1	-46.6
book	-45.6	-42.5	-45.9
cyclic	-44.8	-42.1	-45.4
boat	-43.6	-41.0	-44.2
Relative Binding Energies with Respect to the Prism Isomer (kcal/mol)			
prism	0.0	0.0	0.0
cage	0.2	0.1	0.0
book	1.0	0.7	0.6
cyclic	1.8	1.1	1.2
boat	3.0	2.2	2.4

^a The binding energies represent the energy difference between the water hexamer isomer and six isolated water molecules.

energies is larger: 0.4 kcal/mol for the book isomer and 0.6 kcal/mol for the cyclic and boat isomers. While the differences in relative binding energies between CCSD(T) and MP2 (0.2–0.6 kcal/mol) are modest when the aug'-cc-pVTZ basis set is used, it is unclear how basis set improvements will affect these energy differences.

Another interesting issue is the accuracy of the CCSD method with respect to CCSD(T) and MP2. The CCSD(T) and MP2 binding energies agree to within ~0.5 kcal/mol for both the aug- and aug'-cc-pVTZ basis sets (Tables 6 and 7). However, the CCSD binding energies differ from the CCSD(T) binding energies by 2.7–3.5 kcal/mol (Tables 6 and 7). Assuming that CCSD(T) provides the most accurate binding energies, these calculations suggest that the MP2 method can more accurately predict the binding energies than the CCSD method. Kim et al.⁶⁸ reported such a difference between CCSD and MP2 for cyclic water hexamer. This is surprising, since the CCSD method is often considered to be more reliable than MP2.

Table 8 describes in more detail how the prism isomer is stabilized by the CCSD(T) method based on differences between CCSD(T) and CCSD (column 2) and differences between CCSD and MP2 (column 3). The triples correction to the CCSD energy (column 2, Table 8) plays an increasingly larger role in stabilizing the prism structure relative to higher energy isomers. The difference between CCSD and MP2 (column 3, Table 8) stabilizes the prism structure over the cage and book structures but decreases the stability of the prism structure relative to the cyclic and boat structures. The effects of the triples approach 1 kcal/mol and should not be overlooked, especially for larger water clusters.

The two basis sets employed here exhibit very similar trends in the differences of relative binding energies for all

Table 8. Difference in Relative Binding Energies between the CCSD(T), CCSD, and MP2 Methods with Respect to the Basis Set Measured in kcal/mol^a

	CCSD(T)-MP2	aug'-cc-pVTZ (T)-CCSD	CCSD-MP2
prism	0.00	0.00	0.00
cage	0.19	0.07	0.12
book	0.39	0.33	0.07
cyclic	0.60	0.71	-0.12
boat	0.59	0.78	-0.19

	CCSD(T)-MP2	aug'-cc-pVTZ (T)-CCSD	CCSD-MP2
prism	0.00	0.00	0.00
cage	0.18	0.07	0.12
book	0.39	0.32	0.07
cyclic	0.58	0.69	-0.10
boat	0.58	0.76	-0.18

	CCSD(T)-MP2	difference (T)-CCSD	CCSD-MP2
prism	0.00	0.00	0.00
cage	0.01	0.00	0.01
book	0.01	0.01	0.00
cyclic	0.01	0.02	-0.01
boat	0.01	0.02	-0.01

^a The first column [CCSD(T)-MP2] shows how the relative binding energies differ between the CCSD(T) and the MP2 method. The second column [CCSD(T)-CCSD] shows the effect of the triples correction on the relative binding energies. The last column [CCSD-MP2] shows the differences between CCSD and MP2 on the relative binding energies. The section subtitled "difference" subtracts the values from the aug'-cc-pVTZ set from the corresponding aug'-cc-pVTZ set.

methods. Csonka and co-workers⁶⁹ have suggested that including diffuse functions in the oxygen atom basis set is important. In the present work, the aug'-cc-pVTZ basis set only includes diffuse functions on the oxygen atoms. The omitted hydrogen diffuse functions in the aug'-cc-pVTZ basis set were found to increase the binding energies of the water clusters by 1.0–1.4 kcal/mol (Table 6 and Table 7); therefore the diffuse functions on the hydrogen atoms do seem to be important for calculating the *absolute* binding energies. However, the *relative* binding energies (Table 8) predicted by the aug- and aug'-cc-pVTZ basis sets are very similar. For example, column 1 of Table 8 describes the differences in relative binding energies between CCSD(T) and MP2. These values are virtually identical for each isomer for both basis sets. This consistency in the differences between CCSD(T) and MP2 for the aug- and aug'-cc-pVTZ basis sets suggests that the CCSD(T)/aug'-cc-pVTZ binding energies can be accurately estimated using computationally less intensive CCSD(T)/aug'-cc-pVTZ and MP2/aug'-cc-pVTZ calculations. As illustrated in Table 9, the following additive scheme,

$$\text{CCSD(T)/aug} = \text{CCSD(T)/aug}' + [\text{MP2/aug} - \text{MP2/aug}'] \quad (29)$$

where the -cc-pVTZ extension to the basis set is implied, estimates the actual CCSD(T)/aug'-cc-pVTZ binding energies to within less than 0.2 kcal/mol. A future study will examine

Table 9. Estimated CCSD(T)/aug-cc-pVTZ Binding Energies (kcal/mol) Using Eq 29 Compared to the Actual CCSD(T)/aug-cc-pVTZ at the MP2/DH(d,p) Optimized Geometries^a

	prism	cage	book	cyclic	boat
est. CCSD(T)/ aug-cc-pVTZ	-47.9	-47.6	-46.7	-45.7	-44.6
actual CCSD(T)/ aug-cc-pVTZ	-48.1	-47.8	-46.9	-45.9	-44.8
error	0.2	0.2	0.2	0.2	0.2

^a The differences were rounded up.**Table 10.** Parallel Speedup (*S*) and Parallel Efficiency (*E*) for the MP-CCSD(T) Algorithm as a Function of the Number of Processors per Node (PPN) and the Number of Nodes for Calculations Performed on the Prism Isomer Using the aug-cc-pVTZ Basis Set^a

processes per node	1		2		4		8	
	<i>S</i>	<i>E</i> (%)	<i>S</i>	<i>E</i> (%)	<i>S</i>	<i>E</i> (%)	<i>S</i>	<i>E</i> (%)
1 Node								
CCSD-AO	1.00	100	1.90	95	3.70	92	6.18	77
CCSD-MO	1.00	100	1.87	93	3.11	78	4.21	53
CCSD-total	1.00	100	1.86	93	3.58	89	5.68	71
triples correction (T)	1.00	100	1.78	89	2.59	65	4.06	51
2 Nodes								
CCSD-AO	2.00	100	3.76	94	7.43	93	12.31	77
CCSD-MO	1.38	69	2.46	62	4.10	51	6.21	39
CCSD-total	1.88	94	3.34	84	6.53	82	9.56	60
triples correction (T)	1.94	97	3.38	85	4.73	59	7.13	45
3 Nodes								
CCSD-AO	3.00	100	5.85	97	11.07	92	18.48	77
CCSD-MO	1.68	56	2.96	49	4.56	38	6.91	29
CCSD-total	2.55	85	4.80	80	8.28	69	14.57	61
triples correction (T)	2.95	98	5.24	87	7.63	64	11.82	49

^a CCSD-AO represents the AO-driven four virtual term of the MP-CCSD algorithm; CCSD-MO represents all the other MO-based terms of the MP-CCSD algorithm. CCSD-total is the overall scalability for the MP-CCSD algorithm. The speedup and efficiency is also given for the triples correction. Intranode trends are observed across rows, while internode trends are observed down the columns. The benchmark calculations are based on MP-CCSD(T) calculations of the water hexamer (prism isomer) with $N_o = 24$, $N_v = 408$, and $N_{br} = 510$ run on nodes containing a total of 8 processors.

the extrapolation of the CCSD(T) binding energies of water hexamer isomers to the complete basis set limit (CBS).

Parallel Performance. The speedup and efficiency values for the four virtual term (CCSD-AO) and the remaining MO terms (CCSD-MO) from the MP-CCSD method on the benchmark calculation run on the IBM Power4 platform are given in Table 10. Speedup is defined as the ratio of the measured execution time to the execution time on a single processor; efficiency is the ratio of the measured speedup compared to the ideal speedup.

The intranode scalability of the MP-CCSD method was measured by the speedup and efficiency of the benchmark calculation as the number of processors within a single node was increased. The intranode scalability of the AO driven

four-virtual term (CCSD-AO) is better than 90% of ideal over two and four processors within one node; however, the efficiency drops to approximately 77% when all eight processors within the node are used (Table 10). The drop in performance when using all eight processors with a single node is likely due to memory bandwidth limitations; i.e. all eight processors within the node were accessing and utilizing the same local system memory. The scalability of the MO based terms of the MP-CCSD algorithm is approximately 93%, 77%, and 52% efficient when run on 2, 4, and 8 processors, respectively, within the same node (Table 10). The intranode scalability of the MO based MP-CCSD terms suffers due to the high degree of synchronization needed between local processes; the lock-step manner in which the terms are calculated results in deviations from ideal speedup. The MO-based terms also require a significant number of cache unfriendly rearrangements of the T_2 amplitudes. These operations, similar to the four-virtual term, stress the memory bandwidth of the system and result in less than ideal scalability.

The internode scalability of the MP-CCSD method was measured as the number of nodes was increased, while the number of processors per node (PPN) was kept fixed. The internode scalability of the AO driven four-virtual term (CCSD-AO) is extremely good (Table 10), i.e., the parallel efficiency measured on one node stays approximately the same as the number of nodes is increased. This high degree of internode scalability is expected because very little communication is required relative to the amount of computational effort needed for the four-virtual term. The internode speedup is expected to extend well beyond three nodes, since the four-virtual term was modeled upon, and has a better computational vs communication ratio than the direct four-index integral transformation.

The internode scalability of the MO based terms suffers due to a low computation vs communication ratio. As mentioned earlier, the MO terms of the MP-CCSD method require a high degree of synchronization. Some of these synchronization points in the MO based MP-CCSD algorithm are collective operations which require a considerable amount of network communication. The lower computation vs communication ratio resulting from higher internode communication, combined with smaller computational workloads, significantly reduces the internode scalability of the MO based terms in the MP-CCSD program.

Despite the poor scaling of the MO-based terms, reasonable overall scalability is achieved for the MP-CCSD algorithm due to the highly scalability and overwhelmingly dominant four-virtual term. On a single processor, 88% of the execution time of the MP-CCSD algorithm was spent calculating the four-virtual term in the benchmark calculation. The outlook for the MP-CCSD algorithm for larger calculations is good, since the four-virtual term becomes increasingly dominant for larger calculations.

The performance of the triples (T) correction in the MP-CCSD(T) algorithm falls in between that of the four-virtual term and the MO-based terms of the MP-CCSD algorithms. Similar to the four-virtual term, the MP-(T) algorithm scales well as the number of nodes is increased; i.e., the efficiency

does not change considerably as the number of nodes is increased for a given number of processes per node (PPN). This is expected due to the general independence of the work distribution. The intranode scaling of the (T) part suffers in the benchmark calculation from intranode synchronization and a relatively small value of N_v . That is $N_v = 408$ is at the low end of the range of N_v for which the algorithm was designed ($300 < N_v < 1000$), and, as such, the computational effort needed to evaluate the intranode parallel DGEMMs does not scale well because the subdivided DGEMMs evaluated per process are too small in size to gain a significant advantage from the highly optimized BLAS library. In the benchmark calculations, the parallel efficiency drops to just under 90% for PPN = 2, 65–70% for PPN = 4, and approximately 50% for PPN = 8 (Table 10). Larger values of N_v would provide a greater amount of computational work, and better scaling is expected.

Overall, the MP-CCSD(T) routine is dominated by two key terms: the four-virtual term and the (T) term. The intranode scaling of both of these terms is the major performance limiting variable. However, based on a fixed number of processors per node, the internode scaling is very good. Therefore, in general, a constant speedup is expected as the number of nodes is increased, even though this speedup is less than ideal due to the less than desirable intranode scaling.

Future Enhancements. The four-virtual term was designed based on the premise that quality disk I/O would not be generally available, so the method, as presented, is a fully direct algorithm. This decision was deliberate since many of the next-generation MP platforms may not have local scratch disks. However, a considerable saving in the cost of recalculating the AO integrals might be achieved by making use of a local scratch disk to store integrals or intermediates. One way in which a local scratch disk could be utilized to reduce the computational cost of recalculating the AO integrals would be to selectively store those sets of half-transformed integrals that are the *most* expensive to recalculate. Using the angular momentum quantum number (l) for the basis set shells ν and o , then for each set of ν and o in which the sum $l(\nu) + l(o)$ is larger than a user defined input parameter, the half-transformed integrals for the set of ν and o would be saved on the local disk during the first CCSD iteration. Subsequent CCSD iterations process all “local” sets of half-transformed integrals stored on a disk before processing the remaining sets of half-transformed integrals that must be calculated directly. There are a variety of ways in which load balancing might be achieved in such a scheme. One method would be to statically distribute the disk-based tasks while dynamically distributing the direct tasks. This would ensure that a similar amount of scratch disk is used on each node, while the dynamically distributed direct task would compensate for any potential load imbalances from the disk-based portion of the algorithm.

The limited scalability of the MO-based terms of the MP-CCSD method represents one of the major limitations in the current MP-CCSD algorithm. Each CCSD iteration performs the computationally demanding four-virtual term using every parallel process, and, when complete, every parallel process

is then used to calculate the MO-based terms. Since the four-virtual term scales extremely well with the number of parallel processes, it is desirable to utilize a large number of CPUs to gain significant computational speedup. However, since the MO-based terms reach asymptotic scaling with significantly fewer processes, performing these operations sequentially results in a loss of efficiency due to the MO-based terms. To compensate for this limitation, the MO-based terms could be calculated concurrently with the AO-based terms. Using n nodes to calculate the MO-based terms, where n is the maximum number of nodes for which the MO-based terms achieve better than 50–75% parallel efficiency, the remaining nodes would then immediately begin work on the computationally dominant AO-based terms. Since the AO-based tasks are so computationally dominant, the n nodes used to calculate the MO-based terms could potentially finish before the AO-based terms are completed. In that case, those nodes would assist in the completion of the AO-based terms. This scheme would maximize the efficiency of the MP-CCSD algorithm.

Finally, improvements in the intranode scaling would benefit every step in the MP-CCSD(T) algorithm. However, in terms of an overall reduction in wall time, the biggest computational saving could be gained by improving the intranode performance of the MP-(T) algorithm. One means of improving the intranode performance of the MP-(T) algorithm (and also the intranode MP-CCSD algorithm) is to explore the use of a shared-memory model based on threads rather than processes. Thread-based models like OpenMP³⁸ and/or POSIX threads (Pthreads) offer a greater set of tools which are generally more robust and better performing than the limited capabilities of the System V model. Improved synchronization routines and better tuning of the intranode portion of the (T) algorithm should result in the biggest overall performance improvements.

VII. Conclusions

The MP-CCSD(T) algorithm was shown to achieve reasonable scalability for chemically interesting systems, i.e., water hexamer. The most computationally challenging portions of the algorithm, the four-virtual term and the triples corrections, achieve good internode scalability, which implies that the performance will scale well up to a large number of nodes. In general, the intranode scalability for both the MP-CCSD and MP-(T) was found to be less than optimal. However, it was only the use of the node-based model that provided the ability to perform these calculations by making it possible to store all of the various data structures. Careful consideration of the data and storage model is as crucial to the algorithm design as is CPU scaling.

The CCSD(T) calculations on isomers of water hexamer show good agreement between the CCSD(T) and MP2 methods, while the CCSD method predicts significantly worse binding energies than either CCSD(T) or MP2. While the differences between the CCSD(T) and MP2 methods are small, these differences could be important at the CBS limit or for larger water clusters, since the geometric isomers are themselves very similar in energy. Diffuse functions on the hydrogen atoms are important for calculating accurate

binding energies; however, the contributions of these diffuse functions to the total energy of higher level methods, like CCSD(T), can be accurately estimated using energy differences from calculations performed at a lower level of theory, e.g., MP2.

Overall, the MP-CCSD(T) algorithm offers a node-based parallel algorithm designed to take advantage of modern cluster of SMPs. With the ever increasing trend toward more intranode compute power, most notably with the advent of multicore processors, the distinction between internode and intranode parallelism will become more important. The present work provides an initial analysis of how effectively this dual-level parallelism can be applied to modern state-of-the-art ab initio methods. While further optimizations to improve the algorithm, especially the intranode portions, should be considered, the MP-CCSD(T) method presented here is capable of calculating CCSD(T) energies for a system up to approximately 1000 basis functions in a massively parallel environment.

Acknowledgment. The authors are grateful for many illuminating discussions with Professor Alistair Rendell. The authors thank the Scalable Computing Laboratory of the Ames Laboratory for the generous allotment of computer resources allocated for use on this project. The authors also thank Professor Donald Truhlar and Ms. Erin Dalke for helpful discussion regarding basis sets, in particular the aug'-cc-pVTZ basis set. This work was supported by an ITR (Information Technology Research) grant from the National Science Foundation.

References

- (1) Cížek, J.; Paldus, J. *Int. J. Quantum Chem.* **1971**, *5*, 359.
- (2) Cížek, J. *Adv. Chem. Phys.* **1969**, *14*, 35.
- (3) Cížek, J. *J. Chem. Phys.* **1966**, *45*, 4256.
- (4) Purvis, G. D., III; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910.
- (5) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
- (6) Scuseria, G. E.; Lee, T. J. *J. Chem. Phys.* **1990**, *93*, 5851.
- (7) Hoffmann, M. R.; Schaefer, H. F., III. *Adv. Quantum Chem.* **1986**, *18*, 207.
- (8) Geertsen, J.; Rittby, M.; Bartlett, R. J. *Chem. Phys. Lett.* **1989**, *164*, 57.
- (9) Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 7029.
- (10) Comeau, D. C.; Bartlett, R. J. *Chem. Phys. Lett.* **1993**, *207*, 414.
- (11) Piecuch, P.; Bartlett, R. J. *Adv. Quantum Chem.* **1999**, *34*, 295.
- (12) Kowalski, K.; Piecuch, P. *J. Chem. Phys.* **2004**, *120*, 1715.
- (13) Krylov, A. I. *Chem. Phys. Lett.* **2001**, *338*, 375.
- (14) Krylov, A. I.; Sherrill, C. D. *J. Chem. Phys.* **2002**, *116*, 3194.
- (15) Kowalski, K.; Piecuch, P. *J. Chem. Phys.* **2000**, *113*, 18.
- (16) Piecuch, P.; Wloch, M.; Gour, J. R.; Kinal, A. *Chem. Phys. Lett.* **2005**, *418*, 463.
- (17) Scuseria, G. E.; Lee, T. J.; Schaefer, H. F., III. *Chem. Phys. Lett.* **1986**, *130*, 236.
- (18) Scuseria, G. E.; Scheiner, A. C.; Lee, T. J.; Rice, J. E.; Schaefer, H. F., III. *J. Chem. Phys.* **1987**, *86*, 2881.
- (19) Lee, T. J.; Rice, J. E. *Chem. Phys. Lett.* **1988**, *150*, 406.
- (20) Scuseria, G. E.; Janssen, C. L.; Schaefer, H. F., III. *J. Chem. Phys.* **1988**, *89*, 7382.
- (21) Stanton, J. F.; Gauss, J.; Watts, J. D.; Bartlett, R. J. *J. Chem. Phys.* **1991**, *94*, 4334.
- (22) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; et al. *J. Comput. Chem.* **1993**, *14*, 1347.
- (23) MOLPRO is a package of ab initio programs written by H.-J. Werner, P. J. Knowles, R. Lindh, F. R. Manby, M. Schütz, P. Celani, T. Korona, G. Rauhut, R. D. Amos, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, C. Hampel, G. Hetzer, A. W. Lloyd, S. J. McNicholas, W. Meyer, M. E. Mura, A. Nicklaß, P. Palmieri, R. Pitzer, U. Schumann, H. Stoll, A. J. Stone, R. Tarroni, and T. Thorsteinsson.
- (24) ACES II is a program product of the Quantum Theory Project, University of Florida. Authors: J. F. Stanton, J. Gauss, J. D. Watts, M. Nooijen, N. Oliphant, S. A. Perera, P. G. Szalay, W. J. Lauderdale, S. A. Kucharski, S. R. Gwaltney, S. Beck, A. Balková D. E. Bernholdt, K. K. Baeck, P. Rozyczko, H. Sekino, C. Hober, and R. J. Bartlett. Integral packages included are VMOL (J. Almlöf and P. R. Taylor); VPROPS (P. Taylor); and ABACUS (T. Helgaker, H. J. Aa. Jensen, P. Jørgensen, J. Olsen, and P. R. Taylor).
- (25) Kong, J.; White, C. A.; Krylov, A. I.; Sherrill, D.; Adamson, R. D.; Furlani, T. R.; Lee, M. S.; Lee, A. M.; Gwaltney, S. R.; Adams, T. R.; Ochsenfeld, C.; Gilbert, A. T. B.; Kedziora, G. S.; Rassolov, V. A.; Maurice, D. R.; Nair, N.; Shao, Y.; Besley, N. A.; Maslen, P. E.; Dombroski, J. P.; Daschel, H.; Zhang, W.; Korambath, P. P.; Baker, J.; Byrd, E. F. C.; Van Voorhis, T.; Oumi, M.; Hirata, S.; Hsu, C.-P.; Ishikawa, N.; Florian, J.; Warshel, A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M.; Pople, J. A. *J. Comput. Chem.* **2000**, *21*, 1532.
- (26) Crawford, T. D.; Sherrill, C. D.; Valeev, E. F.; Fermann, J. T.; King, R. A.; Leininger, M. L.; Brown, S. T.; Janssen, C. L.; Seidl, E. T.; Kenny, J. P.; Allen, W. D. *PSI*, *3.2*; 2003.
- (27) Aprà, E.; Windus, T. L.; Straatsma, T. P.; Bylaska, E. J.; de Jong, W. A.; Hirata, S.; Valiev, M.; Hackler, M.; Pollack, L.; Kowalski, K.; Harrison, R. J.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Auer, A. A.; Brown, E.; Cisneros, G.; Fann, G.; Fruchtl, H.; Garza, J.; Hirao, K.; Kendall, R. A.; Nichols, J.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D. E.; Borowski, P.; Clark, T.; Clerc, D.; Daschel, H.; Deegan, M.; Dyall, K. G.; Elwood, D.; Glendenning, E. D.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R. J.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers, Version 4.7*; Pacific Northwest National Laboratory: Richland, Washington 99352-0999, U.S.A., 2005.
- (28) Kendall, R. A.; Aprà, E.; Bernholdt, D. E.; Bylaska, E. J.; Dupuis, M.; Fann, G. I.; Harrison, R. J.; Ju, J.; Nichols, J. A.; Nieplocha, J.; Straatsma, T. P.; Windus, T. L.; Wong, A. T. *Comput. Phys. Commun.* **2000**, *128*, 260.
- (29) Dalton, a molecular electronic structure program, Release 2.0, 2005. See <http://www.kjemi.uio.no/software/dalton/dalton.html> (accessed May 2007).

- (30) 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.; 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, A.; Peng, C. Y.; Nanyakkara, 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.
- (31) Olson, R. M.; Varganov, S.; Gordon, M. S.; Metiu, H.; Chretien, S.; Piecuch, P.; Kowalski, K.; Kucharski, S. A.; Musial, M. *J. Am. Chem. Soc.* **2005**, *127*, 1049.
- (32) Harrison, R. J. *Int. J. Quantum Chem.* **1991**, *40*, 847.
- (33) *SHMEM Library Reference*, Cray Research Incorporated.
- (34) TheMPIForum, MPI: a message passing interface. In *Proceedings of the 1993 ACM/IEEE conference on Supercomputing*, ACM Press: Portland, OR, United States, 1993.
- (35) Nieplocha, J.; Harrison, R. J.; Littlefield, R. J. *J. Supercomput.* **1996**, *10*, 169.
- (36) Olson, R. M.; Schmidt, M. W.; Gordon, M. S.; Rendell, A. P. Enabling the Efficient Use of SMP Clusters: The GAMESS/DDI Approach. In *Supercomputing, 2003 ACM/IEEE Conference*, Phoenix, AZ, 2003; p 41.
- (37) Fletcher, G. D.; Schmidt, M. W.; Bode, B. M.; Gordon, M. S. *Comput. Phys. Commun.* **2000**, *128*, 190.
- (38) Dagum, L.; Menon, R. *IEEE Comput. Sci. Eng.* **1998**, *5*, 46.
- (39) Rendell, A. P.; Lee, T. J.; Komornicki, A. *Chem. Phys. Lett.* **1991**, *178*, 462.
- (40) Rendell, A. P.; Lee, T. J.; Lindh, R. *Chem. Phys. Lett.* **1992**, *194*, 84.
- (41) Rendell, A. P.; Guest, M. F.; Kendall, R. A. *J. Comput. Chem.* **1993**, *14*, 1429.
- (42) Kobayashi, R.; Rendell, A. P. *Chem. Phys. Lett.* **1997**, *265*, 1.
- (43) Janowski, T.; Ford, A. R.; Pulay, P. Abstracts of Papers, 232nd ACS National Meeting, San Francisco, CA, United States, Sept. 10–14, 2006, 2006, PHYS.
- (44) Koch, H.; Christiansen, O.; Kobayashi, R.; Jorgensen, P.; Helgaker, T. *Chem. Phys. Lett.* **1994**, *228*, 233.
- (45) Koch, H.; Sanchez de Meras, A.; Helgaker, T.; Christiansen, O. *J. Chem. Phys.* **1996**, *104*, 4157.
- (46) Rendell, A. P.; Lee, T. J. *J. Chem. Phys.* **1994**, *101*, 400.
- (47) Ford, A. R.; Janowski, T.; Pulay, P. *J. Comput. Chem.* **2007**, *28*, 1215.
- (48) Auer, A.; Baumgartner, G.; Bernholdt, D. E.; Bibireata, A.; Choppella, V.; Cociorva, D.; Gao, X.; Harrison, R. J.; Krishnamoorthy, S.; Krishnan, S.; Lam, C.-C.; Nooijen, M.; Pitzer, R. M.; Ramanujam, J.; Sadayappan, P.; Sibiryakov, A. *Mol. Phys.* **2006**, *104*, 211.
- (49) Hirata, S. *J. Phys. Chem. A* **2003**, *107*, 9887.
- (50) Piecuch, P.; Hirata, S.; Kowalski, K.; Fan, P.-D.; Windus, T. L. *Int. J. Quantum Chem.* **2005**, *106*, 79.
- (51) Baumgartner, G.; Auer, A.; Bernholdt, D. E.; Bibireata, A.; Choppella, V.; Cociorva, D.; Gao, X.; Harrison, R. J.; Hirata, S.; Krishnamoorthy, S.; Krishnan, S.; Lam, C.; Lu, Q.; Nooijen, M.; Pitzer, R. M.; Ramanujam, J.; Sadayappan, P.; Sibiryakov, A. Synthesis of High-Performance Parallel Programs for a Class of Ab Initio Quantum Chemistry Models. In *Proceedings of the IEEE*; 2005.
- (52) Piecuch, P.; Kucharski, S. A.; Kowalski, K.; Musial, M. *Comput. Phys. Commun.* **2002**, *149*, 71.
- (53) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.
- (54) Moller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.
- (55) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (56) Xantheas, S. S.; Burnham, C. J.; Harrison, R. J. *J. Chem. Phys.* **2002**, *116*, 1493.
- (57) Feller, D. *J. Chem. Phys.* **1992**, *96*, 6104.
- (58) Crawford, T. D.; Schaefer, H. F., III. *Rev. Comput. Chem.* **2000**, *14*, 33.
- (59) Piecuch, P.; Kowalski, K.; Pimienta, I. S. O.; McGuire, M. J. *Int. Rev. Phys. Chem.* **2002**, *21*, 527.
- (60) Urban, M.; Noga, J.; Cole, S. J.; Bartlett, R. J. *J. Chem. Phys.* **1985**, *83*, 4041.
- (61) Fletcher, G. D.; Schmidt, M. W.; Gordon, M. S. *Adv. Chem. Phys.* **1999**, *110*, 267.
- (62) Wong, A. T.; Harrison, R. J.; Rendell, A. P. *Theor. Chim. Acta* **1996**, *93*, 317.
- (63) Olson, R. M.; Gordon, M. S. *J. Chem. Phys.*, in press.
- (64) Bentz, J. L.; Olson, R. M.; Gordon, M. S.; Schmidt, M. W.; Kendall, R. A. *Comput. Phys. Commun.* **2007**, *176*, 589.
- (65) Kudo, T.; Gordon, M. S. *J. Phys. Chem. A* **2001**, *105*, 11276.
- (66) Day, P. N.; Pachter, R.; Gordon, M. S.; Merrill, G. N. *J. Chem. Phys.* **2000**, *112*, 2063.
- (67) Dunning, T. H., Jr.; Hay, P. J. In *Methods of Electronic Structure Theory*; Plenum Press: New York, 1977; p 1.
- (68) Kim, J.; Kim, K. S. *J. Chem. Phys.* **1998**, *109*, 5886.
- (69) Csonka, G. I.; Ruzsinszky, A.; Perdew, J. P. *J. Phys. Chem. B* **2005**, *109*, 21471.

CT600366K

Critical Role of the Correlation Functional in DFT Descriptions of an Agostic Niobium Complex

Dimitrios A. Pantazis,[†] John E. McGrady,^{*,†} Feliu Maseras,[‡] and Michel Etienne[§]

WestCHEM, Department of Chemistry, Joseph Black Building, University of Glasgow, Glasgow G12 8QQ, U.K., Institute of Chemical Research of Catalonia (ICIQ), Avda. Països Catalans 16, 43007, Tarragona, Spain, Unitat de Química Física, Edifici Cn, Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain, and Laboratoire de Chimie de Coordination du CNRS, UPR 8241 liée par conventions à l'Université Paul Sabatier et à l'Institut National Polytechnique de Toulouse, 205 Route de Narbonne, Toulouse Cedex 4, France

Received February 21, 2007

Abstract: In previous studies of the agostic bonding in $\text{Tp}^{\text{Me}_2}\text{NbCl}(\text{R}'\text{CCR}'')(\text{R})$, we have made use of a hybrid QM/MM protocol (B3LYP:UFF) where the QM partition ($[\text{Nb}(\text{Cl})(\text{iPr})(\text{HCCH}-(\text{NHCH}_2)_3]^+$) was rather small, but the optimized structures were nevertheless in apparently good agreement with experiment. In attempting to improve this model by expanding the size of the QM region, we were surprised to discover that a full QM treatment of the whole molecule using the B3LYP functional failed to locate an agostic structure of any kind. A systematic assessment of density functionals reveals that the poor performance of B3LYP in these systems is typical of all DFT methods that do not obey the uniform electron gas (UEG) correlation limit. Those that do obey the UEG limit, in contrast, provide an excellent description of the agostic structure when the complete ligand system is treated at the QM level. The apparently good performance of our original (B3LYP:UFF) hybrid method can be traced to a cancellation of errors: the B3LYP functional underestimates the intrinsic strength of the agostic interaction relative to competing Nb–Cl π bonding, but this is offset by an additional but unphysical electrostatic component to the agostic bond introduced by the presence of a positive charge in the QM region.

Introduction

The nature of the agostic bond in transition-metal organometallic compounds continues to excite debate in the literature.^{1–3} The earliest, and perhaps simplest, model of a C–H \cdots M interaction⁴ is as a 3-center 2-electron interaction between a C–H (or C–C) σ bond and an electron-deficient metal center, a picture that was supported by early studies of the β -C–H agostic interaction using Extended Hückel

theory,⁵ along with a more recent analysis using the Atoms in Molecules (AIM) procedure.⁶ Scherer and McGrady have, however, arrived at a rather different model of the agostic bonding in electropositive d^0 metal alkyls based on hyperconjugative stabilization of the M–C bonding electrons by antibonding orbitals localized on the alkyl group.^{7,8} This model is somewhat reminiscent of the early work of Eisenstein and co-workers on α -C–H agostics,⁵ where delocalization of the M–C bonding electrons, in this case into vacant orbitals on the metal, was identified as the major driving force for the α -C–H agostic structure of $\text{H}_5\text{TiCH}_3^{2-}$. In more electron-rich late transition metals, back-bonding into the C–H σ^* orbitals can also stabilize agostic bonds,⁹ often leading to substantial elongation of the C–H bonds (~ 1.25 Å¹⁰ compared to ~ 1.12 Å and ~ 1.10 Å in electron-

* Corresponding author e-mail: j.mcgrady@chem.gla.ac.uk.

[†] University of Glasgow.

[‡] Institute of Chemical Research of Catalonia (ICIQ) and Universitat Autònoma de Barcelona.

[§] UPR 8241 liée par conventions à l'Université Paul Sabatier et à l'Institut National Polytechnique de Toulouse.

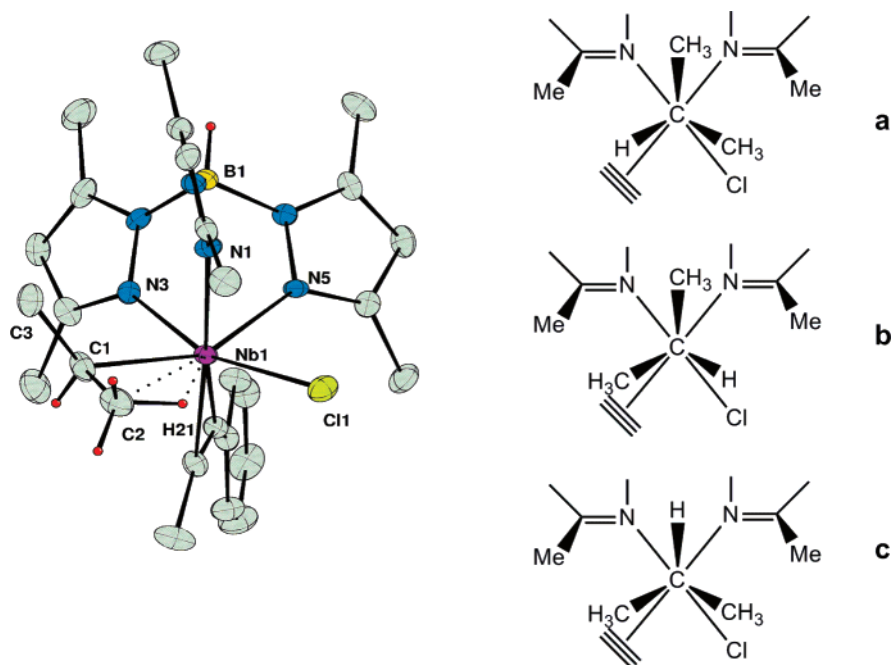
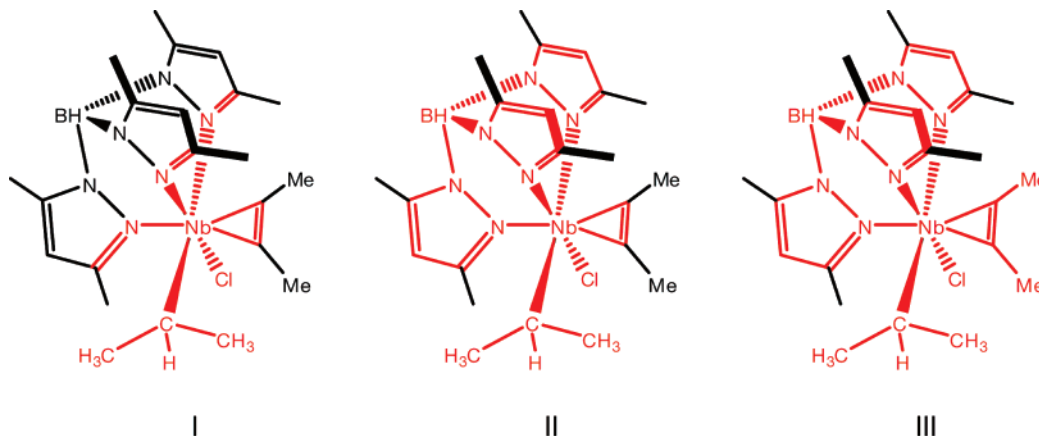


Figure 1. Crystal structure of $\text{Tp}^{\text{Me}_2}\text{NbCl}(\text{PhCCMe})\text{Pr}$ (rotamer a) and the relationship between the rotamers (a)–(c) (viewed along the Nb–C $_{\alpha}$ bond).

Scheme 1. QM/MM Partitions in Models I–III^a



^a The QM region is shown in red, MM in black.

deficient agostic and nonagostic C–H bonds, respectively). In light of the varied electronic mechanisms that can potentially stabilize an agostic bond, it is perhaps unsurprising that many authors have settled for a phenomenological definition based on structure rather than a specific electronic mechanism. Thus the combination of a short metal–hydrogen separation, an elongated C–H bond and relatively small angles in the M–C–C–H unit, are generally taken as indicative of an agostic structure, whatever its origin.

Over the past 8 years we have published a number of papers where we have adopted this approach to probe the agostic bonding in a series of niobium–alkyl complexes, $\text{Tp}^{\text{Me}_2}\text{NbCl}(\text{R}'\text{CCR}'')(\text{R})$ (Figure 1).^{11–15} The facial $\text{Tp}^{\text{Me}_2}\text{M}$ fragment provides a unique platform to investigate the nature of agostic interactions, because the pendant methyl groups define a tight steric pocket on the opposite face of the metal and so restrict rotation about the Nb–C $_{\alpha}$ bond. As a result, isomers differing only in the torsion angles about this Nb–C $_{\alpha}$ bond form distinct minima on the potential energy surface.

For example, in the case of the isopropyl complex (Figure 1), rotation of the alkyl moiety generates three distinct minima (a–c), two of which (a and b) are in equilibrium in solution.

In the work described in refs 11–15, we have employed a hybrid QM/MM approach (B3LYP:UFF), wherein the molecule is divided into a ‘core’, treated at the quantum mechanical level, and a ‘periphery’, described using the more tractable molecular mechanics protocol. In our original 1998 work, we defined the core as $[\text{Nb}(\text{Cl})(i\text{Pr})(\text{HCCH})(\text{NHCH}_2)_3]^+$, implying cuts through the N–N, C–C, and C–Me bonds of the Tp^{Me_2} ligand as well as the C–R bonds of the alkyne (Scheme 1, Model I). Such a dramatic simplification of the ligand is clearly far from ideal because replacing the anionic Tp^{Me_2} ligand with three neutral imine groups introduces a positive charge in the QM partition. Moreover, making cuts across bonds with aromatic character necessarily leads to localization of the π character in the core partition and hence underestimation of N=C bond lengths. Such gross simpli-

fications were, however, necessary and pragmatic given the computational resources available when we started this work in 1998, and we were greatly encouraged by the apparently good agreement between experiment and theory, both in terms of structural parameters and the energetic separation of the different minima.¹²

Very recently, we turned our attention to the dynamic exchange between the different agostic rotamers, a problem that requires the accurate computation of transition structures as well as the minima we have considered previously. The rapid advances in hardware since our initial work mean that a full QM treatment of the whole system is now computationally tractable, and, given the deficiencies of the original QM/MM protocol highlighted in the previous paragraph, we decided that the time was right to adopt a full QM approach. As a starting point, we naturally reoptimized the equilibrium structures of the various minima using our apparently improved full QM protocol, expecting to see an even closer agreement with experimental structures and energies. To our great surprise, we discovered that the agreement was instead dramatically worse: in fact, a full QM calculation using the B3LYP functional and basis sets similar to those employed in our previous hybrid approach completely failed to locate an agostic minimum of any kind.

In view of this apparent failure of the B3LYP functional, we now report a systematic survey of density functionals, with the aim of establishing a suitable protocol for describing the agostic bonds in these systems. Although the performance of different density functionals is well documented in a variety of chemical contexts,¹⁶ it seems that no systematic evaluation of their performance has been reported for agostic interactions. In this contribution, we show that only those functionals with correlation parts that obey the Uniform Electron Gas (UEG) limit lead to minima with an agostic structure. Moreover, these functionals reproduce certain structural features (bond lengths and angles) that were, in hindsight, rather poorly represented in our original B3LYP:UFF calculations. Ultimately, we conclude that functionals which do not obey the UEG limit fail completely to describe the balance between agostic and Nb–Cl π bonding in these systems.

Computational Methodology

Basis Sets. In view of the weak nature of agostic interactions it is important to address adequately the issue of basis set selection from the outset. Practical limitations arise due to the large size of the system, so we sought to ensure a high quality description of the most critical features while remaining within the limits of computational feasibility. We therefore chose the valence triple- ζ polarized basis sets of Ahlrichs (TZVP) for all the atoms of the alkyl moiety as well as Cl¹⁷ and valence double- ζ polarized (SVP) basis sets for all atoms of the alkyne ligand and the Tp^{Me2} backbone, with unpolarized SV basis sets for Tp^{Me2} substituents.¹⁸ Niobium was described with the [6s5p3d] SDD valence basis set and the quasi-relativistic ECP28MWB effective core potential of Andrae et al.¹⁹

QM/MM Calculations. Hybrid quantum mechanical/molecular mechanical calculations were carried out with

the ONIOM method,²⁰ using the B3LYP functional^{21–23} for the quantum mechanics partition and the UFF force field²⁴ for the molecular mechanics partition. Basis sets for the QM region were identical to those described above. Details regarding the different partitioning schemes used are described in detail in the text. All ONIOM calculations used microiterations²⁵ for the optimization procedure and did not employ electrostatic embedding.

DFT Methods. Twenty-four density functionals were included in our study, sampled from all current DFT implementations. LSDA was evaluated in the form of the SVWN5 functional.^{26,27} GGA functionals tested include the nonempirical PBE functional of Perdew, Burke, and Ernzerhof;²⁸ the BP86 functional incorporating Becke (B88) exchange²⁹ and Perdew correlation;³⁰ two functionals based on Perdew–Wang 1991 correlation,³¹ BPW91 and mPWPW91;³² and three functionals with the Lee–Yang–Parr expression for correlation,²³ BLYP,²⁹ OLYP,³³ and G96LYP.³⁴ Also from the GGA family we assessed the highly parametrized HCTH/147 and HCTH/407 modifications^{35,36} of the Hamprecht–Cohen–Tozer–Handy functional,³⁷ itself an elaboration of Beckes’s 1997 10-parameter functional form.³⁸ Hybrid GGA functionals included in our study were B3LYP (20% HF exchange),^{21–23} B3PW91,^{21,22,31} O3LYP (11.61% HF exchange),^{23,39} X3LYP (21.8% HF exchange),^{23,40} mPW1PW91 (25% HF exchange),^{31,32} mPW1LYP,^{23,32} and PBE1PBE (25% HF exchange),²⁸ and also two functionals based on modifications of Becke’s B97³⁸ expression, B97-2⁴¹ (21% HF exchange) and B98⁴² (21.98% HF exchange), were included. From the field of meta-GGA (τ -dependent) functionals we have tested the VSXC functional of van Voorhis and Scuseria⁴³ and the nonempirical functional of Tao, Perdew, Staroverov, and Scuseria (TPSS).⁴⁴ Last, from the hybrid-meta-GGA family we assessed Becke’s B1B95 functional⁴⁵ (28% HF exchange), mPW1B95 (25% HF exchange), and TPSSh⁴⁶ (10% exact exchange).

We use the isopropyl complex, Tp^{Me2}NbCl(PhCCMe)*i*Pr, where structural data are available for the dominant β -C–H agostic isomer (Figure 1), as a test case for this study. The phenyl group of the alkyne was replaced by a methyl in the computational model in order to reduce computational cost; test calculations confirm that this simplification has negligible impact on the optimized structural parameters. Full geometry optimizations with no restrictions were carried out for each density functional. Calculations were initiated from either the experimental structure or from previously optimized geometries and were always allowed to proceed to convergence, even when initial divergence indicated the absence of a corresponding stationary point for a specific conformation. Optimized structures were confirmed to be genuine minima by analytic calculation of their harmonic vibrational frequencies. All calculations were performed with the Gaussian03 series of programs, Revision D.02.⁴⁷

Results

In previous contributions we modeled the agostic complexes with QM/MM calculations of the IMOMM(B3LYP:UFF) type. As a first step in this systematic investigation, we have repeated these calculations using the same [Nb(Cl)(*i*Pr)-

Table 1. Optimized QM/MM (Models I–III) and Full QM (Model IV) Structural Parameters^c

	NbC _α C _β	Nb–C _β	Nb–H _β	NbC _α C _β H _β	C _β –H _α	C _β –H' ^b	Nb–C _α	C _α –C _β	C _α –C' _β	Nb–Cl
expt	87.0(3)	2.608(4)	2.17(5)	2.4	1.11(5)		2.228(4)	1.476(7)	1.535(6)	2.493(1)
B3LYP:UFF										
Model I	90.7	2.734	2.366	1.0	1.110	1.091	2.250	1.528	1.524	2.428
Model II	107.7	3.126	2.976	14.2	1.092	1.093	2.282	1.552	1.531	2.423
Model III	110.1	3.177	3.087	25.0	1.090	1.094	2.291	1.549	1.532	2.430
Model IV	109.3	3.148	3.226	52.2	1.089	1.094	2.281	1.543	1.532	2.440
PBE1PBE:UFF										
Model I	86.5	2.609	2.189	1.3	1.125	1.091	2.224	1.507	1.514	2.424
Model II	88.3	2.648	2.242	1.1	1.119	1.092	2.219	1.512	1.515	2.457
Model III	87.9	2.641	2.232	1.2	1.120	1.092	2.220	1.513	1.515	2.470
Model IV	87.2	2.620	2.204	3.5	1.123	1.091	2.217	1.511	1.514	2.494
B3LYP:UFF on Zirconium Analogue										
Model I	110.1	3.266	3.218	40.0	1.091	1.097	2.399	1.540	1.532	2.529

^a Bond length of agostic C–H bond or of C–H bond with the shortest Nb–H distance. ^b Mean value of the other two C–H bond lengths. ^c Angles in degrees, distances in Å.

(HCCH)(NHCH₂)₃⁺ ‘core’ but employing the larger basis sets described above and following the ONIOM(B3LYP:UFF) protocol (Scheme 1, Model I and Table 1). The QM region was then expanded in three steps to include the following: (i) the whole Tp backbone but not the methyl groups (Model II), (ii) the Tp backbone and the substituents on the alkyne (Model III), and (iii) the entire molecule (Model IV). The results for Model I are very similar to those reported in our original work, where the partition was identical but the basis set rather more limited. When we originally published these results, we were encouraged by the fact that the gross geometric features of the agostic unit were reproduced with reasonable accuracy with even this small core. A careful examination of bond lengths and angles, however, reveals a number of discrepancies, all of which can be traced to the partitioning of the system into a QM and an MM region. For example, the optimized N–C bond lengths of 1.28 Å are some 0.06 Å shorter than in the crystal structure—a direct consequence of the loss of aromatic character—while the C–C–C bond angles in the alkyne moiety are overestimated by approximately 10°. More subtly, the two C–C bond lengths are almost identical, whereas in the crystal structure the C–C bond in the agostic position (C_α–C_β) is somewhat shorter than the other (C_α–C'_β). Finally, but most importantly, the optimized Nb–Cl bond length is some 0.07 Å shorter than the experimental value, suggesting that Nb–Cl π bonding is overestimated in this case. Taken as a whole, these results imply a somewhat imbalanced description of the electronic structure within the Nb coordination sphere.

Incorporation of the Tp backbone and the alkyne substituents into the QM partition (Models II and III) eliminates the inaccuracies in the pyrazolyl N–C bond lengths (optimized values 1.34 Å) and alkyne angles. Surprisingly, however, these improvements come at the expense of the agostic interaction, which disappears completely: the agostic C–H bond length decreases from 1.110 Å (Model I) to 1.090 Å (Model III), while the Nb–C_β distance increases from 2.734 Å to 3.177 Å. These changes are accompanied by rotation of the agostic methyl group which removes the hydrogen atom from the agostic plane. The full B3LYP

calculation (Model IV) also fails to reproduce the agostic structure, converging instead to an anagostic minimum very similar to that of Model III, with an Nb–C_β distance of 3.128 Å and an NbC_αC_βH_β dihedral of 52.2°. We emphasize that an agostic stationary point proved impossible to locate even after extensive sampling of the potential energy surface in the vicinity of the agostic structure. It should also be stressed that this result was unaffected by basis set extension (the Martin–Sundermann (2fg) polarization set⁴⁸ for Nb and a QZVP basis set for *i*Pr gave a similar structure), so we conclude that the source of this surprising failure is the density functional.

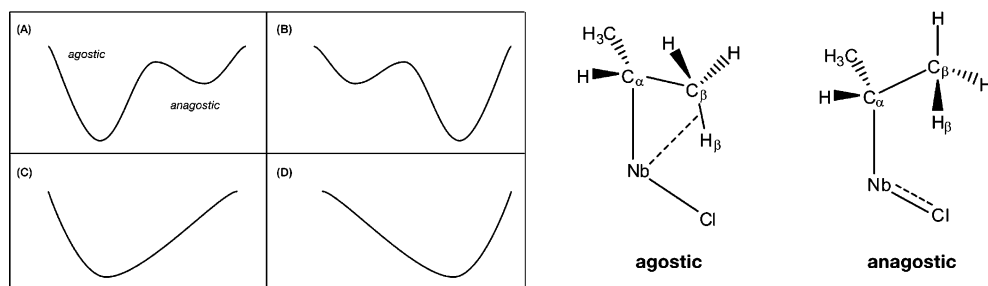
In order to obtain insight into the origin of this failure we undertook a systematic comparison of 24 distinct density functionals. Optimized structural parameters for all 24 are collected in Table 2. The functionals clearly fall into four distinct groups (A–D), based on the shape of the resultant potential energy surface (Figure 2). There are 11 density functionals that identify both an agostic and an anagostic minimum on the potential energy surface, six of which predict the agostic minimum to be more stable (group A: PBE1PBE, PBEPBE, B1B95, mPW1B95, TPSS, and TPSSh), while the other five (group B: BP86, BPW91, mPWPW91, B3PW91, and mPW1PW91) favor the anagostic structure, albeit only marginally. In the third group (C) we find two functionals (VSXC and SVWN5) that identify the agostic minimum but do not locate an anagostic alternative. The 11 functionals in the fourth group (D) predict the anagostic structure to be the unique minimum on the potential energy surface, failing completely to identify the experimentally observed agostic geometry. These functionals either incorporate the Lee–Yang–Parr (LYP) correlation or are based on modifications and extensions of the multiple-coefficient B97 functional (B98, B97-2, HCTH/147, and HCTH/407).

With the exception of SVWN5, the key optimized parameters for the agostic structure are very consistent among the first three groups of functionals: the agostic C_β–H bonds are significantly elongated in all cases (1.12–1.13 Å), while the Nb–C_β distances lie in the range 2.58–2.67 Å, in excellent agreement with the experimental value of 2.608–(4) Å. The SVWN5 results are qualitatively correct, but the

Table 2. Structural Parameters of Agostic and Anagostic Minima^c

	agostic						anagostic						ΔE	
	NbC _α C _β	Nb–C _β	Nb–H _β	C _β –H ^a	C _β –H' ^b	Nb–Cl	NbC _β C _β	Nb–C _β	Nb–H _β	C _β –H ^a	C _β –H' ^b	Nb–Cl		
expt	87.0(3)	2.608(4)	2.17(5)	1.11(5)		2.493(1)								
A PBE1PBE	87.2	2.620	2.204	1.123	1.091	2.494	107.5	3.082	3.138	1.090	1.095	2.419	0.4	
PBEPBE	86.8	2.640	2.213	1.131	1.098	2.502	106.5	3.089	3.112	1.098	1.102	2.428	0.8	
B1B95	85.6	2.581	2.145	1.124	1.087	2.494	104.7	3.017	3.050	1.088	1.092	2.416	7.5	
mPW1B95	85.7	2.581	2.149	1.123	1.086	2.490	104.6	3.009	3.048	1.087	1.090	2.415	8.3	
TPSS	86.3	2.634	2.189	1.129	1.093	2.502	107.0	3.106	3.129	1.093	1.097	2.429	2.7	
TPSSh	86.5	2.626	2.188	1.126	1.091	2.499	107.5	3.103	3.144	1.091	1.095	2.425	2.4	
B BP86	87.2	2.656	2.230	1.130	1.099	2.504	107.2	3.111	3.166	1.098	1.103	2.431	–0.4	
BPW91	87.3	2.658	2.234	1.128	1.097	2.503	107.4	3.115	3.160	1.096	1.101	2.428	–1.2	
B3PW91	87.9	2.647	2.238	1.120	1.090	2.496	108.0	3.102	3.152	1.090	1.095	2.422	–1.6	
mPWPW91	87.1	2.650	2.225	1.128	1.096	2.503	107.1	3.105	3.156	1.096	1.101	2.428	–0.3	
mPW1PW91	87.5	2.631	2.219	1.120	1.089	2.493	107.7	3.089	3.144	1.089	1.093	2.419	–0.6	
C VSXC	88.3	2.668	2.271	1.122	1.095	2.499								
SVWN5	83.4	2.524	2.063	1.150	1.101	2.466								
D B3LYP							109.3	3.148	3.226	1.089	1.094	2.440		
O3LYP							109.7	3.157	3.215	1.088	1.094	2.419		
X3LYP							109.2	3.143	3.226	1.089	1.094	2.440		
mPW1LYP							109.5	3.149	3.235	1.088	1.093	2.443		
BLYP							109.1	3.178	3.246	1.095	1.101	2.453		
OLYP							110.0	3.179	3.234	1.091	1.097	2.419		
G96LYP							109.1	3.172	3.223	1.094	1.100	2.445		
B98							109.0	3.135	3.209	1.091	1.095	2.434		
B97–2							108.7	3.120	3.184	1.088	1.093	2.421		
HCTH/147							109.0	3.151	3.210	1.091	1.096	2.424		
HCTH/407							109.5	3.162	3.225	1.089	1.095	2.419		

^a Bond length of agostic C–H bond or of C–H bond with the shortest Nb–H distance. ^b Mean value of the other two C–H bond lengths. ^c Angles in degrees, distances in Å, relative energies in kJ mol^{–1}.

**Figure 2.** Schematic depiction of the topology of the potential energy for the four different groups of density functionals (A–D).

optimized parameters suggest that pure LSDA overestimates the strength of the agostic interaction. VSXC also identifies the agostic structure as the sole stationary point on the PES, but in this case the structure agrees more closely with both the experimental data and the predictions of the other functionals in groups A and B. The optimized structural parameters for the anagostic structure are also quite similar across groups A, B, and D: the NbC_αC_β angle is always wider than 104.5°, and in the case of group D functionals it is consistently between 109° and 110°. Comparison of the agostic and anagostic geometries for groups A and B reveals that the Nb–C_β distance is about 0.5 Å longer in the anagostic isomer. The terminal methyl group in the anagostic structure is also rotated along the C_α–C_β axis so as to place the closest H_β atom a further 0.9 Å away from the metal center. Most significantly, the Nb–Cl bond length is always about 0.07 Å shorter in the anagostic structure, suggesting that π -donation from the chloride ligand competes with the

agostic bond for vacant orbital space on the Nb center. A Natural Bond Orbital (NBO)⁴⁹ analysis of the agostic and anagostic structures (PBE1PBE) clearly illustrates this competition. The dominant donor–acceptor interaction between the C_β–H σ bonding orbital and a formally empty Nb d orbital (99.6 kJ mol^{–1}) in the agostic structure disappears almost completely at the anagostic minimum (7.1 kJ mol^{–1}) and is replaced by an interaction of similar magnitude between a chloride lone pair and the same Nb d orbital (58.6 kJ mol^{–1}).

The direct competition between agostic interactions and Nb–Cl π bonding means that the Nb–Cl bond length provides a sensitive probe of the strength of the agostic bond: a short Nb–Cl bond in the region of 2.42 Å indicates weak (or nonexistent) agostic bonding, while a bond length around 2.49 Å is more consistent with a significant attraction between the metal and the C–H bond. This conclusion further increases our uneasiness about the small QM/MM

partition (Model I) used in our previous work, where we noted that the optimized Nb–Cl bond length was precisely 0.07 Å shorter than experiment. The optimized structure of the isopropyl complex (Table 1) therefore appears to feature *both* a strong agostic interaction (C–H 1.110 Å, Nb–H_β = 2.366 Å) and a strong Nb–Cl π bond (2.428 Å), despite the fact that the two electron pairs (C–H σ and Cl lone pair) are competing for a single vacant orbital on the metal.¹¹ We will return to this issue in our concluding remarks.

Close inspection of the members in each group of functionals reveals certain regularities that lead us to an understanding of the decisive factor governing the performance of the DFT methods. First of all, we see that all functionals that incorporate the Lee, Yang, and Parr (LYP) correlation²³ belong to group D and therefore fail to locate the agostic minimum. It is clear from Table 2 that the performance of LYP-containing methods remains the same irrespective of the exchange functional (pure GGA or hybrid) and regardless of the percent of exact exchange in the latter. The minimal effect of the exchange functional is also apparent in the other groups: for example, there is little difference between BPW91 and B3PW91 or between TPSS and TPSSH. The poor performance of the group D functionals can be traced to the approximations used to generate the correlation functionals. In the local density approximation (LDA) the exchange and correlation energies of a system at a given point in space are assumed to be those of a homogeneous electron gas of uniform density at that point. Consequently, LDA provides the correct results for uniform densities, or, in other words, it satisfies the uniform electron gas (UEG) limit. It also performs well for slowly varying densities, but in cases of more rapid density variations which are typical of molecular systems it usually underestimates the exchange energy and significantly overestimates the correlation energy. This deficiency is addressed by the generalized gradient approximation (GGA) functionals, which are usually constructed as corrections of the LDA with terms that depend on the gradients of the density. What distinguishes LYP from other correlation functionals of this kind is that it was not formulated as a correction of LDA but was instead constructed by recasting the Colle–Salvetti correlation energy formula of the helium atom⁵⁰ in terms of gradient expansions. Thus, LYP does not obey the uniform electron gas (UEG) limit. The other four members of group D share the same feature: B98 and B97-2 as well as the HCTH functionals are all based on the B97 functional, which does not satisfy the UEG limit. Conversely, the correlation parts of density functionals that successfully identify an agostic minimum (groups A, B, and C) do obey the UEG condition. These include the Perdew family of P86, PW91, and PBE GGA functionals as well as the meta-GGA correlation functionals B95 and TPSS. The only apparent exception to the rule regarding the UEG limit requirement is the VSXC functional, which, along with SVWN5, is the only one to locate only the agostic minimum. VSXC was developed on the basis of the density matrix expansion to model the exchange-correlation hole and in order to improve performance for molecular systems the UEG constraint was relaxed by reducing the LDA coefficients to 70% for

opposite-spin and 33% for same-spin correlation. Despite that, VSXC manages to reproduce the agostic structure quite accurately, yielding only a marginally wider agostic angle compared to the other functionals. Therefore, the failure of group D functionals to predict a stationary point corresponding to the agostic structure can be traced to their divergence from the uniform electron gas (UEG) limit. We note that Schaeffer and co-workers have recently reported similar trends in a study of Ag₃ and Ag₄ clusters, where non-UEG functionals also yield poor results.⁵¹ The nature of the bonding in these silver clusters is clearly rather different from the agostic interactions that are the focus of this study, suggesting that the choice of correlation functional may have important implications in a wider variety of chemical contexts.

A complementary perspective on the role of correlation can be obtained from wave function-based *ab initio* approaches. HF optimizations locate only the anagostic structure ($\angle \text{NbC}_\alpha\text{C}_\beta = 112.7^\circ$, Nb–C_β = 3.205 Å). Geometry optimizations at the MP2 level were not computationally feasible, so we chose to perform the comparison using single-point energy calculations on typical agostic and anagostic structures (optimized using the PBE1PBE functional). At the HF level the anagostic structure is 30.1 kJ mol⁻¹ lower in energy than the agostic one, but the order is reversed at the MP2 level: the agostic structure is estimated to be 11.0 kJ mol⁻¹ more stable than the anagostic one, an energy difference comparable to that predicted by the B95-based hybrid meta-GGA functionals. This stabilization of the agostic structure by 40 kJ mol⁻¹ at the MP2 level emphasizes the importance of a correct description of dynamic correlation.

Having established that non-UEG correlation functionals are poorly suited to describing the agostic interactions in this class of systems, we can now return to our original hybrid QM/MM results and ask why we obtained apparently good results using this functional in combination with the highly simplified Model I partition. We have noted previously that a close inspection of the optimized structure of Model I (Table 1) reveals inconsistencies, the most serious being the short Nb–Cl bond, which effectively blocks donation of electron density from the C–H bond to the metal. So why then do we also observe a short Nb–H_β separation, generally considered to be characteristic of an agostic interaction? The answer lies in the simplifications implicit in our Model I QM/MM partition and, in particular, the presence of a positive charge in the QM region. This introduces an unphysical electrostatic component to the agostic interaction, which is sufficient to hold the C–H bond in the vicinity of the metal center even in the absence of significant charge transfer between the C–H bond and the Nb center. The electrostatic nature of the agostic interaction in this case is reflected in the agostic C–H bond length of 1.110 Å which, although longer than its nonagostic counterpart (1.091 Å), is still significantly shorter than those predicted by the group A functionals (1.12–1.13 Å). The impact of the spurious positive charge in the QM region is clearly illustrated by a further series of calculations on Models I–IV using the PBE1PBE functional (Table 1) where, in contrast to B3LYP,

the functional *does* provide a good description of the agostic bond at the full QM limit (Model IV). The competition between agostic and Nb–Cl π bonding for vacant orbital space is apparent in the trends in C $_{\beta}$ –H $_{\beta}$, Nb–H $_{\beta}$, and Nb–Cl bond lengths in Models IV–II, all of which have a neutral QM region: the contraction of the Nb–Cl bond (2.494–2.457 Å) is accompanied by an increase in Nb–H $_{\beta}$ (2.204–2.242 Å) and a decrease in C $_{\beta}$ –H $_{\beta}$ (1.123–1.119 Å). The Nb–Cl bond length for Model I is even shorter, at 2.424 Å, and is typical of an anagostic structure. On the basis of the competition for vacant orbital space on Nb that we have emphasized above, we would therefore anticipate a further contraction of C $_{\beta}$ –H $_{\beta}$ in Model I, along with an increase in Nb–H $_{\beta}$. In fact, precisely the opposite is observed: the C $_{\beta}$ –H $_{\beta}$ bond length increases to 1.125 Å while the Nb–H $_{\beta}$ bond length contracts to 2.189 Å. This discontinuity in structural trends provides clear evidence that the electronic mechanism responsible for holding the C–H bond close to the metal center is fundamentally different in Model I and Models II–IV. As a final confirmation of the critical role of the positive charge, we have optimized the structure of the isoelectronic zirconium complex, [Tp^{Me2}ZrCl(MeCCMe)*i*Pr][–], using the Model I partition, where the QM region is now neutral rather than cationic. At the (B3LYP:UFF) level, this gives an anagostic minimum, quite distinct from the Model I niobium analogue (Table 1). We therefore conclude that the apparent success of our early B3LYP:UFF hybrid calculations was a result of the presence of a positive charge into the QM partition, which introduces an unrealistic electrostatic component to the agostic interaction. This additional attractive component compensates for the intrinsically imbalanced description of agostic and Nb–Cl π bonding afforded by this and other non-UEG functionals and leads to structures in qualitative agreement with experiment. Only a very close inspection of the niobium coordination sphere reveals the telltale signs of error compensation.

Conclusions

The calculations reported in this paper suggest that functionals that obey the UEG correlation limit provide an accurate picture of the subtle balance between Nb–C–H agostic and Nb–Cl π bonding in [Tp^{Me2}NbCl(MeCCMe)*i*Pr] and so give excellent structure predictions. In contrast, functionals that do not obey the UEG limit appear to overestimate the strength of the Nb–Cl π bond relative to the agostic Nb–C–H, leading to anagostic structures, at odds with experiment. The popular B3LYP functional falls into the second category, and so these results serve as a warning that this functional may be inappropriate, at least in cases where a competition between agostic and π -donor interactions for vacant orbital space is important. Somewhat surprisingly, an agostic structure is recovered by the B3LYP functional when the system is greatly simplified using the QM/MM methodology but only in cases where a positive charge is present in the QM region. The positive charge introduces an additional but unphysical electrostatic component to the agostic interaction which holds the C–H bond close to the metal even though strong Nb–Cl π bonding effectively blocks any sharing of electron density. The net

result is an apparently good agreement with experiment, but only if we restrict our attention to the gross features of the alkyl chain. Our experience serves as a clear warning that the optimization of an ‘agostic structure’ in apparently good agreement with experiment does not necessarily mean that the chosen theoretical method has captured the true nature of the ‘agostic bond’. It may instead simply reflect the fact that distortion of the angles within the alkyl group is relatively easy, and so a wide range of physical mechanisms can induce the bending regarded as typical of an agostic structure. The true nature of the agostic interaction is, however, revealed by the more subtle features of the metal coordination sphere, in this case the Nb–Cl bond length, which acts as a sensitive indicator of the presence or absence of shared electron density between the C–H bond and the metal center.

References

- (1) Brookhart, M.; Green, M. L. H. *J. Organomet. Chem.* **1983**, *250*, 395.
- (2) Brookhart, M.; Green, M. L. H.; Wong, L. L. *Prog. Inorg. Chem.* **1988**, *36*, 1.
- (3) Clot, E.; Eisenstein, O. *Struct. Bonding* **2004**, *113*, 1.
- (4) (a) Cotton, F. A.; Stanislawski, A. G. *J. Am. Chem. Soc.* **1974**, *96*, 754. (b) Cotton, F. A. *Inorg. Chem.* **2002**, *41*, 643.
- (5) (a) Eisenstein, O.; Jean, Y. *J. Am. Chem. Soc.* **1985**, *107*, 1177. (b) Demolliens, A.; Jean, Y.; Eisenstein, O. *Organometallics* **1986**, *5*, 1457.
- (6) Popelier, P. L. A.; Logothetis, G. *J. Organomet. Chem.* **1998**, *555*, 101.
- (7) Scherer, W.; McGrady, G. S. *Angew. Chem., Int. Ed.* **2004**, *43*, 1782.
- (8) (a) Haaland, A.; Scherer, W.; Ruud, K.; McGrady, G. S.; Downs, A. J.; Swang, O. *J. Am. Chem. Soc.* **1998**, *120*, 3762. (b) Scherer, W.; Priermeier, T.; Haaland, A.; Volden, H. V.; McGrady, G. S.; Downs, A. J.; Boese, R.; Bläser, D. *Organometallics* **1998**, *17*, 4406. (c) Scherer, W.; Hieringer, W.; Spiegler, M.; Sirsch, P.; McGrady, G. S.; Downs, A. J.; Haaland, A.; Pedersen, B. *Chem. Commun.* **1998**, 2471.
- (9) (a) Butts, M. D.; Bryan, J. C.; Luo, X.-L.; Kubas, G. *Inorg. Chem.* **1997**, *36*, 3341. (b) Nikonov, G. I.; Mountford, P.; Ignatov, S. K.; Green, J. C.; Leech, M. A.; Kuzmina, G. L.; Razuvaev, A. G.; Rees, N. H.; Blake, A. J.; Howard, J. A. K.; Lemenovskii, D. A. *J. Chem. Soc., Dalton Trans.* **2001**, 2903.
- (10) (a) Ziegler, T.; Tschinke, V.; Becke, A. *J. Am. Chem. Soc.* **1987**, *109*, 1351. (b) Han, Y.; Deng, L.; Ziegler, T. *J. Am. Chem. Soc.* **1997**, *119*, 5939.
- (11) Jaffart, J.; Mathieu, R.; Etienne, M.; McGrady, J. E.; Eisenstein, O.; Maseras, F. *J. Chem. Soc., Chem. Commun.* **1998**, 2011.
- (12) Jaffart, J.; Etienne, M.; Maseras, F.; McGrady, J. E.; Eisenstein, O. *J. Am. Chem. Soc.* **2001**, *123*, 6000.
- (13) Jaffart, J.; Etienne, M.; Reinhold, M.; McGrady, J. E.; Maseras, F. *Chem. Commun.* **2003**, 876.
- (14) Jaffart, J.; Cole, M. L.; Etienne, M.; Reinhold, M.; McGrady, J. E.; Maseras, F. *Dalton Trans.* **2003**, 4057.
- (15) Besora, M.; Maseras, F.; McGrady, J. E.; Oulié, P.; Dinh, D. H.; Duhayon, C.; Etienne, M. *Dalton Trans.* **2006**, 2362.

- (16) (a) Riley, K. E.; Op't Holt, B. T.; Merz, K. M., Jr. *J. Chem. Theory Comput.* **2007**, *3*, 407. (b) Jacquemin, D.; Femenias, A.; Chermette, H.; Ciofini, I.; Adamo, C.; André, J.-M.; Perpète, E. A. *J. Phys. Chem. A* **2006**, *110*, 5952. (c) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *124*, 224105. (d) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 415. (e) Brothers, E. N.; Merz, K. M., Jr. *J. Phys. Chem. A* **2004**, *108*, 2904. (f) Raymond, K. S.; Wheeler, R. A. *J. Comput. Chem.* **1999**, *20*, 207. (g) Ghosh, A.; Taylor, P. R. *Curr. Opin. Chem. Biol.* **2003**, *7*, 113. (h) Holthausen, M. C. *J. Comput. Chem.* **2005**, *26*, 1505. (i) Buhl, M.; Kabrede, H. *J. Chem. Theory Comput.* **2006**, *2*, 1282. (j) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 11127.
- (17) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.
- (18) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571.
- (19) Andrae, D.; Haeussermann, U.; Dolg, M.; Stoll, H.; Preuss, H. *Theor. Chim. Acta* **1990**, *77*, 123.
- (20) (a) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170. (b) Dapprich, S.; Komaromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *J. Mol. Struct. (Theochem)* **1999**, *461*, 1.
- (21) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (22) Stevens, P. J.; Devlin, J. F.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.
- (23) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (24) Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A., III; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024.
- (25) Vreven, T.; Morokuma, K.; Farkas, Ö.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, *24*, 760.
- (26) Slater, J. C. *Quantum Theory of Molecules and Solids, Vol. 4: The Self-Consistent Field for Molecules and Solids*; McGraw-Hill: New York, 1974; pp 12–79.
- (27) Vosko, S. J.; Wilk, L.; Nusair, M. *Can. J. Chem.* **1980**, *58*, 1200.
- (28) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (29) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (30) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.
- (31) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244.
- (32) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.
- (33) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403.
- (34) Gill, P. M. W. *Mol. Phys.* **1996**, *89*, 433.
- (35) Boese, A. D.; Doltsinis, N. L.; Handy, N. C.; Sprik, M. J. *Chem. Phys.* **2000**, *112*, 1670.
- (36) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2001**, *114*, 5497.
- (37) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264.
- (38) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554.
- (39) Cohen, A. J.; Handy, N. C. *Mol. Phys.* **2001**, *99*, 607.
- (40) Xu, X.; Goddard, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673.
- (41) Wilson, P. J.; Bradley, T. J.; Tozer, D. J. *J. Chem. Phys.* **2001**, *115*, 9233.
- (42) Schmider, H. L.; Becke, A. D. *J. Chem. Phys.* **1998**, *108*, 9624.
- (43) van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400.
- (44) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (45) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040.
- (46) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129.
- (47) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision D.02*; Gaussian, Inc.: Wallingford, CT, 2004.
- (48) Martin, J. M. L.; Sundermann, A. *J. Chem. Phys.* **2001**, *114*, 3408.
- (49) (a) Weinhold, F.; Landis, C. *Valency and Bonding. A Natural Bond Orbital Donor-Acceptor Perspective*; Cambridge University Press: Cambridge, U.K., 2005; pp 45–86. (b) Glendening, E. D.; Badenhoop, J. K.; Reed, A. E.; Carpenter, J. E.; Bohmann, J. A.; Morales, C. M.; Weinhold, F. *NBO 5. G*; Theoretical Chemistry Institute, University of Wisconsin: Madison, WI, 2001.
- (50) Colle, R.; Salvetti, D. *Theor. Chim. Acta* **1975**, *37*, 329.
- (51) Zhao, S.; Li, Z.-H.; Wang, W.-N.; Liu, Z.-P.; Fan, K.-N.; Xie, Y.; Schaefer, H. F., III *J. Chem. Phys.* **2006**, *124*, 184102.

Kernel Energy Method: The Interaction Energy of the Collagen Triple Helix

Lulu Huang,[†] Lou Massa,[‡] and Jerome Karle^{*,†}

*Laboratory for the Structure of Matter, Naval Research Laboratory,
Washington, D.C. 20375-5341, and Hunter College and the Graduate School, City
University of New York, New York, New York 10021*

Received March 16, 2007

Abstract: There is a rapid growth in computational difficulty with the number of atoms when quantum mechanics is applied to the study of biological molecules. This difficulty may be alleviated in two different ways. One is the advance of parallel supercomputers. And the second is the use of a quantum crystallographic formalism based upon quantum kernels. The kernel methodology is well suited for parallel computation. Recently published articles have applied these advances to calculate the quantum mechanical ab initio molecular energy of peptides, protein (insulin), DNA, and RNA. The results were found to have high accuracy. This paper shows that it is possible to use the full power of ab initio quantum mechanics to calculate the interaction of long chain molecules of biological and medicinal interest. Such molecules may contain thousands or even tens of thousands of atoms. In the approach presented here the computational difficulty of representing a molecule increases only modestly with the number of atoms. The calculations are simplified by representing a full molecule by smaller “kernels” of atoms. The general case is illustrated by a specific example using an important protein, viz., a triple helix collagen molecule of known molecular structure. In order for such a molecule to be a stable helix, the overall interactions among the chains must be attractive. The results show that such interactions are accurately represented by application of the KEM to this triple helix.

I. Introduction

The Kernel Energy Method (KEM) calculates the quantum mechanical molecular energy by the use of the parts of a whole molecule, called kernels. The kernels are chosen to be much smaller than a full biological molecule. Thus the calculations of kernels and double kernels are in a practical way, doable. The kernel contributions are summed to obtain the energy of a whole molecule. In this way the task of calculating a quantum mechanical energy is simplified. Also, the computational time is much reduced. Previous work has shown that the accuracy obtained appears to be satisfactory.

The first applications of the KEM¹ referred to above involved a number of peptides. Good accuracy was retained throughout a wide range of basis functions and all of the most commonly used computational methods² that were studied. It was also found that good results were obtained in application of the KEM to the protein, insulin,³ and also to A, B, and Z DNA,⁴ RNA,⁵ and the rational design of drug.⁶ Theoretical background for the application of quantum mechanics to known molecular structures may be found in refs 7–14. References, that review the quantum mechanical methods related to computing the properties of large molecules from fragments, may be found in two articles referenced in this paper.^{7,9}

This paper combines a collagen molecule of given structure¹⁵ with quantum-mechanical KEM calculations to obtain the energies and interaction energies of a triple helix.

* Corresponding author e-mail: Jerome.karle@nrl.navy.mil.

[†] Naval Research Laboratory.

[‡] City University of New York.

It is knowledge of such energetics which allows one to understand the stability of known structures and the rational design of new protein interacting chains. It is shown that the kernel energy method accurately represents the energies and interaction energies of each of the chains separately and in combinations with one another. This is a challenging problem for the case of large molecular protein chains. But here the computational chemistry calculations are simplified, and the information derived from the atomic coordinates of the structure is enhanced by quantum mechanical information extracted there from. For the sake of completeness, the main ideas of the KEM are reviewed in the next section.

II. Review of the Kernel Energy Method

In the KEM, the knowledge of atomic coordinates is combined with quantum mechanics. Central to the KEM is the concept of the kernel. These are the quantum pieces which, when summed together, represent the whole molecule. Quantum calculations are carried out on kernels and double kernels only. All properties of the full molecule may be reconstructed from those of the kernels and double kernels. Given a known molecular structure, a molecule may be mathematically broken into tractable pieces called kernels, all of whose atomic coordinates are known. Each atom occurs in only one kernel. The total molecular energy is calculated in this paper by summation over the energy contributions of all double kernels reduced by those of any single kernels which have been over counted in the sum over double kernels.

In connection with the kernels and double kernels, we mention that in the KEM, the fragment calculations are carried out on double kernels and single kernels whose ruptured bonds have been mended by the attachment of H atoms. In the summation of energies the contribution of hydrogen atoms introduced to saturate the broken bonds tends to zero, on the assumption that the energy added by hydrogen atoms is transferable among the kernels and double kernels. The energy of the hydrogen atoms added to the double kernels effectively cancels that of the hydrogen atoms added to the pure single kernels, which enter with opposite sign. This cancellation of the mending hydrogen atom energy effects contributes to the accuracy achieved by the KEM.

The total energy is

$$E_{\text{total}} = \sum_{m=1}^{n-1} \left(\sum_{i=1}^{n-m} E_{ij} \right) - (n-2) \sum_{i=1}^n E_i \quad (1)$$

where E_{ij} = energy of a double kernel of name ij ; E_i = energy of a single kernel of name i ; i, j, m = running indices; and n = number of single kernels.

The validity of this approximation, in the case of a variety of peptides, proteins, DNA, RNA, and drug structures, has been shown in previous works.^{1,3-6} In this paper we depend upon the known ab initio accuracy of the KEM to show how it may be applied to a triple helix collagen molecule, 1A89,¹⁵ of given molecular structure, to obtain its relevant interaction energies, defined next.

III. The Interaction Energy

The definition of the interaction energy between any pair of kernels is

$$I_{ij} = E_{ij} - E_i - E_j \quad (2)$$

where the subscript indices name the pair of kernels in question, I_{ij} is the pair interaction energy, E_{ij} is the energy of a double kernel, and E_i and E_j are each the energies of a single kernel. The sign of the interaction energy, I_{ij} , indicates whether the kernels i and j attract (negative I) or repel (positive I). The total interaction energy is a sum of the pair interaction energies of the individual double kernels. The magnitude of a given pair interaction energy I_{ij} determines its relative importance to the total molecular interaction energy.

A generalization of eq 2 gives the interaction energy for a particular pair of molecular chains, a and b , as

$$I_{ab} = E_{ab} - E_a - E_b \quad (3)$$

where the subscript indices name the pair of protein chains in question, I_{ab} is the chain pair interaction energy, E_{ab} is the energy of a chain pair, and E_a and E_b are each the energies of a single protein chain. The sign of the interaction energy, I_{ab} , indicates whether the protein chains, a and b , attract (negative I) or repel (positive I). An equation analogous to eq 3 applies for the other chain pairs ac and bc .

The interaction energy among a triplet of protein chains is generalized to

$$I_{abc} = E_{abc} - (E_a + E_b + E_c) \quad (4)$$

where the subscript indices name the triplet of protein chains in question, I_{abc} is the triplet chain interaction energy, E_{abc} is the energy of a triplet of chains, and E_a , E_b , and E_c are each the energies of a single protein chain. Again importantly, the sign of the interaction energy, I_{abc} , indicates whether the triplet of protein chains a , b , and c altogether attract (negative I) or repel (positive I). The magnitude of the interaction energies flows naturally from implementation of the KEM. The KEM delivers the ab initio quantum mechanical interaction energy between and among protein chains. And, this may be envisioned to be computationally practical for molecular structures containing thousands or even tens of thousands of atoms.

IV. Collagen

Collagen is a protein, essential to the physical structure of the animal body. The molecule is made of three peptide chains forming a triple helix. These are incorporated in a vast number of ways to create structure. Collagen molecular cables provide strength in tendons, resilience to skin, support to internal organs, and a lattice structure to the minerals of bones and teeth. A repeated sequence of three amino acids forms the chains out of which the collagen triple helix is composed. Every third amino acid is glycine. Remaining positions in the chain often contain proline and hydroxyproline.

We have selected for study a particular collagen molecule whose molecular structure is known, 1A89,¹⁵ and whose

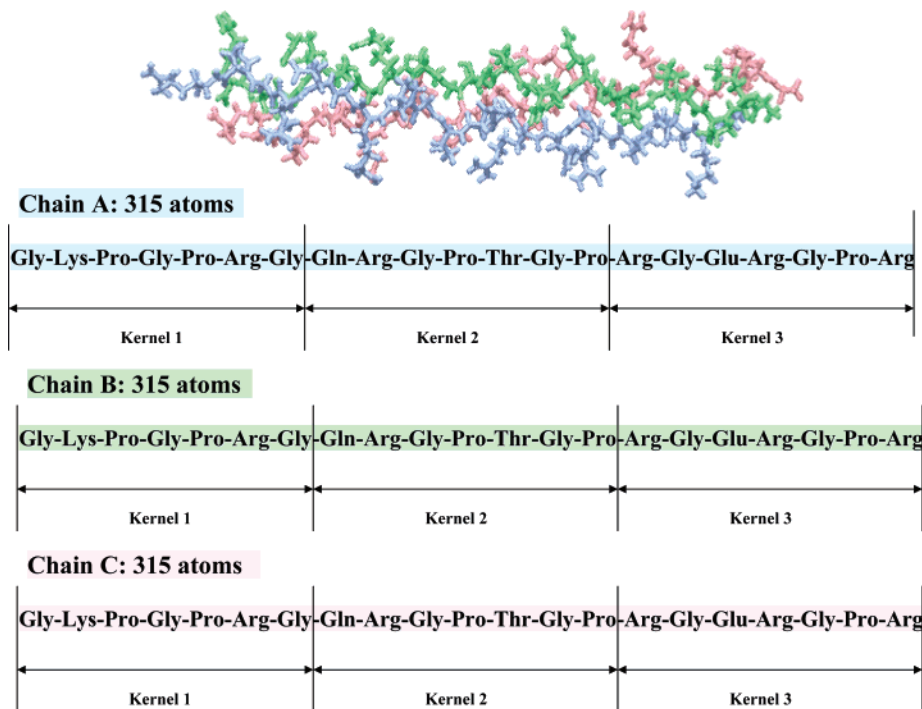


Figure 1. A picture of the collagen triple helix, 1A89, and the primary structure of each of its individual protein chains broken into kernels.

Table 1. Energy Calculations of Collagen Triple Helix by HF/STO-3G

chain	atoms	kernels	E_{HF} [au]	E_{KEM} [au]	$E_{\text{HF}} - E_{\text{KEM}}$ [au]	$E_{\text{HF}} - E_{\text{KEM}}$ [kcal/mol]
A	315	3	-7381.8557	-7381.8557	0.0000	0.0047
B	315	3	-7382.1621	-7382.1621	0.0000	0.0260
C	315	3	-7382.8332	-7382.8330	-0.0002	-0.1027
triple helix	945	9	-22146.9171	-22146.9112	-0.0059	-3.7332

atomic coordinates are readily available in the Protein Data Bank. The atomic coordinates are the starting information from which the KEM proceeds. From the structural role that collagen plays in the animal body, it is clear that it must be a stable molecule, with the chains of the triple helix structure adhering to one another. We apply the KEM to the molecular structure, 1A89, to see if the approximation is sufficiently accurate to reveal the expected adhesion of the collagen triple chains.

V. Results

Figure (1) shows a triple helix of protein chains that make up the collagen molecule under study. Also shown is the primary structure of the 3 identical protein chains that make up the triple helix, and each protein chain is broken into 3 kernels. The total triplex contains 945 atoms, each chain contains 315 atoms, with kernels 1, 2, and 3 containing 96, 98, and 121 atoms, respectively.

Table 1 contains the KEM calculations for each of the protein chains considered as a single entity. All calculations of this paper are of quantum mechanical Hartree Fock type, using an STO-3G limited basis of atomic orbitals. An “exact” result refers to the Hartree Fock calculation of an entire molecule, including all of its atoms together, without use of the kernel approximation. The KEM calculated energies are

meant to approximate the “exact” results. The difference between the two types of calculation are listed in both [au] and [kcal/mol]. One may conclude that the KEM calculation well represents the “exact” result. The percentage difference between the two types calculation is small. For the single chains A, B, and C the percentage differences are $1.0 \times 10^{-7}\%$, $5.6 \times 10^{-7}\%$, and $2.2 \times 10^{-6}\%$, respectively. Notice also that the percentage difference for the entire triple helix is only $2.7 \times 10^{-5}\%$. This level of accuracy accords with our previous experiences.¹⁻⁶

In Table 2 we list the calculation results for the triplex protein chains considered in pairs. The rows and columns are arranged as in Table 1, except that a new quantity, the interaction energy between the chains of the pairs, is also listed. As before the accuracy of the KEM energies is as expected, with differences for pairs AB, AC, and BC of approximately $2.6 \times 10^{-5}\%$, $2.2 \times 10^{-5}\%$, and $2.8 \times 10^{-5}\%$, respectively. Notice especially that not only do we obtain the chain pair interaction energies but also, as expected, the interaction is attractive.

In Table 3 we list the calculation results for the full triple helix of the collagen structure. As indicated above, the KEM result for the total energy is accurate. The HF and KEM interaction energies of the triple helix are also listed.

Table 2. Interaction Energy Calculations^a of Chain Pairs by HF/STO-3G

chains	atoms/ kernels	E_{HF} [au]	E_{KEM} [au]	I_{HF} [kcal/mol]	I_{KEM} [kcal/mol]	$I_{\text{HF}} - I_{\text{KEM}}$ [kcal/mol]
AB	630/6	-14764.0547	-14764.0508	-23.1488	-20.7075	-2.4413
AC	630/6	-14764.7100	-14764.7067	-13.2896	-11.2950	-1.9946
BC	630/6	-14765.0091	-14765.0050	-8.6151	-6.2123	-2.4028

^a Interaction energies, $I_{ab} = E_{ab} - E_a - E_b$.

Table 3. Interaction Energy Calculations^a of Collagen Triple Helix (945 Atoms and 9 Kernels) by HF/STO-3G

$E_{\text{HF}(abc)}$ [au]	$E_{\text{HF}(a+b+c)}$ [au]	$E_{\text{KEM}(abc)}$ [au]	$E_{\text{KEM}(a+b+c)}$ [au]	I_{HF} [kcal/mol]	I_{KEM} [kcal/mol]	$I_{\text{HF}} - I_{\text{KEM}}$ [kcal/mol]
-22146.9171	-22146.8510	-22146.9112	-22146.8508	-41.4778	-37.9010	-3.5768

^a Interaction energies, $I_{abc} = E_{abc} - E_{a+b+c}$, $E_{a+b+c} = E_a + E_b + E_c$.

VI. Discussion and Conclusions

A limited basis (STO-3G) was chosen simply to make the energy calculations as convenient as possible, for a protein structure of this size. Previous numerical experience has shown that the KEM can be applied to a wide variety of molecules with good accuracy, and such expectations were realized in this instance.

We have shown how to begin with a known molecular structure and obtain there from quantum mechanical information not otherwise known from the structure alone. With collagen, such information includes the energy of the individual protein chains and their combinations in pairs and as a triplex. Importantly, the interaction energy between chains of a pair or among those of a triplex are well represented, by the KEM. Notably, the KEM approximation is sufficiently accurate to reveal the expected adhesion which must prevail among the collagen triple chains. This forms the basis of an understanding of the structure of collagen in particular but more generally of a rational design of protein chain interactions.

The advantageous contribution which derives from the KEM is the interaction energy between and among protein chains when the molecular structure might contain tens of thousands of atoms. In such a case, if an ab initio quantum mechanical description of the interaction is to be obtained, then an approximation such as that of the KEM is indicated. Such calculations have typically been computationally impractical. The use of the KEM alleviates much of the computational difficulty by dividing a system into kernels, each smaller than the whole. Computations with each of the kernels can be assigned individually to separate nodes of a parallel processor. Thus, two advantages accrue to the KEM, since calculations are smaller and may be computed in parallel. The entire molecular structure is reconstituted from a sum over kernels. What has been shown by the calculations of this paper is that the KEM may be applied for purposes of obtaining the interaction energy between protein chains for understanding of known molecular structures and rational design of proposed structures of considerable size.

Acknowledgment. The collagen structure, 1A89, used in this article has been taken from the Protein Data Bank (PDB). The research reported in this article was supported by the Office of Naval Research. One of us (L.M.) wishes to thank the U.S. Navy Summer Faculty Research

Program administered by the American Society of Engineering Education for the opportunity to spend summers at NRL. L.M. thanks NIH for grants (NIGMS MBRS SCORE5 S06GM606654 and RR-03037 the National Center For Research Resources) and NSF for CREST grant support.

References

- (1) Huang, L.; Massa, L.; Karle, J. Kernel Energy Method Illustrated with Peptides. *Int. J. Quantum Chem.* **2005**, *103*, 808.
- (2) Huang, L.; Massa, L.; Karle, J. The Kernel Energy Method: Basis Functions and Quantum Methods. *Int. J. Quantum Chem.* **2006**, *106*, 447.
- (3) Huang, L.; Massa, L.; Karle, J. Kernel Energy Method: Application to Insulin. *PNAS* **2005**, *102*, 12690.
- (4) Huang, L.; Massa, L.; Karle, J. Kernel Energy Method: Application to DNA. *Biochemistry* **2005**, *44*, 16747.
- (5) Huang, L.; Massa, L.; Karle, J. Kernel Energy Method: Application to a tRNA. *PNAS* **2006**, *103*, 1233.
- (6) Huang, L.; Massa, L.; Karle, J. Drug-Target Interaction Energies by the Kernel Energy Method: Aminoglycoside Drugs and Ribosomal A-site RNA Target. *PNAS* **2007**, *104*, 4261.
- (7) Massa, L.; Huang, L.; Karle, J. Quantum Crystallography and the Use of Kernel Projector Matrices. *Int. J. Quantum Chem. Quantum Chem. Symp.* **1995**, *29*, 371.
- (8) Huang, L.; Massa, L.; Karle, J. Kernel Projector Matrices for Leu¹- Zervamicin. *Int. J. Quantum Chem. Quantum Chem. Symp.* **1996**, *30*, 1691.
- (9) Huang, L.; Massa, L.; Karle, J. Kernel Projector Matrices: Application to Leu¹- Zervamicin. *Encyclopedia of Computational Chemistry*, 1st ed.; von Schleyer, P., Ed.; John Wiley & Sons: New York, 1998; pp 1457–1470.
- (10) Karle, J.; Huang, L.; Massa, L. Quantum Crystallography, a Technique for Extending the Concept of Structure. *J. Pure Appl. Chem.* **1998**, *70*, 319.
- (11) Huang, L.; Massa, L.; Karle, L. Quantum Crystallography Applied to Crystalline Maleic Anhydride. *Int. J. Quantum Chem.* **1999**, *73*, 439.
- (12) Karle, J.; Huang, L.; Massa, L. Quantum Crystallography: Features and Application, in Current Challenges on Large Supramolecular Assemblies. *NATO Sci. Ser., Ser. C* **1999**, *519*, 1.

- (13) Huang, L.; Massa, L.; Karle, J. Quantum Crystallography, a Developing Area of Computational Chemistry extending to Macromolecules. *IBM J. Res. Dev.* **2001**, 45, 409.
- (14) Karle, J.; Huang, L. The glue that holds crystals together: A review. *J. Mol. Struct.* **2003**, 746, 9.
- (15) Delacoux, F.; Fichard, A.; Geourjon, C.; Garrone, R.; Ruggiero, F. Molecular Features of the Collagen V Heparin Binding Site Theoretical Model. *J. Biol. Chem.* **1998**, 273, 15069.

CT7000649

Electrostatically Embedded Many-Body Correlation Energy, with Applications to the Calculation of Accurate Second-Order Møller–Plesset Perturbation Theory Energies for Large Water Clusters

Erin E. Dahlke and Donald G. Truhlar*

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

Received March 9, 2007

Abstract: The electrostatically embedded many-body expansion (EE-MB), previously applied to the total electronic energy, is here applied only to the electronic correlation energy (CE), combined with a Hartree–Fock calculation on the entire system. The separate treatment of the Hartree–Fock and correlation energies provides an efficient way to approximate correlation energy for extended systems. We illustrate this here by calculating accurate Møller–Plesset second-order perturbation theory (MP2) energies for a series of clusters ranging in size from 5 to 20 water molecules. In this new method, called EE-MB-CE, where MB is pairwise additive (PA) or three-body (3B), the full Hartree–Fock energy of a system of N monomers is calculated (i.e., the many-body expansion is carried out to the N th order), while the EE-MB method is used to calculate the correlation energy of the system. We find that not only does this new method lead to better energetics than the original EE-MB method but also that one is able to obtain excellent agreement with full MP2 calculations by considering only a two-body expansion of the correlation energy, leading to a considerable savings in computational time as compared to the three-body expansion. Additionally, we propose the use of a cutoff to further reduce the number of two-body terms that must be calculated, and we show that if a cutoff of 6 Å is used, then one can eliminate up to 44% of the pairs and still calculate energies to within 0.1% of the net interaction energy of the full cluster.

1. Introduction

The application of post-Hartree–Fock correlated levels of electronic structure theory (e.g., second-order Møller–Plesset perturbation theory, MP2,¹ coupled cluster theory with single and double excitations, CCSD,² or CCSD with quasiperturbative connected triple excitations, CCSD(T)³) to systems containing tens to hundreds of atoms provides a grand challenge to the chemical community because of the rapid scaling of the computational cost of such methods with respect to system size. For example, CCSD(T), CCSD, and MP2 scale as N^7 , N^6 , and N^5 , respectively, where N is the number of atoms.⁴ To meet the challenge of calculating the correlation energy of large systems, there has been considerable interest in trying to develop methods to make the

problem more tractable. One approach is to reduce the scaling by using localized molecular orbitals. Such methods include the natural scaling coupled-cluster,⁵ divide-and-conquer methods,⁶ and cluster-in-molecules methods⁷ as well as many others (See, for example, refs 8–12 and references within.). Another is to break up a large system into many smaller and more manageable subsystems as in the fragment molecular orbital,¹³ many-body expansion,¹⁴ systematic molecular fragmentation,^{15,16} and conjugated caps methods.^{17–19}

In past work we have presented our own fragmentation-based method, the electrostatically embedded many-body (EE-MB) method,²⁰ for calculating the energies of large molecular clusters. The EE-MB method calculates the total energy of a large cluster by taking a linear combination of

the energies of monomers and dimers (in the case of the EE-PA method, where PA denotes pairwise additive) or monomers, dimers, and trimers (in the case of the EE-3B method, where 3B denotes a three-body approximation), with a key element being that each monomer, dimer, or trimer is embedded in a field of point charges representing the other $N - 1$, $N - 2$, or $N - 3$ monomers. (A monomer can be defined as a single molecule or as a collection of molecules, and the method can be extended to allow monomers to be portions of large molecules, such as the monomers of a polymer. In the examples discussed in the present paper, a monomer will be a single water molecule.) Using the EE-MB method we were able to reproduce the absolute correlated interaction energy of a cluster of 21 water molecules to within 2%, by using the EE-PA method, and to within 0.2% when the EE-3B method was used.²⁰ In the present article we present an extension of the electrostatically embedded many-body method, to be called electrostatic embedding of the many-body correlation energy (EE-MB-CE), to predict the MP2 correlation energy for a series of water clusters ranging in size from 5 to 20 water molecules. Because MP2 is the simplest of the correlated methods it provides a good starting point for testing this new method, and since it is the least expensive of the post-Hartree–Fock methods, it allows us to compare our results directly to the MP2 energy for clusters containing 10–20 water molecules; this would not be possible for the more expensive post-Hartree–Fock methods.

2. Theory

By using the electrostatically embedded many-body expansion, the energy of a system of N interacting particles (monomers) can be written as

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

where

$$V_1 = \sum_i E_i \quad (2)$$

$$V_2 = \sum_{i>j} E_{ij} - E_i - E_j \quad (3)$$

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

and so forth, where E_i , E_{ij} , and E_{ijk} are the energies of monomers, dimers, and trimers that are embedded in a sea of point charges representing the other $N - 1$, $N - 2$, or $N - 3$ monomers, and V_n (with $n > 1$) denotes the difference between the n -body approximation and the $(n-1)$ -body approximation. If the series in eq 1 is truncated after the second term one is said to have made the electrostatically embedded pairwise additive approximation; then the total energy of the system can be written as

$$E_{EE-PA} = \sum_{i>j} E_{ij} - (N-2) \sum_i E_i \quad (5)$$

where N is the number of particles in the system, and E_i and E_{ij} have the same meaning as above. If no embedding charges are used, then the subscript on the left side of eq 5 can be changed to PA, and one is said to have made the pairwise approximation.²¹ If one also considers the three-body terms, the electrostatically embedded three-body energy can be written as

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

where E_i , E_{ij} , and E_{ijk} have the same meanings as in eqs 2–5. As in the case of the EE-PA energy, if no point charges are used one can write the subscript on the left-hand side of eq 6 as 3B, and one is said to have made the three-body approximation.

The electronic energy for any correlated level of electronic structure theory can be written as

$$E_X = E_{HF} + \Delta E_{\text{corr},X} \quad (7)$$

where E_X is the electronic energy of correlated method X ($X = \text{MP2}$, CCSD , CCSD(T) , etc.), E_{HF} is the Hartree–Fock energy of the system, and $\Delta E_{\text{corr},X}$ is the correlation energy for method X. Since the n th term of the many body expansion of eq 1 is simply a linear combination of energies for the 1- to n -body clusters, V_n can be rewritten, using eq 7, for any correlated level of theory as

$$V_n = V_{n,HF} + \Delta V_{n,\text{corr}} \quad (8)$$

As a consequence of eq 8 the total energy of the system can be written as

$$V = (V_{1,HF} + V_{2,HF} + V_{3,HF} + \dots + V_{N,HF}) + (\Delta V_{1,\text{corr}} + \Delta V_{2,\text{corr}} + \Delta V_{3,\text{corr}} + \dots + \Delta V_{N,\text{corr}}) \quad (9)$$

where the first term in parentheses is the many-body expansion of the Hartree–Fock energy, and the second term in parentheses is the many-body expansion of the correlation energy.

The Hartree–Fock energy contains electrostatic and inductive terms that can be long-range (e.g., the electrostatic interaction between dipolar monomers dies off only as R^{-3} , where R is the distance between monomers, and the charge–induced dipole interaction dies as R^{-4}), whereas the terms due *entirely* to correlation energy are known to decay as R^{-6} , which is a medium-ranged interaction. However, the inclusion of correlation energy does change the dipole moment of a monomer, leading to changes in the long-range dipole–dipole interactions, which, as mentioned above, die off as R^{-3} . But, if the change in dipole moment between the correlated level of electronic structure theory and Hartree–Fock theory is small, then this effect is also “small”, despite being long-range in nature. Höffinger et al.²² have tested the accuracy of a series of electronic structure methods, including both wave function methods and density functional methods with a variety of basis sets, for predicting dipole moments for a test set of small molecules (N_2 , CO_2 , SO_2 , HF , HCl ,

H₂O, NH₃, PH₃). They found that, on average, the dipole moments of these molecules change by 5–11% when one goes from Hartree–Fock theory to MP2, with the largest mean percent changes using the aug-cc-pVDZ and aug-cc-pVTZ basis sets (11% and 10%, respectively). If one looks only at the water molecule, it was found that for any of the five basis sets tested the percent change in the dipole moment, as one goes from Hartree–Fock to MP2 theory, is never more than 5%. Moreover, they found that as one considers more highly correlated levels of electronic structure theory (e.g., MP4SDQ or QCISD) the mean percent error changes by at most only an additional 2%. Therefore, since most of the change in the dipole moment due to correlation energy is present at the MP2 level of theory, the use of MP2 theory to test methods such as those presented here should provide good insights into the performance of other post-Hartree–Fock methods.

Given the differing nature of the Hartree–Fock and correlation energies, it is not unreasonable to treat their many-body expansions differently by considering more terms in the many-body expansion of the Hartree–Fock energy (in order to better account for the long-range electrostatic and inductive terms) than in the expansion of the correlation energy. Fortunately, since Hartree–Fock theory formally scales as N^4 , where N is again the number of atoms, it is less computationally demanding to consider larger clusters with Hartree–Fock theory than it is for the correlated methods. In practice, one can use Hartree–Fock theory for the calculation of moderately sized systems (up to a few hundred atoms) with a large basis set at an affordable cost.²³ Therefore, we propose to calculate the complete Hartree–Fock energy for the system (i.e., to carry out the many-body expansion to N th order) and calculate only the correlation energy of the system by using a truncated many-body series. If the many-body expansion is used for the correlation energy without the presence of point charges, the result is denoted MB-CE, where MB is PA if the first two terms in the series are kept, and MB is 3B if the first three terms are kept. If the electrostatically embedded many-body expansion is used for the correlation energy, then the results are denoted EE-MB-CE, where MB has the same subcases as above.

3. Computational Methods

In order to test the accuracy of the new methods described in section 2, a series of water clusters ranging in size from 5 to 20 water molecules was taken from the Cambridge Cluster Database.²⁴ These clusters are the global-minimum-energy structures at the HF/6-31G(d,p) level of theory.²⁵ Since water clusters are known to exhibit large many-body effects,^{26,27} this set of clusters should provide a good test of the different methods described here. Eight different many-body methods were applied to these systems: PA, 3B, EE-PA, EE-3B, PA-CE, 3B-CE, EE-PA-CE, and EE-3B-CE, where each method has been described in the previous section. The full cluster calculations were performed using the *Gaussian 03*²⁸ software package. All many-body calculations were carried out with the *MBPAC 2007*²⁹ software package, which uses *Gaussian 03* to perform all electronic structure calculations. For the EE-MB and EE-MB-CE

Table 1. Comparison of Mean Errors^a (kcal/mol) and Mean Percent Errors^b (%) for Different Many-Body Methods, as Compared to Full Cluster Calculations^c

	MSE	MUE	RMSE	MPSE	MPUE	RMPSE
PA	15.95	15.95	17.55	15.40	15.40	15.47
3B	0.55	0.56	0.71	0.69	0.69	0.88
EE-PA	0.80	0.80	0.84	0.82	0.82	0.83
EE-3B	-0.34	0.35	0.51	-0.24	0.26	0.33
PA-CE	0.22	0.22	0.24	0.22	0.22	0.23
3B-CE	-0.05	0.17	0.24	0.01	0.15	0.18
EE-PA-CE	-0.10	0.10	0.11	-0.09	0.09	0.10
EE-3B-CE	-0.23	0.23	0.34	-0.16	0.17	0.21

^a MSE, MUE, and RMSE denote mean signed, mean unsigned, and root mean squared errors, respectively. ^b MPSE, MPUE, and RMPSE denote mean percent signed, mean percent unsigned, and root mean percent squared errors, respectively. ^c All calculations correspond to the MP2/aug'-cc-pVTZ level of theory, which uses the aug-cc-pVTZ basis set on oxygen, and the cc-pVTZ basis set on hydrogen.³¹

calculations, point charges of -0.778 and 0.389 were used for the oxygen atoms and hydrogen atoms, respectively, as in ref 20.

4. Results and Discussion

Table 1 shows the mean errors and mean percent errors for the eight different many-body methods as compared to the full MP2 calculations. One of the striking results of Table 1 is the improvement of the EE-PA method as compared to the PA approximation. The inclusion of point charges changes the mean unsigned error from 15.95 kcal/mol to only 0.80 kcal/mol, which is consistent with previous results.²⁰ Considering that the binding energies range from 33.52 to 196.02 kcal/mol with an average of 105.46 kcal/mol, a mean unsigned error of 0.80 kcal/mol is an impressive result. One can also see that inclusion of the three-body terms improves the energy by only 0.5–0.6 kcal/mol. Since the trimer calculations are the most numerous and most expensive calculations considered in this article, the EE-PA may be sufficient for many applications.

The second significant result is the large reduction in error between the PA and PA-CE methods. The mean unsigned error is reduced by a factor of 72 by including the full Hartree–Fock energy! The other methods show a much smaller change in their mean errors when the full Hartree–Fock energy is included, with changes of a factor of 3, 8, and 1.5 for the 3B, EE-PA, and EE-3B methods, respectively. The EE-PA-CE method is the most accurate method with a mean unsigned error of only 0.10 kcal/mol, which represents a mean percent unsigned error of only 0.09% of the net interaction energies. As mentioned previously, the ability to consistently calculate total energies that are within less than 0.1% of the full cluster calculation by only having to consider two-body terms represents a significant savings in the total computational time needed to carry out the calculation. For example, if one assumes that the time needed to calculate the energy of a water monomer, dimer, and trimer at the MP2/aug'-cc-pVTZ level of theory is 30 s, 2 min, and 5 min respectively, then the total time needed to calculate the MP2 correlation energy for a cluster of 20 water molecules is 6.5 h with the EE-PA-CE method as compared to 4.2 days

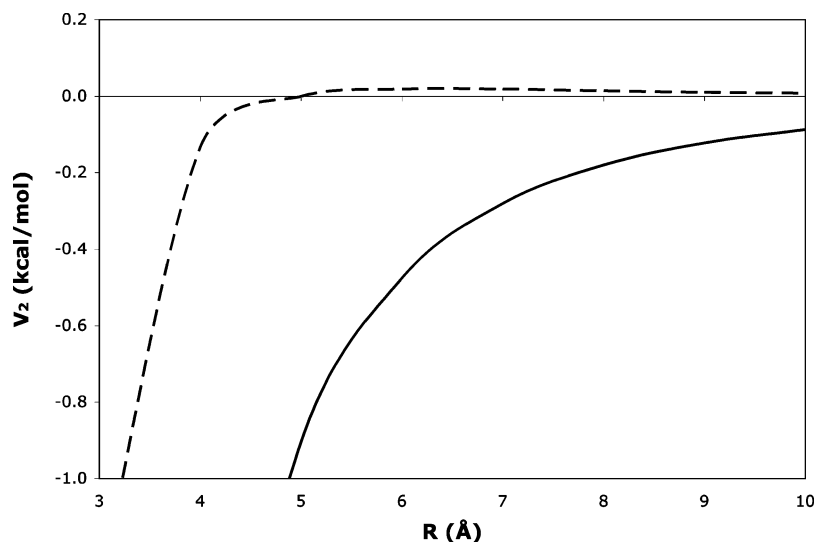


Figure 1. Two-body energy versus center-of-mass separation for the water dimer. The solid line is the HF/aug'-cc-pVTZ result; the dashed line is the result for the MP2 correlation energy also using the aug'-cc-pVTZ basis set.

Table 2. Comparison of Mean Errors^a (kcal/mol) for PA-CE and EE-PA-CE Methods with $R_{\text{cut}} = 5 \text{ \AA}$, $R_{\text{cut}} = 6 \text{ \AA}$, and $R_{\text{cut}} = \infty$

	$R_{\text{cut}} = 5 \text{ \AA}$			$R_{\text{cut}} = 6 \text{ \AA}$			$R_{\text{cut}} = \infty$		
	MSE	MUE	RMSE	MSE	MUE	RMSE	MSE	MUE	RMSE
PA-CE	0.58	0.58	0.71	0.38	0.38	0.46	0.22	0.22	0.23
EE-PA-CE	0.31	0.33	0.46	0.02	0.07	0.09	-0.09	0.09	0.10

^a MSE, MUE, and RMSE denote mean signed, mean unsigned, and root mean squared errors, respectively.

to calculate the correlation energy with the EE-3B-CE method. The fact that good energetics can be determined using only a two-body approximation is consistent with other fragment-based methods that have been proposed in the literature.^{13,18,30}

The result that the EE-3B-CE method has a larger average error than the EE-PA-CE method is somewhat surprising, as one might expect the EE-3B-CE method to give a smaller average error than the EE-PA-CE method since it contains more terms in the many-body expansion. If the errors for each individual structure are examined, then one finds that of the 16 structures considered, eight have a larger error at the EE-3B-CE level than at the EE-PA-CE level. Similarly, four of the 16 structures have larger errors at the 3B-CE than at the PA-CE level, despite the 3B-CE method having a lower average error. Since the EE-MB-CE methods use embedding and the MB-CE methods do not, the use of embedding cannot be the sole source of this error. In fact, if the many-body expansion given in eq 1 is not truncated, then the result is exact and is independent of whether or not embedding is used. We can also exclude double counting as a source of error. At the EE-PA-CE level we account accurately for the two-body terms and approximate the higher order many-body terms, then, at the EE-3B-CE level, we subtract the two-body terms with approximate three-body effects, and we treat the three-body terms (and lower-order terms) exactly and estimate the higher-order terms. This continues at higher orders so that the method is free of any double counting. The source of the errors for both the EE-3B-CE and 3B-CE methods would be an interesting topic

for further study; however, for now we proceed with examining the EE-PA-CE method.

If one is interested in trying to further decrease the cost of the calculation one could consider implementing a cutoff to reduce the number of pairs that one must calculate. Since correlation energy is typically short-ranged as compared to the Hartree–Fock energy (see the Background section for more discussion of this point), it may be reasonable to assume that for a dimer with a large intermolecular distance the contribution of the correlation energy to the two-body term might be very small. In order to determine a reasonable cutoff for water, we next examine the magnitude of the two-body contribution to the energy (V_2) as a function of distance. In order to examine this, the global-minimum-energy structure of the water dimer was optimized at the CCSD(T)/aug-cc-pVTZ level of theory and was separated along the vector connecting the centers of mass from 4 to 10 Å in intervals of 1 Å. For each of these seven geometries as well as at the optimized geometry an MP2/aug'-cc-pVTZ single-point calculation was carried out, and the V_2 term was calculated. Figure 1 shows the plot of V_2 as a function of this separation, for both the HF/aug'-cc-pVTZ energy and the MP2/aug'-cc-pVTZ correlation energy. Figure 1 clearly shows that the V_2 term for the MP2 correlation energy goes to zero much more rapidly than for the Hartree–Fock energy. In fact, by $\sim 4.5 \text{ \AA}$ the V_2 term for the MP2 correlation energy is approximately zero. While this plot does not take into account any type of rotational averaging, and any one orientation cannot be fully representative, it does suggest that considering a cutoff between 5 and 6 Å might be a reasonable starting point.

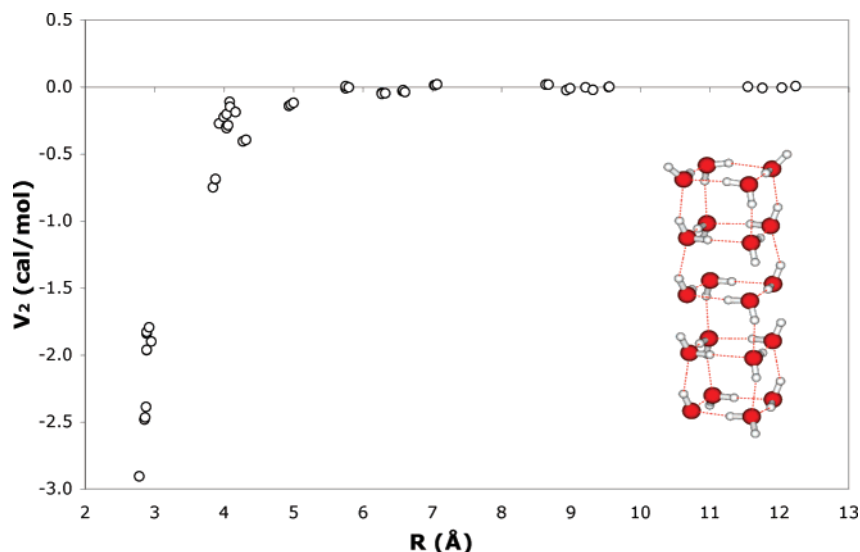


Figure 2. Two-body energy versus center-of-mass separation for the 190 dimers of $(\text{H}_2\text{O})_{20}$ structure shown. Each circle represents the MP2/aug'-cc-pVTZ two-body correlation energy for one of the dimers. In Figures 2 and 3, many of the circles cannot be seen because they are obscured by other circles.

Table 2 compares the mean errors obtained using cutoffs of 5 and 6 Å for the V_2 term to the mean errors if no cutoffs are used. If a cutoff of 5 Å is used, at least one pair can be disregarded in 13 of the 16 structures studied; however, one must go up to clusters of 10 water molecules before a significant number (5 or more) of pairs can be ignored. If a cutoff of 6 Å is used one must consider structures containing 11 water molecules or more before a significant number of pairs can be ignored; however, by the time 20 water molecules are present only 56% of the total number of pairs need to be considered. Additionally, one can see that for the EE-PA-CE method the use of a cutoff of 6 Å is able to reproduce the accuracy obtained when no cutoff is used. Based on the timing arguments presented in the second paragraph of this section it would take only approximately 3.6 h to calculate the EE-PA-CE correlation energy for a cluster of 20 water molecules with a cutoff of 6 Å.

Figure 2 shows a plot of the two-body correlation energy versus the center-of-mass separation for each of the 190 dimers in the 20-mer. Based on this figure it is clear that the two-body correlation energy for this cluster goes to zero at approximately 6 Å, as opposed to the gas-phase water dimer, which becomes negligible at about 4.5 Å. As mentioned previously, consideration of the gas-phase water dimer was used as a guide to approximate where an appropriate cutoff might be; however, it is clear from Figure 2 that while such a rudimentary example can give some insight into the choice of cutoff one may still need to consider several cutoffs or carry out a full analysis on a large cluster to obtain the best possible cutoff for the system of interest.

Figure 3 shows a plot of the electrostatically embedded two-body correlation energy versus the center-of-mass separation of the same cluster as in Figure 2. A comparison of Figures 2 and 3 shows that the addition of the embedded charges does not change the range of the interaction—both decrease to zero at approximately 6 Å; however, it is evident that the addition of the point charges does introduce some

of the higher-order many-body terms as the magnitude of the two-body term in the range of $\sim 3\text{--}6$ Å is noticeably different between the two plots, particularly in the region of the plot from 3.9 to 4.2 Å. If one considers the cluster shown to be a series of cubes stacked on top of each other, all the dimers in the region from 3.9 to 4.2 Å are pairs of water molecules that form diagonals across the faces of these cubes. The different orientations and distances between the water molecules give rise to three clusters of points in Figure 2. However, all of these dimers are a part of larger tetrameric clusters, making up the faces of the cubes, which have cooperative hydrogen bonding around the cycle, leading to large many-body effects. Addition of the embedded point charges helps to mimic these effects, making the series of dimers converge to a smoother envelope of points in Figure 3.

The last issue we would like to discuss is the efficiency with which gradients can be calculated since gradients are necessary for carrying out geometry optimizations or molecular dynamics calculations. In previous work²⁰ we discussed the linearity of the original EE-MB method and the ease with which gradients could be implemented. For example, the gradient of the EE-PA energy can be written as

$$\nabla E_{\text{EE-PA}} = \sum_{i < j}^N \nabla E_{ij} - (N - 2) \sum_i^N \nabla E_i \quad (10)$$

where an analytic gradient for the EE-PA method is available for any method that has analytic gradients for the monomer and dimer calculations, provided the program allows for fractionally charged point charges as pseudonuclei. For the EE-MB-CE methods presented here, the total energy can still be written as a linear combination of energies. For example,

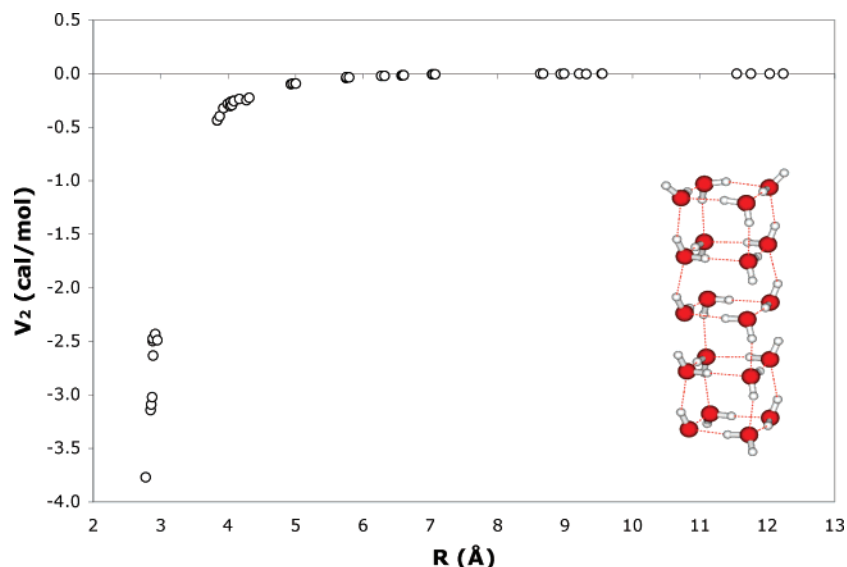


Figure 3. Plot of the electrostatically embedded two-body energy versus center-of-mass separation for the 190 dimers of $(\text{H}_2\text{O})_{20}$ structure shown. Each circle represents the electrostatically embedded MP2/aug'-cc-pVTZ two-body correlation energy for one of the dimers.

the total energy at the EE-PA-CE level, using correlated method X, is given by

$$\begin{aligned}
 E_{\text{EE-PA-CE}} &= E_{\text{HF}} + \Delta E_{\text{EE-PA,corr}} \\
 &= E_{\text{HF}} + \sum_{i < j}^N \Delta E_{ij,\text{corr}} - (N - 2) \sum_i^N \Delta E_{i,\text{corr}} \\
 &= E_{\text{HF}} + \sum_{i < j}^N (E_{ij,X} - E_{ij,\text{HF}}) - \\
 &\quad (N - 2) \sum_i^N (E_{i,X} - E_{i,\text{HF}}) \quad (11)
 \end{aligned}$$

Because the gradient is a linear operator, the gradient of the energy given in eq 11 can be written as

$$\begin{aligned}
 \nabla E_{\text{EE-PA-CE}} &= \nabla E_{\text{HF}} + \sum_{i < j}^N (\nabla E_{ij,X} - \nabla E_{ij,\text{HF}}) - \\
 &\quad (N - 2) \sum_i^N (\nabla E_{i,X} - \nabla E_{i,\text{HF}}) \quad (12)
 \end{aligned}$$

and it is again true that the method will have analytic gradients so long as the electronic structure methods used have analytic gradients. A key point here is that the values of our point charges are fixed. Since the magnitude of these charges are fixed, the embedding charges are like fractionally charged nuclei with no basis functions, and so the only extension of the usual gradient routines that is required is to allow fractionally charged “nuclei”. This is an important advantage of the present method over some alternative many-body schemes.

5. Summary and Conclusions

We have presented here an extension of the electrostatically embedded many-body method that calculates the full Hartree–Fock energy of the system and applies the EE-MB

method only to the correlation energy of the system. We have found that for MP2 correlation energies the inclusion of the full Hartree–Fock energy reduces the error of the standard pairwise additive approximation by a factor of 72, and the error of the EE-PA method by a factor of 8. We have also found that one can accurately calculate the energies of clusters containing up to 20 water molecules to within 0.09%, on average, of the net interaction energy by considering only the two-body terms for the correlation energy. Since the calculations needed to evaluate the three-body terms in many-body expansion are both the most numerous and most expensive, this constitutes a substantial savings in time.

Finally, we have demonstrated that the use of a cutoff for evaluation of the two-body term can reduce the number of dimer terms that need to be calculated substantially, without having a large impact on the accuracy of the EE-PA-CE method. Using a cutoff of 6 Å we are able to reproduce the total energy of a cluster of 20 water molecules to within 0.1% of the net interaction energy by calculating the correlation energy of only 106 of the 190 possible pairs of water molecules. In the future we hope to extend this work both to larger systems and to other levels of correlated electronic structure theory.

The *MBPAC 2007* software package for running EE-MB and EE-MB-CE calculations, where MB is PA or 3B, is available free of charge and may be downloaded at <http://comp.chem.umn.edu/mbpac>.

Acknowledgment. The authors thank Ryan Olson for stimulating discussions. This work was supported in part by the National Science Foundation under grant nos. CHE03-49122 and ITR-0428774.

Supporting Information Available: Binding energies at the MP2, PA, 3B, EE-PA, EE-3B, PA-CE, 3B-CE, EE-PA-CE, and EE-3B-CE levels of theory for each of the clusters considered in this work (Table S1). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.
- (2) Cizek, J. *Adv. Chem. Phys.* **1969**, *14*, 35.
- (3) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head, Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
- (4) Raghavachari, K.; Anderson, J. B. *J. Phys. Chem.* **1996**, *100*, 12960.
- (5) Flocke, N.; Bartlett, R. J. *J. Chem. Phys.* **2004**, *121*, 10935.
- (6) Li, W.; Li, S. *J. Chem. Phys.* **2004**, *121*, 6649.
- (7) Li, S.; Shen, J.; Li, W.; Jiang, Y. *J. Chem. Phys.* **2006**, *125*, 074109.
- (8) Saebø, S.; Pulay, P. *J. Chem. Phys.* **1987**, *86*, 914.
- (9) Scuseria, G. E.; Ayala, P. Y. *J. Chem. Phys.* **1999**, *111*, 8330.
- (10) Lee, M. S.; Maslen, P. E.; Head, Gordon, M. *J. Chem. Phys.* **2000**, *112*, 3592.
- (11) Schutz, M.; Werner, H.-J. *J. Chem. Phys.* **2001**, *114*, 661.
- (12) Casassa, S.; Zicovich-Wilson, C. M.; Pisani, C. *Theor. Chem. Acc.* **2005**, *116*, 726.
- (13) Federov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *121*, 2483.
Federov, D. G.; Kitaura, K. *J. Chem. Phys.* **2005**, *123*, 134103.
- (14) Christie, R. A.; Jordan, K. D. *Struct. Bond.* **2005**, *116*, 27.
- (15) Deev, V.; Collins, M. A. *J. Chem. Phys.* **2005**, *122*, 154102.
- (16) Collins, M. A.; Deev, V. A. *J. Chem. Phys.* **2006**, *125*, 104104.
- (17) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2002**, *119*, 3599.
- (18) Li, S.; Li, W.; Fang, T. *J. Am. Chem. Soc.* **2005**, *127*, 7215.
- (19) Jiang, N.; Ma, J.; Jiang, Y. *J. Chem. Phys.* **2006**, *124*, 114112.
- (20) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comp.* **2007**, *3*, 46.
- (21) Elrod, M. J.; Saykally, R. J. *Chem. Rev.* **1994**, *94*, 1975.
- (22) Höffinger, S.; Wendland, M. *Int. J. Quantum Chem.* **2002**, *86*, 199.
- (23) Polly, R.; Werner, H.-J.; Manby, F.; Knowles, P. *Mol. Phys.* **2004**, *102*, 2311.
- (24) Wales, D. J.; Doye, J. P. K.; Dullweber, A., et al. Cambridge Cluster Database. <http://www-wales.ch.cam.ac.uk/CCD.html> (accessed March 8, 2006).
- (25) Maheshwary, A.; Patel, N.; Sathyamurthy, N.; Kulkarni, A. D.; Gadre, S. R. *J. Phys. Chem. A* **2001**, *105*, 10525.
- (26) Xantheas, S. S. *J. Chem. Phys.* **1994**, *100*, 7523.
- (27) Dahlke, E. E.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *110*, 10595.
- (28) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Robb, G. E. S. 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. *Gaussian03—version c01*; Gaussian Inc.: Wallingford, CT, 2004.
- (29) Dahlke, E. E.; Truhlar, D. G. *MBPAC 2007*; University of Minnesota: Minneapolis, MN, 2007.
- (30) Federov, D. G.; Kitaura, K. *J. Chem. Phys.* **2005**, *123*, 134103.
- (31) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 46.

CT700057X

JCTC Journal of Chemical Theory and Computation

Parameter Calibration of Transition-Metal Elements for the Spin-Polarized Self-Consistent-Charge Density-Functional Tight-Binding (DFTB) Method: Sc, Ti, Fe, Co, and Ni

Guishan Zheng,[†] Henryk A. Witek,[‡] Petia Bobadova-Parvanova, Stephan Irle,[§]
Djamaladdin G. Musaev, Rajeev Prabhakar,^{||} and Keiji Morokuma*

*Department of Chemistry and Cherry L. Emerson Center for Scientific Computation,
Emory University, Atlanta, Georgia 30322*

Marcus Lundberg

*Fukui Institute for Fundamental Chemistry, Kyoto University, Sakyo,
Kyoto 606-8103, Japan*

Marcus Elstner,[⊥] Christof Köhler,[#] and Thomas Frauenheim[#]

Universität Paderborn, Fachbereich Physik, 33095 Paderborn, Germany

Received October 24, 2006

Abstract: Recently developed parameters for five first-row transition-metal elements (M = Sc, Ti, Fe, Co, and Ni) in combination with H, C, N, and O as well as the same metal (M–M) for the spin-polarized self-consistent-charge density-functional tight-binding (DFTB) method have been calibrated. To test their performance a couple sets of compounds have been selected to represent a variety of interactions and bonding schemes that occur frequently in transition-metal containing systems. The results show that the DFTB method with the present parameters in most cases reproduces structural properties very well, but the bond energies and the relative energies of different spin states only qualitatively compared to the B3LYP/SDD+6-31G(d) density functional (DFT) results. An application to the ONIOM(DFT:DFTB) indicates that DFTB works well as the low level method for the ONIOM calculation.

1. Introduction

Molecules that contain transition-metal atoms play an important role in catalysis, material science, drug design, and

enzymatic reactions. Theoretical modeling of such systems is challenging due to their large size and complexity of their electronic structure arising from the presence of chemically active d-electrons. Despite the advent of fast computers and advanced techniques, high level ab initio methods are prohibitively expensive to treat very large molecular systems. A partial remedy for this problem can be the density functional theory (DFT),^{1–3} which can be used routinely to systems containing a few hundred atoms with the present computers.

Efforts to reduce computational cost associated with quantum chemical calculations have led in last several decades to development of a large number of semiempirical methods, such as MNDO,^{4,5} SINDO/1,^{6,7} AM1,⁸ PM3,^{9,10} SAM1,^{11,12} MNDO/d,^{13,14} PM3/tm,¹⁵ NDDO-G,¹⁶ PDDG/PM3,^{17–20} NO-MNDO,²¹ and RM1²² which can routinely

* Corresponding author e-mail: morokuma@emory.edu. Also at Fukui Institute for Fundamental Chemistry, Kyoto University, Kyoto, Japan.

[†] Present address: Department of Chemistry, University of Illinois, Urbana, IL 61801.

[‡] Present address: Institute of Molecular Science and Department of Applied Chemistry, National Chiao Tung University, Hsinchu, Taiwan.

[§] Present address: Institute for Advanced Research and Department of Chemistry, Nagoya University, Nagoya 464-8602, Japan.

^{||} Present address: Department of Chemistry, University of Miami, Coral Gables, FL 33124.

[⊥] Present address: Technical University, Braunschweig, Germany.

[#] Present address: University of Bremen, Bremen, Germany.

treat molecular systems containing up to 1000 or so atoms. An alternative approach to perform calculations for such large systems is an approximate density functional technique called the density functional tight-binding (DFTB) method.^{23,24} This method has been applied to calculating energies, geometries, and spectra of organic and inorganic molecules.^{23–29} The accuracy for molecular geometries is comparable to that of DFT-GGA methods, while reaction energies and vibrational frequencies are slightly less accurate.^{20,25,30–36} Recently, a special parametrization for vibrational frequencies has shown that DFTB can approach the DFT accuracy,³⁷ while heats of formation are still slightly less accurate than those determined at recently optimized MNDO approaches.²¹

In the present article, we will use a specific version of the series of DFTB methods, the spin-polarized self-consistent charge DFTB,³⁸ which is based on a second-order expansion of the Kohn–Sham total energy with respect to spin densities. This method introduces a self-consistent calculation of the spin density using Mulliken populations. The SCF procedure minimizes the dependence of the results on the choice of the zero-order initial density and substantially increases the transferability of the parameters in comparison with the non-self-consistent-charge approach.²⁴ In addition, the spin-polarized version of DFTB distinguishes different spin distributions (whereas spin-unpolarized DFTB depends only on the total electron density) and can qualitatively describe different spin states, a fact that is essential for transition-metal elements. All the needed one- and two-center integrals are precomputed for a large number of grid points, and, in practical calculations, the actual values of integrals are obtained by a suitable interpolation scheme, usually a cubic spline function fitting. All electronic parameters of the spin-polarized DFTB model are calculated from DFT using the PBE functional,³⁹ while two-center repulsive potentials are fitted to results using hybrid functional, i.e., B3LYP,^{40,41} no fitting to experimental data is involved. Since only valence electrons are considered in a minimal basis set and explicit integral evaluations are not required, DFTB is computationally comparable to semiempirical methods (like MNDO, AM1, PM3) and 2–3 orders of magnitude faster than ab initio Hartree–Fock (HF) and density functional theory (DFT) methods.²⁵ As a result, the computational speed of DFTB is determined to a large extent by the solution of the generalized eigenvalue problem.

Up to now, the only transition metals available in DFTB are Zn²⁶ and some other scattered atom pair parameters,^{27,42–58} and therefore one of the serious drawbacks of the DFTB method was the lack of parameters for further transition-metal elements, which play an important role in many inorganic, organometallic, and metalloprotein problems. This situation has restricted the active use of DFTB methods from many interesting applications.

In the present paper, we present our recent work on extending the currently available spin-polarized DFTB parameter database in the form of the Paderborn group to five additional elements: Sc, Ti, Fe, Co, and Ni, which are parametrized in combination with C, H, O, and N nonmetal elements as well as with the element itself (dimer). In section 2, we give an overview and procedure for the parametrization

procedure. In section 3, test calculations using the new parameters are discussed and analyzed. Here, the performance of the parameter sets in different chemical environments is discussed in detail, focusing on calculated molecular geometries and energies. In section 3, we present a sample application of the new parameters in ONIOM(DFT:DFTB) method, and in section 4, we summarize the performance and problems of the new parameters.

2. Method and Parametrization

A. Spin-Polarized Self-Consistent-Charge DFTB Approach. A detailed description of the spin-polarized self-consistent-charge density-functional tight-binding (DFTB) method has been given elsewhere.^{38,59,60} Here a brief review is presented. The total spin-polarized DFTB energy is given by

$$E_{\text{tot}}^{\text{SDFTB}} = \sum_{\sigma=\uparrow,\downarrow} \sum_i^{\text{MO}} n_i^\sigma \langle \psi_i^\sigma | \hat{H}^0 | \rho_0 | \psi_i^\sigma \rangle + \frac{1}{2} \sum_{A,B}^{\text{atom}} \gamma_{AB} \Delta q_A \Delta q_B + \frac{1}{2} \sum_A^{\text{atom}} \sum_l \sum_{l'} p_{Al} p_{Al'} W_{All'} + E^{\text{rep}} \quad (1)$$

where \uparrow and \downarrow denote the up and down spin orientation, γ_{AB} is a distance-dependent interaction parameter between induced Mulliken charges Δq_A , Δq_B on atoms A and B , $W_{All'}$ is a one-center interaction parameter between the l and l' shell spin densities p_{Al} , $p_{Al'}$ on atom A , E^{rep} is a sum of two-center core–core repulsive potentials:

$$E^{\text{rep}} = \sum_{A < B}^{\text{atom}} E_{AB}^{\text{rep}} \quad (2)$$

and n_i^σ is the occupation number of the spin orbital ψ_i^σ that is given as a linear combination of localized pseudoatomic Slater orbitals χ_{μ}

$$\psi_i^\sigma = \sum_{\mu}^{\text{AO}} c_{\mu i}^\sigma \chi_{\mu} \quad (3)$$

The induced Mulliken charge Δq_A on atom A is given by

$$\Delta q_A = \sum_i^{\text{MO}} \sum_{\mu \in A} \sum_v^{\text{AO}} (n_i^\uparrow c_{\mu i}^\uparrow c_{vi}^\uparrow + n_i^\downarrow c_{\mu i}^\downarrow c_{vi}^\downarrow) S_{\mu v} - q_A^0 \quad (4)$$

and the spin density p_{Al} of shell l on atom A is given by

$$p_{Al} = \sum_i^{\text{MO}} \sum_{\mu \in l} \sum_v^{\text{AO}} (n_i^\uparrow c_{\mu i}^\uparrow c_{vi}^\uparrow - n_i^\downarrow c_{\mu i}^\downarrow c_{vi}^\downarrow) S_{\mu v} \quad (5)$$

where S is the overlap matrix of pseudoatomic Slater orbitals, and q_A^0 is the valence charge on the neutral atom A . The effective Kohn–Sham Hamiltonian \hat{H}^0 depends only on the reference density ρ_0 .

The derivation of the spin-polarized DFTB energy with respect to nuclear coordinate a yields the DFTB energy gradient acting on atom A . The exact formula and its derivation can be found elsewhere.⁶⁰

Table 1. Chemical Hardness or Hubbard Parameters U_{MI} and the Atomic Spin-Dependent Constants W_{MIr} (both in Hartree) for M = Sc, Ti, Fe, Co, and Ni

element	Sc	Ti	Fe	Co	Ni
U_s	0.18881	0.20020	0.20050	0.26064	0.23145
U_p	0.13784	0.14432	0.20050	0.11593	0.18913
U_d	0.32717	0.35522	0.36342	0.38599	0.40632
W_{ss}	-0.013	-0.014	-0.016	-0.016	-0.016
W_{sp}	-0.011	-0.012	-0.012	-0.012	-0.012
W_{sd}	-0.005	-0.004	-0.003	-0.003	-0.003
W_{pp}	-0.014	-0.014	-0.029	-0.033	-0.022
W_{pd}	-0.002	-0.001	-0.001	-0.001	-0.001
W_{dd}	-0.013	-0.014	-0.015	-0.016	-0.018

B. Development of Atomic and Diatomic Parameters

Sets. We develop the M–M and M–X diatomic parameter sets, where M = Sc, Ti, Fe, Co, and Ni and X = H, C, O, and N. Atomic valence orbitals are obtained by solving an all-electron Kohn–Sham atomic eigenvalue problem with an additional confining potential. The repulsive potentials for each pair of atoms are obtained by reproducing DFT energies and geometries for a number of carefully selected molecular systems; their choice is ideally meant to represent the most important chemical compounds created by a given pair of atoms.

In the present paper, spin-polarized DFTB parametrization is performed in the same way as in the standard, nonspin-polarized self-consistent-charge (SCC-) method.⁶¹ Here, we will review briefly the main ideas of the parametrization procedure together with necessary modifications required by the introduction of the spin-polarization term. There are two families of parameters necessary to construct the spin-polarized DFTB Hamiltonian, namely (1) atomic parameters obtained from calculations for confined pseudoatoms and (2) diatomic distance-dependent parameters obtained from diatomic calculations.

Atomic Parameters. The required spin-polarized DFTB atomic parameters comprise atomic basis functions χ_{μ} , chemical hardness or Hubbard parameters U_{Al} , and the atomic spin-dependent constants W_{Alr} . U_{Al} is determined by taking the second derivative of the total atomic energy with respect to the total charge on orbital l of atom A . The values of W_{Alr} are calculated by taking the second derivatives of the total atomic energy with respect to the spin density; at the point where the spin density is zero, this derivative reduces to³⁸

$$W_{Alr} = \frac{1}{2} \left(\frac{\partial \epsilon_{Al}^{\uparrow}}{\partial n_r^{\uparrow}} - \frac{\epsilon_{Al}^{\uparrow}}{\partial n_r^{\downarrow}} \right)_{\rho=0} = W_{Alr} \quad (6)$$

where n_l and n_r are the occupation numbers of atomic shells l and l , respectively, and ϵ_{Al}^{\uparrow} is the atomic Kohn–Sham orbital energy for alpha (\uparrow) spin. The second derivative values are computed using finite difference method. The determined values of U_{Al} and W_{Alr} are listed in Table 1 for all the considered metal elements.

We use a standard procedure to construct the atomic basis set. It is expressed as a linear combination of Slater spherical harmonics; the coefficients are obtained from atomic Kohn–

Sham calculations with the PBE functional³⁹ and an additional confining potential $(r/r_0)^2$ (in Hartree). The confinement mimics the behavior of atoms in molecular systems and in solids. The values of r_0 (4.86 for Sc, 3.6 for Ti, 3.2 for Fe, 4.38 for Co, and 3.2 for Ni, all in bohr) have been selected out of a large number of trials and ensure that SCC-DFTB reproduces accurate DFT electronic band structures for solid-state metals to the highest possible degree.

Diatomic Parameters. The overlap $S_{\mu\nu}$ and Hamiltonian $H_{\mu\nu}^0$ matrix elements are obtained from two-center approximate noniterative DFT calculations on the corresponding diatomic compounds for a large number of different interatomic distances, i.e., the two-center integral calculations using atomic wavefunctions from previous pseudoatomic calculations. The term “approximate DFT” refers to the fact that the exchange-correlation functional is built from approximate electronic density obtained as a simple sum of unperturbed atomic densities. The atomic densities for these transition-metal atoms are obtained from an auxiliary pseudoatomic calculation with an additional confining potential $(r/r_0)^2$ where a universal value of $r_0 = 14$ bohr is adopted for all studied transition metals. It is important to stress that the confinement radius r_0 used previously to construct valence atomic orbitals is different from the confinement radius used here to generate the zero-order unperturbed atomic density. The confinement radius for the orbitals is used to generate a minimal LCAO basis set that is appropriate for the target molecular systems, i.e., the choice of this parameter for the basis set determination can be compared to the procedure of basis set construction for HF or DFT calculations and r_0 has originally been treated as a variational parameter.⁶² The choice of the confinement radius for the density is different in its nature; it can be interpreted as an empirical value to generate an optimal starting (input) density that is characteristic to tight-binding methods.⁶³

These two confinement radii can be treated as parameters used to enhance the performance of new DFTB parameter sets. However, the influence of these values on molecular properties is rather small. In the present parametrization procedure the PBE functional is used.³⁹ The orbitals employed in this calculation are atomic Slater orbitals with confined potential discussed in the paragraph above. The values of $S_{\mu\nu}$ and $H_{\mu\nu}^0$ are represented numerically on a grid of atomic distance.

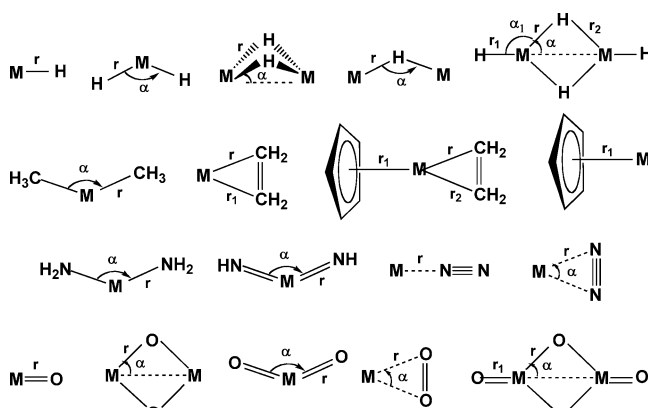
Determination of the two-center repulsive potentials E_{AB}^{rep} is the most labor-intensive and therefore most time-consuming step in the parametrization procedure. The repulsive potential function is the difference between the DFT energy and the DFTB electronic energy as a function of atomic distance. At first, segments of each repulsion potential were calculated for a carefully selected group of molecules (called tier 1 molecules in Table 2) representing a large spectrum of bonding situations (covalent single and multiple bonds, ionic bonds, back-donation bonds, π -interactions, etc.) for a given pair of elements A–B. We have used small molecules containing only a few atoms, in which typically additional hydrogen or other atoms have been used to saturate the unfilled valences of given transition-metal and nonmetal atoms. In general, mainly closed-shell molecules have been

Table 2. List of Molecules and Their Spin States Used in the Parametrization Procedure^a

	M–M	M–H	M–C	M–N	M–O
M = Sc					
tier 1	¹ Sc ₂	¹ ScH ₃	¹ HScCH ₂ ¹ H ₂ ScCH ₃	¹ ScN ¹ H ₂ ScN ₂	¹ HScO ¹ H ₂ ScOH
M = Ti					
tier 1	¹ Ti ₂	¹ TiH ₂	¹ HTiCH ¹ H ₂ TiCH ₂ ¹ H ₃ TiCH ₃	¹ HTiN ¹ H ₂ TiNH ¹ H ₃ TiNH ₂	¹ H ₂ TiO ¹ H ₃ TiOH
tier 2			¹ Ti(CO) ₂ ⁺⁴ ¹ Ti(CO) ₃ ⁺⁴ ¹ Ti(CO) ₄ ⁺⁴ ¹ Ti(CO) ₅ ⁺⁴ ¹ Ti(CO) ₆ ⁺⁴	¹ Ti(NH ₃) ₂ ⁺⁴ ¹ Ti(NH ₃) ₃ ⁺⁴ ¹ Ti(NH ₃) ₄ ⁺⁴ ¹ Ti(NH ₃) ₅ ⁺⁴ ¹ Ti(NH ₃) ₆ ⁺⁴	¹ Ti(H ₂ O) ₂ ⁺⁴ ¹ Ti(H ₂ O) ₃ ⁺⁴ ¹ Ti(H ₂ O) ₄ ⁺⁴ ¹ Ti(H ₂ O) ₅ ⁺⁴ ¹ Ti(H ₂ O) ₆ ⁺⁴
M = Fe					
tier 1	¹ Fe ₂	¹ FeH ₂	¹ FeCH ₂ ¹ FeCH ₃ ⁺ ¹ HFeCO	¹ FeNH ¹ HFeNH ₂ ¹ FeNH ₃ ⁺²	¹ FeO ¹ HFeOH ¹ FeOH ₂ ⁺²
tier 2			⁶ Fe(CO) ₂ ⁺³ ⁶ Fe(CO) ₃ ⁺³ ⁶ Fe(CO) ₄ ⁺³ ⁶ Fe(CO) ₅ ⁺³ ⁶ Fe(CO) ₆ ⁺³	⁶ Fe(NH ₃) ₂ ⁺³ ⁶ Fe(NH ₃) ₃ ⁺³ ⁶ Fe(NH ₃) ₄ ⁺³ ⁶ Fe(NH ₃) ₅ ⁺³ ⁶ Fe(NH ₃) ₆ ⁺³	⁶ Fe(H ₂ O) ₂ ⁺³ ⁶ Fe(H ₂ O) ₃ ⁺³ ⁶ Fe(H ₂ O) ₄ ⁺³ ⁶ Fe(H ₂ O) ₅ ⁺³ ⁶ Fe(H ₂ O) ₆ ⁺³
M = Co					
tier 1	¹ Co ₂	² CoH ₂ ¹ CoH ₃	¹ CoCH ² CoCH ₂ ⁺ ² CoCO ²⁺	¹ CoN ¹ HCoNH ² CoNH ₃ ⁺²	¹ HCoO ¹ HOCO ₂
tier 2			² Co(CO) ₁ ⁺² ² Co(CO) ₂ ⁺² ² Co(CO) ₃ ⁺² ² Co(CO) ₄ ⁺² ² Co(CO) ₅ ⁺² ² Co(CO) ₆ ⁺²	² Co(NH ₃) ₁ ⁺² ² Co(NH ₃) ₂ ⁺² ² Co(NH ₃) ₃ ⁺² ² Co(NH ₃) ₄ ⁺² ² Co(NH ₃) ₅ ⁺² ² Co(NH ₃) ₆ ⁺²	² Co(H ₂ O) ₁ ⁺² ² Co(H ₂ O) ₂ ⁺² ² Co(H ₂ O) ₃ ⁺² ² Co(H ₂ O) ₄ ⁺² ² Co(H ₂ O) ₅ ⁺² ² Co(H ₂ O) ₆ ⁺²
M = Ni					
tier 1	¹ Ni ₂	¹ NiH ₂	¹ NiCH ₂ ¹ CH ₃ NiCO ⁺	¹ NiN ⁺ ¹ NiN ₂ ⁺²	³ NiO ¹ HNiOH
tier 2			¹ Ni(CO) ₁ ⁺² ¹ Ni(CO) ₂ ⁺² ¹ Ni(CO) ₃ ⁺² ¹ Ni(CO) ₄ ⁺² ¹ Ni(CO) ₅ ⁺² ¹ Ni(CO) ₆ ⁺²	¹ Ni(NH ₃) ₁ ⁺² ¹ Ni(NH ₃) ₂ ⁺² ¹ Ni(NH ₃) ₃ ⁺² ¹ Ni(NH ₃) ₄ ⁺² ¹ Ni(NH ₃) ₅ ⁺² ¹ Ni(NH ₃) ₆ ⁺²	¹ Ni(H ₂ O) ₁ ⁺² ¹ Ni(H ₂ O) ₂ ⁺² ¹ Ni(H ₂ O) ₃ ⁺² ¹ Ni(H ₂ O) ₄ ⁺² ¹ Ni(H ₂ O) ₅ ⁺² ¹ Ni(H ₂ O) ₆ ⁺²

^a Tier 1 molecules are used to generate the diatomic repulsive potential curve, and tier 2 molecules are used to adjust the repulsive curve to reproduce B3LYP binding energies.

used at this stage of the parametrization to avoid additional complication; however, in certain cases, some open-shell molecules have also been included. Following the standard DFTB parametrization procedure, the thereby determined segments of the two-center repulsive potentials were connected to yield a continuous curve $E_{AB}^{\text{rep}}(R)$ that was shifted up or down in energy so that the DFTB energetics of the larger test molecules (called tier 2) and also in some cases for some tier 3 molecules (see the next section for their definition) reasonably reproduces that of the DFT benchmark calculations at the B3LYP/SDD+6-31G(d) level (see the next section for definition of the basis set). Since $E_{AB}^{\text{rep}}(R)$ has to be zero at $R = \infty$, the $E_{AB}^{\text{rep}}(R)$ curve determined above was

Scheme 1. Schematic Representation of the Geometrical Parameters of the Set of Tier 3 Molecules for M = Ti, Fe, Co, and Ni^a

^a For symmetric structures, only the unique parameters are given.

extrapolated smoothly to zero as R becomes large. The choice of various test molecules as well as the amount of the repulsion potential shift and the way of extrapolation are “empirical” procedures to determine the reliability of the DFTB parameters.

In the standard DFTB parametrization procedure all the two-center parameters, i.e., the overlap $S_{\mu\nu}$ and Hamiltonian $H_{\mu\nu}^0$ matrix elements between a set of valence orbitals μ and ν as well as the charge–charge interaction parameter γ_{AB} and the core–core repulsion E_{AB}^{rep} between the two atomic centers A and B are given in the tables as functions of interatomic distances.

Following the procedure outlined above, we have developed new spin-polarized DFTB parameters for transition-metal compounds containing Sc, Ti, Fe, Co, and Ni in combination with C, H, N, and O nonmetal elements as well as with themselves. These parameter tables have already been made available to the public free of charge at www.dftb.org.²⁵

C. Tests of Determined Parameters. The newly developed parameter sets (used together with the DFTB parameters determined previously for the C, H, N, and O set) are tested against DFT results for a set of relatively small (called tier 3) molecules as well as larger, more realistic (tier 4) molecules. In the tier 3 set of molecules, we include strongly bonded small molecules as well as weakly bonded complexes. Some of these molecules are only hypothetical and are not known experimentally. In tier 4, larger compounds that are of greater interest for practical chemistry applications are chosen. Details concerning tier 4 molecules are given in each of the sections dealing with the individual metal atoms. Schematic structures of the molecule sets in tiers 3 and 4 are presented in Scheme 1 and Figures 1–4.

Benchmark DFT calculations are carried out with the B3LYP functional with a mixed basis set, Stuttgart/Dresden ab initio pseudopotential and (8s7p6d1f)/[6s5p3d1f] Gaussian valence basis set (SDD)^{64,65} for transition-metal elements and the popular 6-31G(d) basis set for H, C, N, and O, unless otherwise noted. The mixed basis sets will be denoted as SDD+6-31G(d) in the remainder of this work. All DFT geometry optimizations have been performed using the Gaussian03⁶⁶ suite of programs, and the DFTB geometry

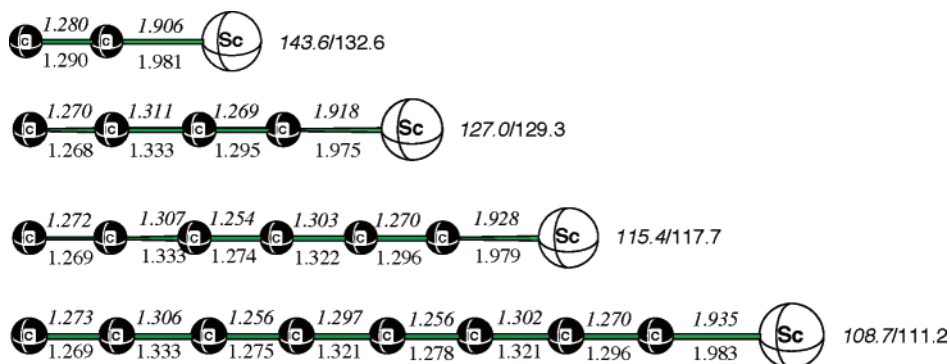


Figure 1. B3LYP/6-311+G(d) and SDFTB optimized bond distances (in Å) and Sc-C binding energies (in kcal/mol) for the electronic 2S ground state of Sc(CC) $_n$ species, $n = 1-4$. Italic and plain values denote the DFT and SDFTB results, respectively.

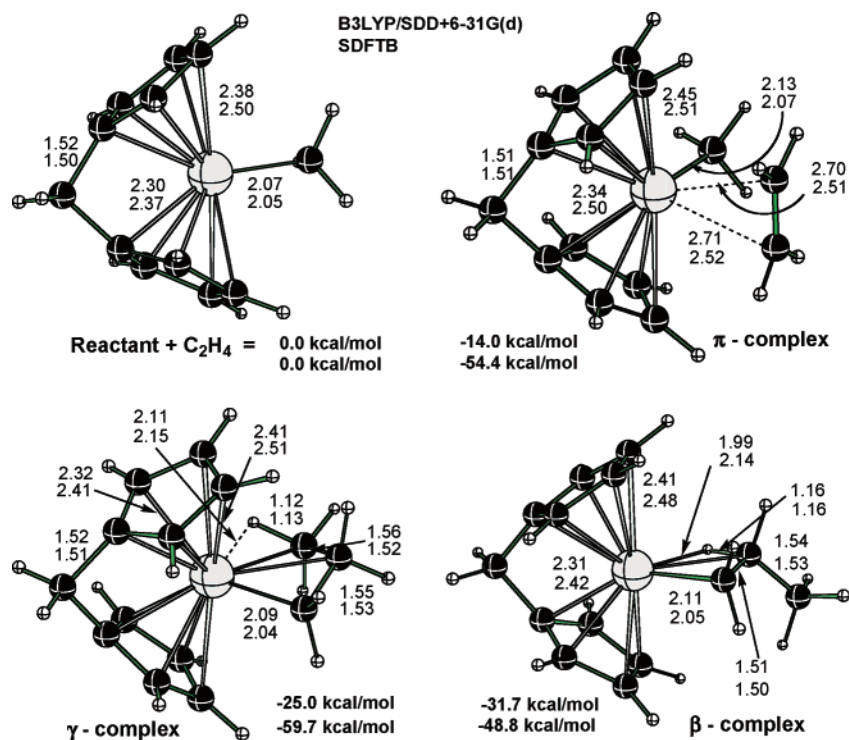


Figure 2. B3LYP/SDD+6-31G(d) (upper numbers) and SDFTB (lower numbers) optimized geometries (distances in Å) and energetics (in kcal/mol) of the reactant, intermediates, and product of the reaction $[(\text{Cp}-\text{CH}_2-\text{Cp})\text{Ti}(\text{CH}_3)]^+ + \text{C}_2\text{H}_4 \rightarrow [(\text{Cp}-\text{CH}_2-\text{Cp})\text{Ti}(\text{CH}_2\text{CH}_2\text{CH}_3)]^+$.

optimizations were carried out using our own DFTB code.³⁸ Default values for gradient and displacement convergence criteria were applied throughout.

3. Results of Test Calculations

In this section we test the ability of DFTB using the presently developed parameters to reproduce common DFT (namely B3LYP/SDD+6-31G(d)) results, such as bond lengths, angles, and relative energetics. We emphasize that it is not the purpose of the present paper to discuss the ability of spin-polarized DFTB to reproduce experimental results but rather to investigate how far the approximations introduced in DFTB cause deviations from the benchmark DFT calculations. Therefore, available literature data on test molecules will not be discussed. We will only check the performance of DFTB based on the results compared with those at the B3LYP/SDD+6-31G(d) level (hereafter this level is simply

called as DFT), unless otherwise noted. This was also the method used for evaluating the repulsive diatomic DFTB potentials. We compare the bond distances and angles for tier 3 molecules as well as the relative energies of low-lying spin states, since these are very important for transition-metal complexes. We did not compare simple bond dissociation energies such that $R_n\text{M}-\text{XR}'_m$, because often single-determinantal wave functions give incorrect spin states and make the direct energy comparison difficult.

A. Scandium. We present the geometrical parameters of Sc-containing tier 3 molecules in Table 3 for DFT and DFTB as well as the respective difference between the two levels of theory. We have dropped Sc_2O_x systems entirely as it was impossible to converge to proper wavefunctions and geometries. The absolute average bond distance difference for Sc-Sc is 0.17 Å (0.09 Å excluding very long distance in triplet $\text{Sc}_2(\text{CH}_3)_4$), Sc-H is 0.02 Å, for Sc-C 0.06 Å, for Sc-N

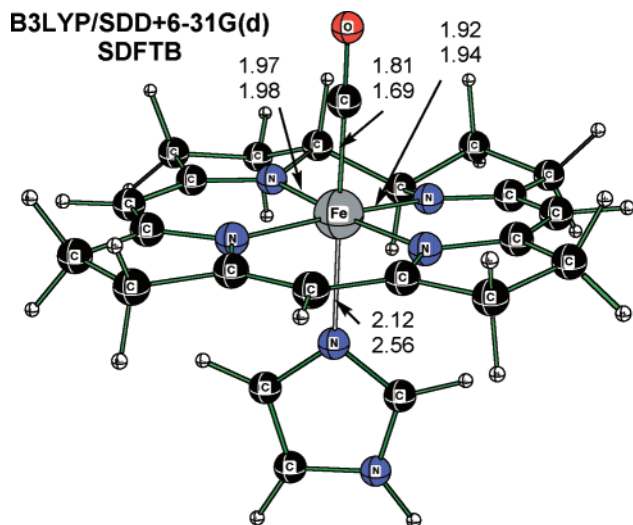


Figure 3. B3LYP/SDD+6-31G(d) (upper numbers) and SDFTB (lower numbers) optimized geometries (distances in Å) for the CO complex of Fe-porphyrin.

0.03 Å, and for Sc–O 0.03 Å. Sc–X bond lengths are therefore well described by the DFTB method with our parameters. Bond angle differences between DFT and DFTB results are on the absolute average 7.5° for Sc–Sc, 4.2° for Sc–H, 13.3° for Sc–C, 3.6° for Sc–N, and 28.4° for Sc–O. These deviations are generally much smaller than those we encountered for more d-electron rich transition-metal elements, further described below, indicating a better performance of the DFTB method when fewer d-electrons are present. The large discrepancy for the O–Sc–O bond in the quartet state of the ScO₂ molecule with 43.7° is an exception; because of the lack of more angle parameters, the average absolute value of Sc–O angle deviations is large. The overall average absolute bond distance difference between DFTB and DFT is 0.04 Å, and the overall average absolute bond angle difference is 12.4°. Therefore, generally speaking, DFTB geometries are in reasonable agreement with those predicted by DFT.

In Table 4 the dissociation energies of Sc-containing molecules are shown. The results are scattered, with some very good values and some poor values. Sc–H bridge bonds are overbound in DFTB, while Sc–N bonds are underbonds. A large error in the ScO₂ (2) → Sc (2) + O₂ (3) proves it is due to the fact SDFTB is unable to describe the triplet state of O₂ correctly.

In Table 5 the relative energies of high-spin and low-spin states of Sc-containing molecules are shown. The DFT energy orders are reproduced by DFTB except for Sc₂H and Sc⁺(η^2 -N₂), where state splittings are relatively small. Although the magnitude of state splitting difference between DFT and DFTB can be as large as 38 kcal/mol (in the case of Sc₂H), the average absolute differences between DFT and DFTB state splitting energies are 13.1 kcal/mol for Sc–Sc, 14.2 kcal/mol for Sc–H, 14.0 kcal/mol for Sc–C, 18.1 kcal/mol for Sc–N, and 15.4 kcal/mol for Sc–O compounds. This performance is better than for d-electron rich transition-metal elements, as we already noted for geometries. The overall average deviation is 15.1 kcal/mol. We report an overall tendency in DFTB to overestimate the binding

energies of low-spin complexes. Consequently, DFTB energetics should be carefully checked in the case of scandium parameters but are more reliable in general than for d-electron rich elements (see below).

As an example of a tier 4 molecule, in Figure 1, we compare B3LYP/6-311+G(d) geometries and energies of linear Sc(CC)_n (*n*=1, 2, 3, 4) in their electronic ²S ground states with the corresponding DFTB results. The DFT results were already partially presented by Redondo et al.²⁸ who however did not provide the Sc–C binding energies for this series of polyne chains. As one can see, DFTB structural results are in reasonable agreement with the B3LYP calculations, with bond differences the largest for the Sc–C bond. Here, DFTB gives bond lengths that are too long by up to 0.08 Å. As for the C–C bond lengths, the DFTB values are consistently longer compared to the B3LYP/6-311+G(d) results for both short/long alternating bond types. Energetics is in excellent agreement, with DFTB overbinding by only 3 kcal/mol, except for the special case of ScC₂ where DFTB underbinds by about 10 kcal/mol in this most strongly bound species due to the over stabilization of the C₂ unit.

B. Titanium. As shown in Table 6, the present set of Ti DFTB spin-polarized parameters leads to optimized geometries close to those obtained by DFT. The average absolute deviations between bond lengths obtained by DFT and DFTB are 0.05 Å for Ti–H, 0.06 Å for Ti–C, 0.02 Å for Ti–N, and 0.03 Å for Ti–O, respectively. Also, bond angles are reasonably described by DFTB when compared with those obtained by DFT, with the average deviation of angles for all tier 3 molecules studied here being 7.0°. However, individual angular deviations can be quite large, for instance the deviation of the DFTB Ti–H–Ti angle from the DFT angle in Ti₂H is 37.4°, leading to a too strongly bent DFTB structure in this case. Other Ti–H–Ti angles are described much better, and their deviations range from 0° to about 15°, following no obvious trend of either too sharp or too flat angles. The same holds true for Ti–X–Ti and X–Ti–X angles with X = C, N, and O, with the only exception of Ti(O₂). In this T-shaped molecule, the main failure lies in the underestimated Ti–O bond distance in DFTB, leading to a too sharp O–Ti–O angle. Problems of DFTB with the Ti–O parameter sets obviously are encountered for such polar Ti π -complexes, which is not surprising considering the fact that tiers 1 and 2 molecule sets did not include such weak bonding situations. Overall, the performance of our Ti parameters for DFTB optimized geometries is very reasonable, especially given the fact that the change of basis sets and density functionals can result in similar deviations among DFT calculations. Therefore, we conclude that the geometry performance of DFTB is acceptable for the Ti–X systems.

In Table 7 the dissociation energies of Ti-containing compounds are shown. All the bonds seem to be substantially overbound with the DFTB methods.

Relative energies (relative to the respective high-spin states) of the lower-lying electronic states of tier 3 Ti-containing molecules for DFT and DFTB as well as the absolute deviation between relative energies for the two respective methods are given in Table 8. The relative energy

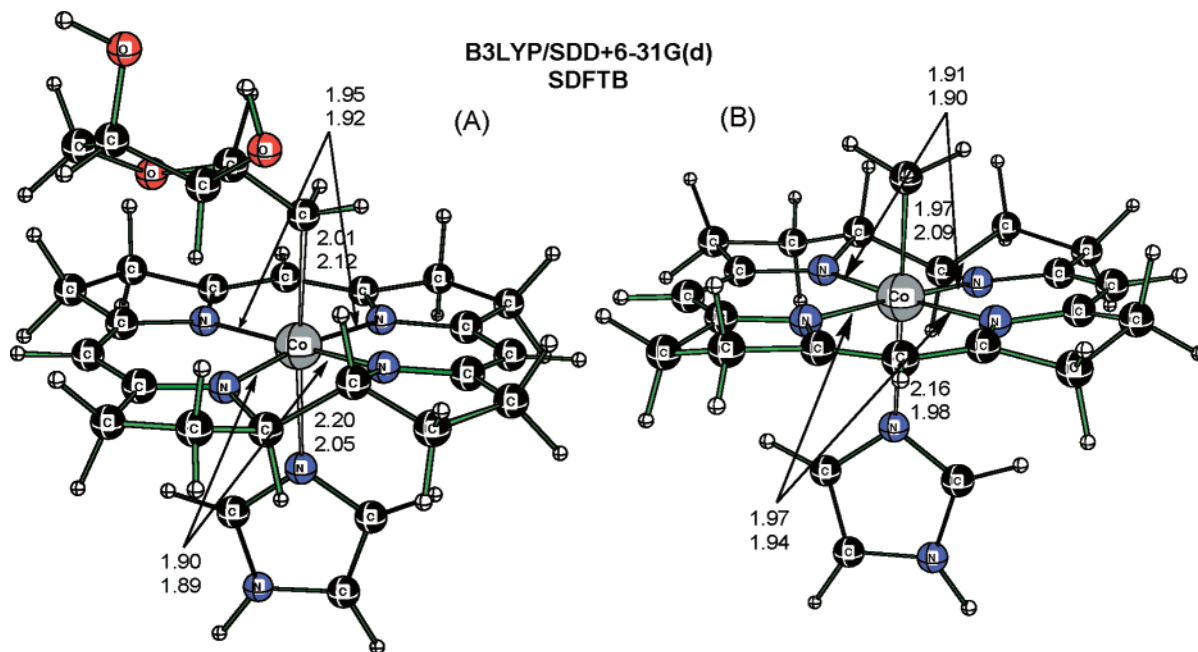


Figure 4. B3LYP/SDD+6-31G(d) (upper numbers) and SDFTB (lower numbers) optimized geometries (distances in Å) geometries of adenosylcobalamin and methylcobalamin.

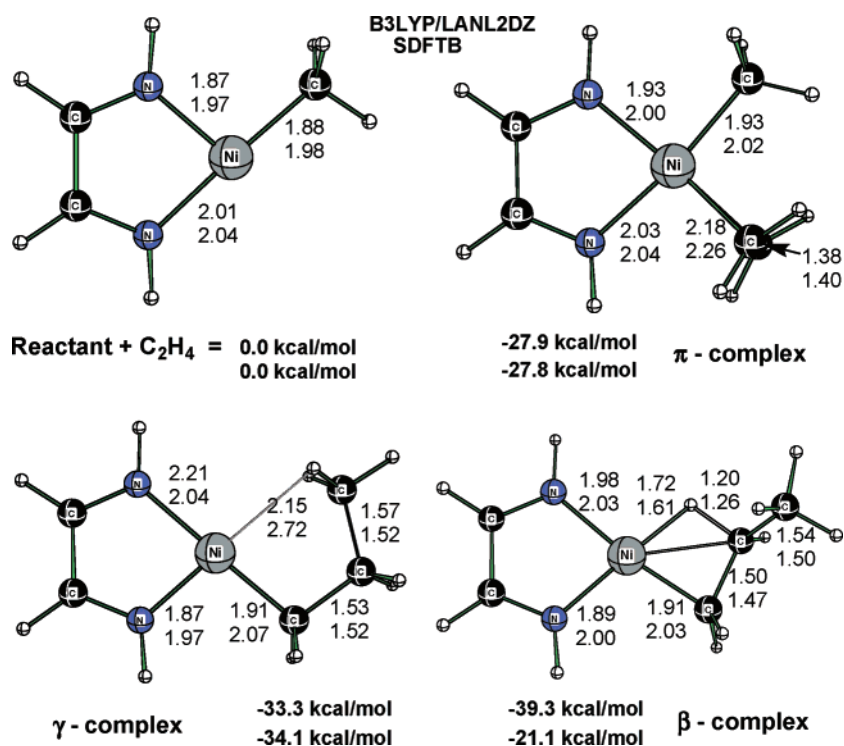


Figure 5. B3LYP/Lan12DZ (upper numbers) and SDFTB (lower numbers) optimized geometries (distances in Å) and energetics (in kcal/mol) of the reactant, intermediates, and product of the ethylene insertion step of ethylene polymerization: $[(\text{NHCHCHNH})\text{NiCH}_3]^+ + \text{CH}_2=\text{CH}_2 \rightarrow [(\text{NHCHCHNH})\text{NiCH}_2\text{CH}_2\text{CH}_3]^+$.

order between high- and low-spin states predicted by DFT is reproduced by DFTB in most cases. However, the difference between DFTB and DFT relative energies can be as large as 25 kcal/mol, as was encountered for the TiO molecule where the low-spin state is appreciably overstabilized in the DFTB method. Partially this difference can be explained by the well-known fact that B3LYP shows a preference for the high-spin state due to the inclusion of exact Hartree–Fock exchange, whereas spin-dependent atomic

parameters in DFTB are derived from the nonhybrid PBE density functional. A similar tendency for low-spin state stabilization is also seen in the case of the molecules Ti_2H_2 , Ti_2H , and $\text{Ti}(\text{C}_2\text{H}_4)^+$, where the B3LYP high-spin states are actually lower in energy than the respective low-spin states, while DFTB predicts a reverse energetic ordering. However, these molecules feature relatively small spin state splittings in DFT (smaller than 10 kcal/mol), and the sign change in DFTB is therefore within the average absolute deviation of

Table 3. DFTB and DFT (B3LYP/SDD+6-31G(d)) Optimized Bond Lengths (Å) and Valence Angles (°) of Sc-Containing Tier 3 Molecules, for the Geometry Parameters Defined in Scheme 1

compound	multiplicity	parameter	DFT	SDFTB	$\Delta(\text{DFTB}-\text{DFT})$	
Sc-Sc						
Sc ₂ (CH ₃) ₂	1	r	2.79	2.75	-0.04	
		r ₁	2.17	2.18	0.01	
		α	109.8	103.1	-6.7	
	3	r	2.57	2.76	0.19	
		r ₁	2.18	2.18	0.00	
		α	180.0	180.0	0.0	
Sc ₂ (CH ₃) ₄	1	r	2.81	2.78	-0.03	
		r ₁	2.16	2.17	0.01	
		α	115.7	106.3	-9.4	
	3	r	3.19	2.77	-0.42	
		r ₁	2.18	2.19	0.01	
		α	121.6	107.9	-13.7	
Sc-H						
ScH	1	r	1.74	1.77	0.03	
	3	r	1.84	1.83	-0.01	
ScH ₂	2	r	1.81	1.81	0.00	
		α	118.5	123.6	-5.1	
	4	r	1.96	1.90	0.06	
		α	180.0	180.0	0.0	
Sc ₂ H ₂	1	r	1.96	1.99	0.03	
		α	75.4	72.7	-2.7	
	3	r	1.97	1.98	0.01	
		α	74.1	84.5	-10.4	
	Sc ₂ H	2	r	1.95	1.98	0.03
			α	73.2	73.8	0.6
Sc ₂ H ₄	4	r	1.93	1.99	0.06	
		α	84.4	87.0	2.6	
	1	r	1.97	1.99	0.02	
		r ₁	1.84	1.83	-0.01	
	r ₂	1.97	1.99	0.02		
	α	48.7	46.5	-2.2		
	α_1	137.7	128.0	-9.7		
Sc-C						
Sc(CH ₃) ₂ ⁺	1	r	2.09	2.08	-0.01	
		α	104.2	105.1	0.9	
	3	r	2.33	2.21	-0.12	
Sc(C ₂ H ₄) ⁺	1	r	2.07	2.08	0.01	
		α	113.1	138.8	25.7	
	3	r	2.36	2.27	-0.11	
CpSc(C ₂ H ₄) ⁺	2	r ₁	2.36	2.27	-0.11	
		r	2.40	2.30	-0.10	
ScCp ⁺	2	r ₁	2.40	2.35	-0.05	
		r ₁	2.35	2.30	-0.05	
4	r ₁	2.35	2.30	-0.05		
Sc-N						
Sc(NH ₂) ₂ ⁺	1	r	1.88	1.81	-0.07	
		α	176.5	180.0	3.5	
	3	r	1.83	1.85	0.02	
		α	180.0	180.0	0.0	
Sc(NH) ₂	2	r	1.84	1.86	0.02	
		α	104.9	115.8	10.9	
	4	r	1.96	1.95	-0.01	
		α	180.0	180.0	0.0	
Sc ⁺ (η^1 -N ₂)	1	r	2.06	2.02	-0.04	
	3	r	2.09	2.04	-0.05	
Sc ⁺ (η^2 -N ₂)	1	r	2.03	2.04	0.01	

Table 3. (Continued)

compound	multiplicity	parameter	DFT	SDFTB	$\Delta(\text{DFTB}-\text{DFT})_{\text{Sc}-\text{O}}$
ScO	3	r	2.17	2.14	-0.03
	2	r	1.66	1.67	0.01
ScO ₂	4	r	1.86	1.92	0.06
	2	r	1.77	1.78	0.01
		α	127.1	114.1	-13.0
Sc(O ₂)	4	r	1.92	1.89	-0.03
		α	122.9	79.2	-43.7
	2	r	1.85	1.90	0.05
	4	r	2.11	2.08	-0.03

Table 4. SDFTB and DFT (B3LYP/SDD+6-31G(d)) Dissociation Energies of Tier 3 Sc-Containing Molecules

dissociation process ^a	dissociation energy (kcal/mol)		
	DFT	DFTB	$\Delta(\text{DFTB}-\text{DFT})$
Sc-H			
Sc ₂ H ₂ (1) → 2 ScH (1)	41.3	64.6	23.3
Sc ₂ H ₄ (1) → 2 ScH ₂ (2)	48.4	68.2	19.8
Sc-C			
Sc(CH ₃) ₂ ⁺ (1) → Sc ⁺ (3) + C ₂ H ₆	34.6	36.7	2.2
Sc(C ₂ H ₄) ⁺ (1) → Sc ⁺ (3) + C ₂ H ₄	35.0	36.4	1.4
Sc-N			
Sc(NH ₂) ₂ ⁺ (1) → Sc ⁺ (3) + N ₂ H ₄	137.1	142.7	5.6
Sc(NH) ₂ (2) → Sc (2) + N ₂ H ₂	180.4	136.5	-43.9
Sc ⁺ (N ₂) (1) → Sc ⁺ (3) + N ₂	9.4	12.5	3.1
Sc-O			
ScO ₂ (2) → Sc (2) + O ₂ (3)	129.0	176.0	47.1

^a The numbers in parentheses represent the spin multiplicity.

13.7 kcal/mol. Therefore we conclude that DFTB predicts the relative energy order between high- and low-spin states in most cases reasonably well.

As to a tier 4 system, we tested one specific reaction [(Cp-CH₂-Cp)TiCH₃]⁺ + C₂H₄ → [(Cp-CH₂-Cp)Ti(CH₂CH₂-CH₃)]⁺ exemplifying a polymerization processes involving a Ti catalyst. DFT and DFTB geometries as well as respective energetics are presented in Figure 2. In this “real-life” scenario, again we find that the DFT geometries of Ti-containing species are reasonably well reproduced by DFTB with bond length differences of at most about 0.1 Å. However, the relative stability of these complexes as predicted by DFT is not reproduced by DFTB, which shows a strong tendency to overbinding of ethylene and results in smearing out subtle energetic differences of a few kcal/mol between isomeric complexes that are predicted by DFT. This finding shows that DFTB binding energies are not as reliable as geometrical parameters and have to be used with great caution.

C. Iron. In Table 9, the geometries of tier 3 molecules optimized at DFT and DFTB levels are listed. The average absolute deviation of DFTB results from DFT is 0.09 Å for Fe-H, 0.08 Å for Fe-C, 0.10 Å for Fe-N, and 0.06 Å for Fe-O bond distances. These values are again within 0.1 Å, which we consider to be acceptable, considering comparable geometrical changes introduced by the change of basis set and/or density functional for DFT calculations. Bond angles perform better for X-Fe-X and Fe-X-Fe than for the

Table 5. SDFTB and DFT (B3LYP/SDD+6-31G(d)) Energies (Relative to the Respective High-Spin States) of the Low-Lying Electronic States of Tier 3 Sc-Containing Molecules

compound	multiplicities ^a	relative energies (kcal/mol)		
		DFT	SDFTB	$\Delta(\text{DFTB}-\text{DFT})$
Sc-Sc				
Sc ₂ (CH ₃) ₂	3 → 1	7.2	-13.0	-20.3
Sc ₂ (CH ₃) ₄	3 → 1	-9.5	-15.3	-5.8
Sc-H				
ScH	3 → 1	-4.3	-2.1	2.2
ScH ₂	4 → 2	-87.6	-63.0	24.7
Sc ₂ H ₂	3 → 1	6.7	0.3	-6.4
Sc ₂ H	4 → 2	8.8	-29.4	-38.2
Sc ₂ H ₄	3 → 1	-3.8	-5.4	-1.6
Sc-C				
Sc(CH ₃) ₂ ⁺	3 → 1	-51.4	-62.4	-10.9
Sc(C ₂ H ₄) ⁺	3 → 1	-1.4	-24.0	-22.6
ScCp ⁺	4 → 2	-74.1	-65.6	8.5
Sc-N				
Sc(NH ₂) ₂ ⁺	3 → 1	-49.2	-71.2	-22.1
Sc(NH) ₂	4 → 2	-45.7	-71.5	-25.8
Sc ⁺ (η^1 -N ₂)	3 → 1	21.5	11.6	-10.0
Sc ⁺ (η^2 -N ₂)	3 → 1	5.9	-8.7	-14.6
Sc-O				
ScO	4 → 2	-76.8	-94.5	-17.7
ScO ₂	4 → 2	-58.5	-86.2	-27.7
Sc(O ₂)	4 → 2	-43.9	-44.7	-0.8

^a For instance, 3 → 1 means that the energy of the singlet (low-spin) state relative to the triplet (high-spin) state.

corresponding Ti systems, with average absolute deviations of 9.6° for Fe-H, 6.5° for Fe-C, 12.9° for Fe-N, and 11.2° for Fe-O systems. The largest deviations in bond angles are actually found for Fe(NH₂)₂ and FeO₂ systems with about 30°. For these compounds, qualitatively different geometries are predicted by DFTB when compared to DFT (bent structure vs linear or vice versa). This difference may stem from the fact that in DFTB parametrization the d⁷s¹ configuration is used, which prefers a linear structure arising from sd hybridization. Concerning the overall performance of DFTB for geometrical parameters however, we find that bond distances and angles of DFT geometries are typically well reproduced by DFTB.

In Table 10 the dissociation energies of Fe-containing compounds are given. Bridged Fe-H bonds are underbound while the terminal Fe-H bonds are overbound, while Fe-C

Table 6. SDFTB and DFT (B3LYP/SDD+6-31G(d)) Optimized Bond Lengths (Å) and Valence Angles (°) of Ti-Containing Tier 3 Molecules, for the Geometry Parameters Defined in Scheme 1

compound	multiplicity	parameter	DFT	SDFTB	$\Delta(\text{DFTB}-\text{DFT})$
Ti-H					
TiH	2	r	1.68	1.74	0.06
	4	r	1.84	1.76	-0.08
TiH ₂	1	r	1.75	1.71	-0.04
	3	α	106.9	111.9	5.0
		r	1.78	1.74	-0.04
Ti ₂ H	2	α	122.3	108.3	-14.0
		r	1.86	1.89	0.03
	4	α	102.0	64.6	-37.4
Ti ₂ H ₂	1	r	1.82	1.93	0.11
		α	83.0	75.1	-7.9
	3	r	1.86	1.80	-0.06
Ti ₂ H ₄	3	α	48.3	57.8	9.5
		r	1.87	1.88	0.01
	1	α	57.2	57.3	0.1
		r	1.85	1.87	0.02
Ti(CH ₃) ₂	1	r ₁	1.74	1.75	0.01
		α	58.0	55.2	-2.8
	3	Ti-C			
Ti(CH ₃) ₂	1	r	2.04	2.05	0.01
		α	110.7	112.5	1.8
Ti(C ₂ H ₄) ⁺	2	r	2.18	2.08	-0.10
		α	117.1	114.8	-2.3
Ti(C ₂ H ₄) ⁺	4	r	2.03	2.00	-0.03
		r	2.34	2.26	-0.08
TiCp ⁺	1	r ₁	2.26	2.20	-0.06
		3	r ₁	2.27	2.27
Ti-N					
Ti(NH ₂) ₂ ⁺	2	r	1.85	1.84	-0.01
		α	118.3	115.1	-3.2
Ti(NH) ₂	1	r	1.71	1.70	-0.01
		α	114.8	117.7	3.1
Ti ⁺ (η^1 -N ₂)	2	r	1.99	2.00	0.01
Ti-O					
TiO	1	r	1.59	1.59	0.00
		3	r	1.61	1.61
Ti ₂ O ₂	1	r	1.81	1.91	0.10
		α	51.4	52.0	0.6
TiO ₂	1	r	1.64	1.64	0.00
		α	117.7	111.5	-6.2
Ti(O ₂)	1	r	1.79	1.81	0.02
		α	49.3	56.4	7.1
Ti ₂ O ₄	1	r	1.84	1.84	0.00
		r ₁	1.63	1.63	0.00
	α	42.6	47.7	5.1	

are acceptable. Both terminal and bridged FeO bonds seem to be grossly overbound.

The relative energies between different spin states of the Fe-containing tier 3 molecules are shown in Table 11. DFTB predicts usually the same energetic order as the one computed by DFT. Fe₂H, FeO, FeO₂, Fe(O₂), and Fe₂O₄ molecules are an exception to this rule with a reversed energy order of low- and high-spin states. Similarly to

Table 7. SDFTB and DFT (B3LYP/SDD+6-31G(d)) Dissociation Energies of Tier 3 Ti-Containing Molecules

dissociation process	dissociation energy (kcal/mol)		
	DFT	DFTB	$\Delta(\text{DFTB}-\text{DFT})$
Ti-H			
Ti ₂ H ₂ (3) → 2 TiH (4)	73.2	106.9	33.7
Ti ₂ H ₄ (1) → 2 TiH ₂ (3)	44.1	111.8	67.7
Ti-C			
Ti(C ₂ H ₄) ⁺ (4) → Ti ⁺ (2) + C ₂ H ₄	75.6	88.9	13.3
CpTi(C ₂ H ₄) (2) → TiCp (4) + C ₂ H ₄	35.9	63.1	27.2
Ti-N			
Ti(NH) ₂ (1) → Ti (3) + N ₂ H ₂	98.8	137.0	38.2

Table 8. SDFTB and DFT (B3LYP/SDD+6-31G(d)) Energies (Relative to the Respective High-Spin States) of the Low-Lying Electronic States of Tier 3 Ti-Containing Molecules

compound	multiplicities ^a	relative energies (kcal/mol)		
		DFT	SDFTB	$\Delta(\text{DFTB}-\text{DFT})$
Ti-H				
TiH	4 → 2	1.2	19.7	18.5
TiH ₂	3 → 1	39.8	14.2	-25.6
Ti ₂ H ₂	3 → 1	2.1	-10.0	-12.1
Ti ₂ H ₄	3 → 1	24.9	45.8	20.9
Ti-C				
Ti(CH ₃) ₂	3 → 1	5.9	12.7	6.8
Ti(C ₂ H ₄) ⁺	4 → 2	8.2	-3.6	-11.8
TiCp ⁺	3 → 1	12.0	13.6	1.6
Ti-O				
TiO	3 → 1	31.4	6.4	-25.0

^a For instance, 3 → 1 means that the energy of the singlet (low-spin) state relative to the triplet (high-spin) state.

Ti, B3LYP generally favors high-spin states when compared with the DFTB approach. This is however not true for all cases; for instance, the quartet state of Fe(η^2 -N₂) is 38.3 kcal/mol lower in energy relative to the doublet state in DFTB than in DFT. In general, relative energy differences between high-spin state and low-spin state between DFT and DFTB can be as large as 40 kcal/mol.

As a tier 4 molecule, binding of CO to a heme molecule with an axial histidine residue has been investigated. The structure of this complex is shown in Figure 3. The DFT geometry is well reproduced by DFTB. The only exceptions are the Fe-N_{imidazole} and Fe-C distance trans to Fe-N_{imidazole}, which are 0.44 Å too long and 0.12 Å too short, respectively, in DFTB. The computed binding energy of CO is 55.7 kcal/mol for DFT, while for DFTB it is only 26.5 kcal/mol despite the short Fe-CO distance. This is in contrast to the case of π and σ bonding of ethylene to a Ti complex discussed above, where DFTB predicts generally too large binding energies. Again, DFTB energetics may have to be used with great caution.

D. Cobalt. The structural parameters of Co-containing tier 3 molecules for both DFT as well as DFTB, and the respective relative differences are listed in Table 12. As observed for the cases of Sc, Ti, and Fe, the Co DFTB

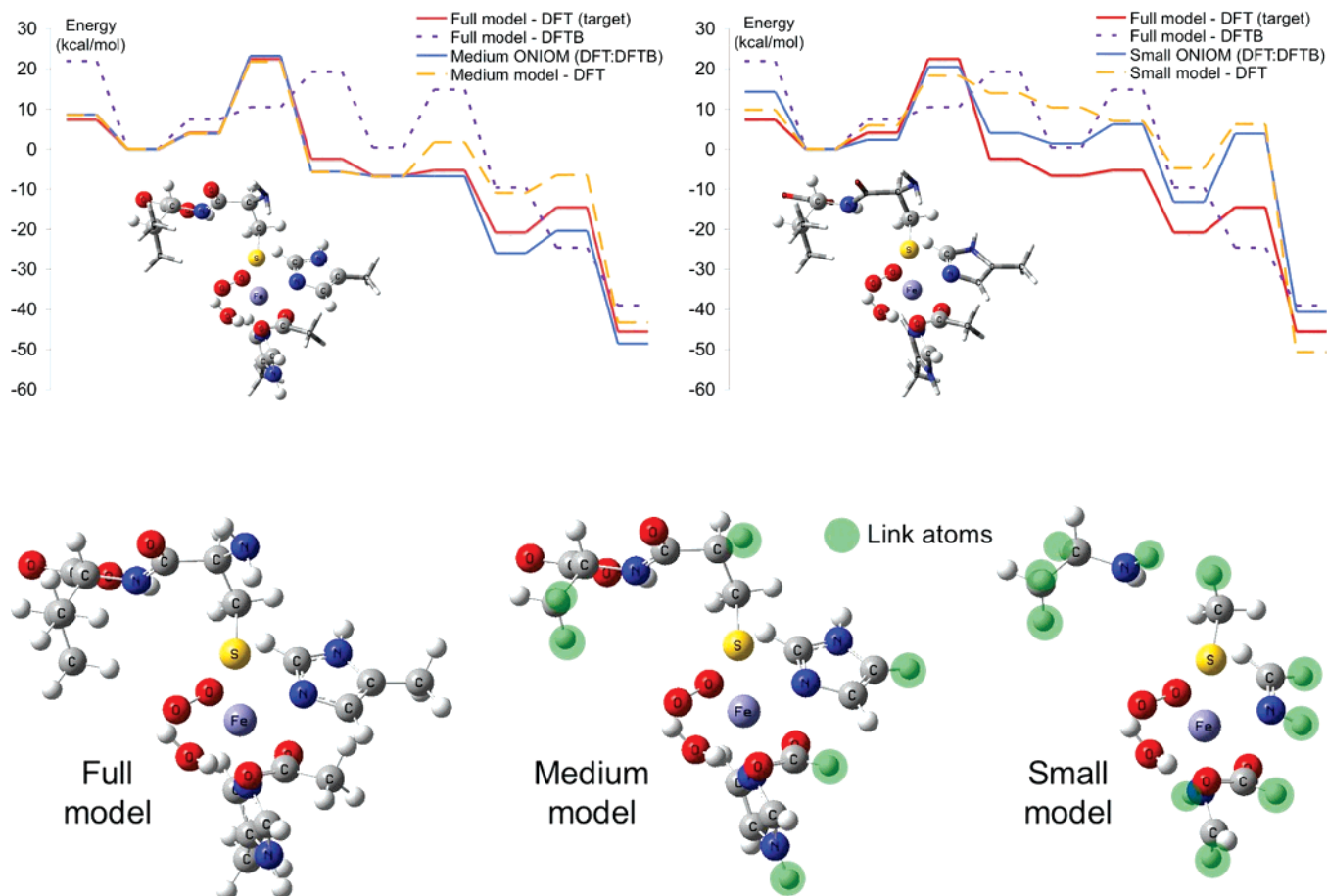


Figure 6. Stationary points along the reaction coordinate for a redox reaction involving iron. The red line represents B3LYP/6-31G(d) energies, the blue line represents DFTB energies, while the purple line shows the results from ONIOM (DFT:DFTB). Energies are aligned at the second stationary point. DFT and DFTB-only calculations include all atoms in the model. For the ONIOM system, the high-level part of is shown in ball-and-stick representation, while the DFTB part is shown in licorice representation.

geometries are in good agreement with DFT optimized structures. Compared with DFT results, the average absolute deviation of bond distance is 0.04 Å for Co–H, 0.06 Å for Co–C, 0.03 Å for Co–N, and 0.01 Å for Co–O. For bond angles, the average absolute deviation of DFTB from DFT is 5.0° for Co–H, 5.7° for Co–C, 9.9° for Co–N, and 6.9° for Co–O parameters. Some linear structures are preferred in DFTB results presumably due to d^8s^1 Co atomic configuration used in parametrization, a phenomenon described above for Fe. Yet, X–Co–X and Co–X–Co angles are generally in better agreement with DFT structural parameters than for corresponding Fe and Ti systems. We cannot comment at this stage on the origin of this exceptional good performance of Co DFTB parameters. Overall, from the average deviation values discussed above, we conclude that DFTB very reasonably reproduces DFT geometries in the case of Co-containing compounds.

The dissociation energies of Co-containing compounds are shown in Table 13. Co–H bonds are overbound, while Co–C, C, and O bonds are all overbound.

The relative energy order between high-spin and low-spin states for different Co containing molecules are summarized in Table 14. The relative energy orders in DFT are well reproduced by DFTB except CoH_2 , $\text{Co}(\text{CH}_3)_2$, $\text{CpCo}(\text{C}_2\text{H}_4)^+$, CoCp^+ , and $\text{Co}(\text{NH}_2)_2$. Considering their high-spin states are

favored by B3LYP, DFTB reasonably predict the relative energy order although the absolute values deviation is 20.7 kcal/mol on the average, with the largest difference being about 54 kcal/mol in the case of CoCp^+ . Particularly noticeable is the DFTB preference for low-spin states in the case of Co–C systems, but noticeable exceptions from this rule exist, for instance Co_2H_4 , where DFTB favors the high-spin state by 21 kcal/mol relative to DFT.

As a real case tier 4 system, binding of adenosyl and methyl groups, respectively, to cobalamin has been investigated. The molecular structures of adenosylcobalamin and methylcobalamin are shown in Figure 4. The geometries in both structures are well described by DFTB compared with the corresponding structures from DFT. The largest bond distance difference is about 0.18 Å in only one case. Again, DFTB predicts Co–N bond distance orders correctly. The binding energy of the adenosyl group to cobalamin is 70.3 kcal/mol in DFTB, that is 12.5 kcal/mol higher in comparison to 57.8 kcal/mol in DFT. Similarly, the binding energy of the methyl group to cobalamin is 94.3 kcal/mol in DFTB, that is an overbinding of 22.2 kcal/mol when compared to 72.1 kcal/mol in DFT. Thus, it is concluded that DFTB performs well in terms of geometries. Again we should

Table 9. SDFTB and DFT (B3LYP/SDD+6-31G(d)) Optimized Bond Lengths (Å) and Valence Angles (°) of Fe-Containing Tier 3 Molecules, for the Geometry Parameters Defined in Scheme 1

compound	multiplicity	parameter	DFT	SDFTB	$\Delta(\text{DFTB}-\text{DFT})$
Fe-H					
FeH	2	r	1.59	1.50	-0.09
	4	r	1.56	1.54	-0.02
FeH ₂	3	r	1.54	1.51	-0.03
		α	102.4	96.9	-5.5
	5	r	1.65	1.62	-0.03
		α	169.7	153.0	-16.7
Fe ₂ H	2	r	1.68	1.66	0.02
		α	46.45	50.79	4.34
Fe ₂ H ₄	1	r	1.52	1.52	0.00
		r ₁	1.65	1.49	0.16
		α	46.79	63.9	17.11
Fe ₂ H ₄	3	r	1.58	1.60	0.02
		r ₁	1.61	1.72	0.11
		α	47.27	48.87	1.60
Fe-C					
Fe(CH ₃) ₂	1	r	1.92	2.06	0.14
		α	117.7	112.2	-5.5
	3	r	1.94	2.07	0.13
		α	112.1	109.3	-2.8
		5	r	2.05	2.13
Fe(C ₂ H ₄) ⁺	2	α	180.0	180.0	0.0
		r	2.05	2.13	0.08
		r	2.07	2.18	0.11
FeCp ⁺	3	r ₁	2.19	2.26	0.07
	5	r ₁	2.23	2.32	0.09
Fe-N					
Fe(NH ₂) ₂	1	r	1.79	1.76	0.03
		α	180.0	150.7	-29.3
	5	r	1.85	1.86	0.01
		α	180.0	180.0	0.0
Fe(NH) ₂	3	r	1.65	1.60	-0.05
		α	171.2	180.0	8.8
	5	r	1.67	1.76	0.09
		α	121.1	142.1	21.0
Fe(η^1 -N ₂) ⁺	4	r	2.09	1.92	-0.17
		α	0.00	0.01	0.01
Fe-O					
FeO	1	r	1.59	1.58	-0.01
	3	r	1.57	1.61	0.04
	5	r	1.61	1.66	0.05
Fe ₂ O ₂	1	r	1.73	1.80	0.07
		α	42.8	51.8	9.0
	3	r	1.76	1.81	0.05
		α	42.8	52.6	9.8
FeO ₂	1	r	1.54	1.61	0.07
		α	145.2	167.6	22.4
	3	r	1.58	1.63	0.05
		α	140.4	149.4	9.0
		5	r	1.60	1.67
Fe(O ₂) ⁺	4	α	118.7	126.0	7.3
		r	1.82	1.81	-0.01
Fe ₂ O ₄	1	α	43.8	44.1	0.3
		r ₁	1.56	1.57	0.01
		r	1.72	1.81	0.09
	3	α	47.6	44.4	-3.2
		r ₁	1.53	1.59	0.06
		r	1.74	1.80	0.06
		α	43.9	45.4	1.5

Table 10. SDFTB and DFT (B3LYP/SDD+6-31G(d)) Dissociation Energies of Tier 3 Fe-Containing Molecules

dissociation process	dissociation energy (kcal/mol)		
	DFT	DFTB	$\Delta(\text{DFTB}-\text{DFT})$
Fe-H			
$\text{Fe}_2\text{H}_4(1) \rightarrow 2\text{FeH}_2(3)$	79.7	54.5	-25.2
$\text{FeH}_2(1) \rightarrow \text{Fe}(1) + \text{H}_2(1)$	86.9	105.2	18.3
$\text{FeH}_2(1) \rightarrow \text{FeH}(2) + \text{H}$	61.9	90.6	28.7
Fe-C			
$\text{Fe}(\text{CH}_3)_2(2) \rightarrow \text{Fe}(2) + \text{C}_2\text{H}_6$	21.3	34.9	13.6
$\text{Fe}(\text{C}_2\text{H}_4)(2) \rightarrow \text{Fe}(2) + \text{C}_2\text{H}_4$	56.3	52.5	-3.8
Fe-O			
$\text{Fe}_2\text{O}_2(1) \rightarrow 2\text{FeO}(1)$	84.6	133.8	49.2
$\text{Fe}_2\text{O}_4(1) \rightarrow 2\text{FeO}_2(1)$	68.9	119.2	50.3

caution about the use of DFTB for the prediction of energetics due to the unforeseeable over- or underbinding errors.

E. Nickel. The geometrical parameters of Ni-containing tier 3 molecules are shown in Table 15 for DFT and DFTB as well as the respective difference between the two levels of theory. In the case of triplet $\text{Ni}_2(\text{CH}_3)_4$, DFT predicts an asymmetric structure with one bridging methyl group, whereas the DFTB triplet geometry resembles more closely the symmetric DFTB singlet geometry. The average absolute bond distance difference is 0.15 Å for Ni-Ni, 0.06 Å for Ni-H, 0.19 Å for Ni-C, 0.04 Å for Ni-N, and 0.02 Å for Ni-O. The Ni-C distance is not well described in cases where cyclopentadienyl (Cp) rings interact with Ni. Here, Ni-C bonds are typically too long by a few tenths of an angstrom, mainly because the position of the Ni on top of the Cp system is very flexible. Bond angle differences between DFT and DFTB results are on the absolute average 16.2° for Ni-Ni, 12.2° for Ni-H, 18.8° for Ni-C, 30.8° for Ni-N, and 7.7° for Ni-O. The rather large deviations are a consequence of the fact that DFTB very often prefers linear arrangements, when the lowest DFT structure is bent (as seen also for Fe and Co above). The overall average absolute bond distance difference between DFTB and DFT is 0.09 Å, and the overall average bond angle difference is 17.1°. Therefore, more generally speaking, DFTB geometries are in reasonable agreement with those predicted by DFT, which is consistent with the findings in the case of other transition-metal elements in this work.

The dissociation energies of Ni-containing compounds are shown in Table 16. Most of the Ni bonds seem to be underbound substantially.

In Table 17, the relative energies of high-spin and low-spin states of Ni-containing molecules are shown. The DFT energy orders are reproduced by DFTB except for Ni_2H_2 , NiCp^+ , $\text{Ni}(\text{NH}_2)_2^+$, NiO_2 , and $\text{Ni}(\text{O}_2)$, where state splittings are generally very small. However, the magnitude of state splitting difference between DFT and DFTB can be as large as 50 kcal/mol (in the case of Ni_2H_4). Average absolute differences between DFT and DFTB state splitting energies are 32.4 kcal/mol for Ni-H, 14.9 kcal/mol for Ni-C, 19.8 kcal/mol for Ni-N, and 20.6 kcal/mol for Ni-O. The overall average absolute deviation is 21.9 kcal/mol. Again we report

Table 11. SDFTB and DFT (B3LYP/SDD+6-31G(d)) Energies (Relative to the Respective High-Spin States) of the Low-Lying Electronic States of Tier 3 Fe-Containing Molecules

compound	multiplicities ^a	relative energies (kcal/mol)		
		DFT	SDFTB	$\Delta(\text{DFTB}-\text{DFT})$
Fe-H				
FeH	4 → 2	43.7	33.5	-10.2
FeH ₂	3 → 1	22.0	11.7	-10.3
Fe-C				
Fe(CH ₃) ₂	3 → 1	33.1	19.7	-13.4
Fe(C ₂ H ₄) ⁺	4 → 2	41.6	32.4	-9.2
FeCp ⁺	5 → 3	13.8	16.7	2.9
Fe-N				
Fe(NH ₂) ₂	5 → 1	33.3	12.4	-20.9
Fe(NH) ₂	5 → 3	8.2	27.6	19.4
Fe(η^1 -N ₂)	4 → 2	25.5	37.1	11.6
Fe-O				
FeO	5 → 1	10.6	0.6	-10.0
Fe ₂ O ₂	3 → 1	40.4	6.0	-34.4
FeO ₂	3 → 1	26.4	8.0	-18.4
Fe(O ₂) ⁺	4 → 2	47.1	14.0	-33.1
Fe ₂ O ₄	3 → 1	7.0	0.7	-6.3

^a For instance, 3 → 1 means that the energy of the singlet (low-spin) state relative to the triplet (high-spin) state.

an overall tendency in DFTB to overestimate the binding energies of low-spin complexes. Consequently, DFTB energetics should be carefully checked in the case of nickel parameters as well.

In Figure 5, as an example of a real case tier 4 system, structures and energetics of the intermediates of an ethylene insertion step of $[\text{C}_2\text{H}_4\text{N}_2\text{NiCH}_3]^+ + \text{C}_2\text{H}_4 \rightarrow [\text{C}_2\text{H}_4\text{N}_2\text{NiCH}_2\text{-CH}_2\text{CH}_3]^+$ are presented. DFT geometries and energetics at the B3LYP/Lanl2DZ level were taken from ref 67. This system features Ni-H, Ni-C, and Ni-N interactions and similar trends as found for molecules in Table 15 can be observed. Ni-X bond lengths are generally too long by about 0.1 Å (with some exceptions). An exception is a very long agostic Ni...H distance of 2.72 Å for DFTB as compared to 2.15 Å for DFT the γ -complex; such a weak interaction does not seem to be properly parametrized. Energetically, DFTB interaction energies for π - and γ -complexes are in almost perfect agreement with DFT, but the β -complex is severely underbound relative to the γ -complex, which is in stark contrast to the DFT results. This finding underlines once again that energetics obtained at the DFTB level of theory are to be trusted only with great caution.

4. Sample Application in ONIOM(DFT:DFTB)

The deficiency in energetic prediction of spin-polarized DFTB with the present transition-metal parameters is expected to be greatly reduced in the ONIOM(QM:QM) scheme adopting DFTB as the low-level method. In this scheme the energetics of the “active” part will be calculated using a more reliable high-level method, and the energetic errors in the DFTB calculations will be mostly canceled out. The use of QM as the low-level method is in some cases essential; QM methods take into account electronic effects

Table 12. SDFTB and DFT (B3LYP/SDD+6-31G(d)) Optimized Bond Lengths (Å) and Valence Angles (°) of Co-Containing Tier 3 Molecules, for the Geometry Parameters Defined in Scheme 1

compound	multiplicity	parameter	DFT	SDFTB	$\Delta(\text{DFTB}-\text{DFT})$
Co-H					
CoH	1	r	1.54	1.52	-0.02
	3	r	1.54	1.52	-0.02
CoH ₂	2	r	1.49	1.47	-0.02
	4	α	97.4	93.9	-3.5
Co ₂ H ₂	1	r	1.59	1.58	-0.01
		α	143.2	140.3	-2.9
	3	r	1.62	1.63	0.01
		α	48.4	47.1	-1.3
Co(CH ₃) ₂	2	r	1.63	1.64	0.01
		α	47.5	47.8	0.3
	4	r	1.89	1.99	0.10
		α	114.0	106.2	-7.8
Co(C ₂ H ₄) ⁺²	4	r	1.99	2.04	0.05
		α	143.9	146.8	2.9
	6	r	2.30	2.08	-0.22
		α	2.19	2.06	-0.13
CoCp ⁺	4	r ₁	1.80	1.79	-0.01
	6	r ₁	2.30	2.52	0.22
Co-N					
Co(NH ₂) ₂	4	r	1.82	1.80	-0.02
	6	α	179.9	179.4	-0.5
Co(NH) ₂ linear	4	r	1.90	1.84	-0.06
		α	97.0	96.3	-0.7
	6	r	1.68	1.66	-0.02
		α	180.0	180.0	0.0
Co(NH) ₂ bent	4	r	1.78	1.73	-0.05
		α	179.1	180.0	0.9
	6	r	1.67	1.65	-0.02
		α	128.4	127.0	1.4
Co ²⁺ (η^1 -N ₂)	2	r	1.75	1.72	-0.03
	4	α	127.8	142.0	13.6
CoO	2	r	2.00	1.95	-0.05
	4	r	1.99	1.98	-0.01
Co-O					
CoO	2	r	1.60	1.58	-0.02
	4	r	1.59	1.61	0.02
Co ₂ O ₂	1	r	1.73	1.77	0.04
		α	45.9	42.6	-3.3
	3	r	1.74	1.78	0.04
		α	44.2	45.1	0.9
CoO ₂ linear	2	r	1.57	1.57	0.00
	α	177.7	179.5	1.8	
CoO ₂ bent	4	r	1.64	1.63	-0.01
	α	101.4	116.7	15.3	
Co(O ₂)	4	r	1.81	1.83	0.02
		α	47.9	44.7	3.2
	6	r	2.03	1.92	-0.11
		α	38.1	40.3	2.2
Co ₂ O ₄	1	r ₁	1.53	1.55	0.02
		r	1.75	1.78	0.03
	3	α	43.5	42.8	-0.7
		r ₁	1.53	1.56	0.03
	5	r	1.76	1.78	0.02
		α	44.7	41.7	-3.0
		r ₁	1.56	1.59	0.03
		r	1.75	1.78	0.03

Table 13. SDFTB and DFT (B3LYP/SDD+6-31G(d)) Dissociation Energies of Tier 3 Co-Containing Molecules

dissociation process	dissociation energy (kcal/mol)		
	DFT	DFTB	$\Delta(\text{DFTB}-\text{DFT})$
Co-H			
CoH ₂ (4) → CoH(3) + H	54.4	22.5	-31.9
Co ₂ H(4) → CoH(3) + Co(4)	13.3	8.7	-4.6
Co-C			
CoC ₂ H ₄ ⁺² → Co ⁺² + C ₂ H ₄	109.9	131.7	21.8
Co(CH ₃) ₂ (4) → Co ⁺² + CH ₃ ⁻	753.8	773.7	19.8
Co-N			
Co(NH ₂) ₂ (4) → Co ⁺² + 2NH ₂ ⁻	386.2	430.2	45.0
Co-O			
CoO ₂ (2) bent → Co + O ₂	17.4	42.7	25.3
Co(O ₂)(2) → Co + O ₂	10.4	23.0	12.6
Co ₂ O ₂ (3) → 2CoO	21.5	52.7	31.2

Table 14. SDFTB and DFT (B3LYP/SDD+6-31G(d)) Energies (Relative to the Respective High-Spin States) of the Low-Lying Electronic States of Tier 3 Co-Containing Molecules

compound	multiplicities ^a	relative energies (kcal/mol)		
		DFT	SDFTB	$\Delta(\text{DFTB}-\text{DFT})$
Co-H				
CoH	3 → 1	54.3	18.2	-36.1
CoH ₂	4 → 2	11.2	-11.2	-22.4
Co ₂ H ₂	3 → 1	25.3	10.5	-14.8
Co-C				
Co(CH ₃) ₂	4 → 2	9.2	-1.1	-10.3
Co(C ₂ H ₄) ⁺²	6 → 4	81.6	95.8	14.2
CoCp ⁺	6 → 4	21.8	46.0	24.2
Co-N				
Co(NH ₂) ₂	6 → 4	-57.7	-74.0	-16.3
Co(NH) ₂ bent	6 → 4	-8.8	-33.8	-25.0
Co(NH) ₂ linear	6 → 4	6.7	49.0	42.3
Co ²⁺ (η^1 -N ₂)	4 → 2	28.4	22.0	-6.4
Co-O				
CoO	4 → 2	39.7	7.8	-31.9
Co ₂ O ₂	3 → 1	5.3	1.1	-4.2
Co(O ₂)	6 → 4	5.2	-26.8	-32.0
Co ₂ O ₄	3 → 1	29.7	3.0	-26.7

^a For instance, 3 → 1 means that the energy of the singlet (low-spin) state relative to the triplet (high-spin) state.

of the environment and are fully polarizable. Both of these important effects are neglected completely when standard MM is used as the low-level method. Since there is virtually no reliable semiempirical method for transition-metal complexes, even "preliminary" spin-polarized DFTB parameters would be useful for ONIOM(QM:QM) calculations.

To illustrate the applicability of the ONIOM(QM:DFTB) scheme, the approach is tested on a proposed mechanism for the iron enzyme isopenicillin N synthase.³⁰ This mechanism involves several metal oxidation and reduction steps, coupled to bond breaking and bond formation in the substrate. The range of different reactions makes it a suitable test system for ONIOM(QM:DFTB).

Table 15. SDFTB and DFT (B3LYP/SDD+6-31G(d)) Optimized Bond Lengths (Å) and Valence Angles (°) of Ni-Containing Tier 3 Molecules, for the Geometry Parameters Defined in Scheme 1

compound	multiplicity	parameter	DFT	SDFTB	$\Delta(\text{DFTB}-\text{DFT})$
Ni-Ni					
Ni ₂ (CH ₃) ₂	1	r	2.19	2.22	0.03
		r ₁	1.85	1.97	0.12
		α	97.3	96.9	-0.4
	3	r	2.35	2.16	-0.19
		r ₁	1.91	2.01	0.10
		α	132.5	114.6	-17.9
Ni ₂ (CH ₃) ₄	1	r	2.46	2.24	-0.22
		r ₁	1.86	1.98	0.12
		α	120.0	109.2	-10.8
	3	r	2.30	2.16	-0.14
		r ₁	1.94	2.08	0.14
		r ₂	1.91	2.08	0.17
		r ₃	2.08	2.08	0.00
		α_1	121.0	114.1	-6.9
		α_2	121.1	114.1	-7.0
α_3	60.1	114.1	54.0		
Ni-H					
NiH	2	r	1.51	1.46	-0.05
	4	r	1.60	1.60	0.00
NiH ₂	1	r	1.53	1.54	0.00
	3	r	1.54	1.52	-0.03
α		180.0	180.0	0.0	
Ni ₂ H ₂	1	r	1.57	1.59	0.02
		α	41.4	45.1	3.3
	3	r	1.67	1.57	-0.10
		α	49.6	43.4	-6.2
Ni-C					
Ni(CH ₃) ₂ ⁺	1	r	1.94	2.04	0.10
		α	180.0	180.0	0.0
	3	r	1.96	2.03	0.06
		α	139.3	142.6	3.3
		r	2.08	2.15	0.07
Ni(C ₂ H ₄) ⁺	2	r	2.08	2.15	0.07
	4	r	2.32	2.28	-0.04
NiCp ⁺	1	r ₁	2.85	3.08	0.23
	3	r ₁	1.71	2.19	0.48
		r ₁	1.77	1.90	0.13
		α	180.0	180.0	0.0
Ni-N					
Ni(NH ₂) ₂ ⁺	2	r	1.88	1.81	-0.07
		α	176.5	180.0	3.6
	4	r	1.83	1.85	0.02
		α	180.0	180.0	0.0
Ni(NH) ₂	1	r	1.60	1.66	0.06
		α	164.9	180.0	15.1
	3	r	1.67	1.72	0.05
		α	125.4	170.6	45.2
		r	1.91	1.94	0.03
Ni ⁺ (η^1 -N ₂)	2	r	1.91	1.94	0.03
	4	r	2.44	2.17	-0.27
Ni-O					
NiO	1	r	1.61	1.62	0.01
	3	r	1.61	1.62	0.01
Ni ₂ O ₂	1	r	1.75	1.79	0.04
		α	50.2	52.1	1.9
	3	r	1.77	1.78	0.01
		α	47.9	50.3	2.4
NiO ₂	1	r	1.58	1.58	0.00
		α	159.4	180.0	30.6
	3	r	1.60	1.61	0.01

Table 15. (Continued)

compound	multiplicity	parameter	DFT	SDFTB	$\Delta(\text{DFTB}-\text{DFT})$
Ni(O ₂)	1	α	132.7	142.0	10.7
		r	1.78	1.79	0.01
	3	r	1.90	1.83	-0.07
Ni ₂ O ₄	1	r	1.76	1.78	0.02
		r ₁	1.58	1.59	0.01
		α	42.7	41.2	-1.5

Table 16. SDFTB and DFT (B3LYP/SDD+6-31G(d)) Dissociation Energies of Tier 3 Ni-Containing Molecules

dissociation process	dissociation energy (kcal/mol)		
	DFT	DFTB	$\Delta(\text{DFTB}-\text{DFT})$
Ni-H			
Ni ₂ H ₂ (3) → 2NiH (2)	38.9	-2.7	-41.6
Ni ₂ H ₄ (1) → 2NiH ₂ (3)	30.3	8.1	-22.1
Ni-C			
Ni(CH ₃) ₂ (3) → Ni (3) + C ₂ H ₆	5.4	14.5	9.1
Ni(C ₂ H ₄) ⁺ (2) → Ni ⁺ (2) + C ₂ H ₄	54.0	36.7	-17.3
CpNi(C ₂ H ₄) (2) → Ni (3) + Cp (2) + C ₂ H ₄	71.4	39.5	-31.9
NiCp ⁺ (3) → Ni ⁺ (2) + Cp (2)	62.2	16.7	-45.5
Ni-N			
Ni(NH ₂) ₂ ⁺ (2) → Ni ⁺ (2) + N ₂ H ₄	66.4	73.3	6.8
Ni(NH) ₂ (1) → Ni (3) + N ₂ H ₂	120.5	136.5	16.0
Ni ⁺ (N ₂) (2) → Ni ⁺ (2) + N ₂	28.4	25.8	-2.7

All stationary points along the reaction coordinate are initially optimized at the B3LYP/6-31G(d) level. The relative B3LYP energies are also taken as target values for the lower-level calculations. ONIOM(B3LYP:DFTB) and DFTB-only energies are then obtained by performing single-point calculations on the DFT geometries. We used two ONIOM “models” to which the “high” level (DFT) is used, while the ONIOM “real” system is the full system, as shown at the bottom of Figure 6. The medium model is a model one would usually adopt in realistic calculations, while the small model is pushing the ONIOM a little further to check what the effect would be. Two resulting potential energy profiles (using medium and small models, respectively) are shown in the top row of Figure 6; due to technical issues with the preliminary ONIOM implementation, DFTB energies could not be properly converged for all stationary points, and those points have therefore been excluded from the results. Note also that the DFTB calculations use iron-sulfur parameters that are not included in the present contribution.

It takes only a brief look at Figure 6 to see that ONIOM-(DFT:DFTB) results are generally in very good agreement with those of the target DFT calculations. For the initial steps of the reaction, deviations between DFT and ONIOM results are in most cases limited to 1 kcal/mol. This accuracy is achieved despite the fact that relative DFTB errors are 10–20 kcal/mol, as could be expected from the data presented in previous sections. The higher accuracy of DFT:DFTB is achieved by the error cancellation that is an inherent part of the ONIOM method. The difference between DFT and DFT:

Table 17. SDFTB and DFT (B3LYP/SDD+6-31G(d)) Energies (Relative to the Respective High-Spin States) of the Low-Lying Electronic States of Tier 3 Ni-Containing Molecules

compound	multiplicities ^a	relative energies (kcal/mol)		
		DFT	SDFTB	$\Delta(\text{DFTB}-\text{DFT})$
Ni-Ni				
Ni ₂ (CH ₃) ₂	3 → 1	19.5	-3.8	-23.3
Ni ₂ (CH ₃) ₄	3 → 1	14.0	-8.4	-22.4
Ni-H				
NiH	4 → 2	-26.3	-47.0	-20.7
NiH ₂	3 → 1	33.8	12.0	-21.8
Ni ₂ H ₂	3 → 1	28.6	-20.0	-48.6
Ni-C				
Ni(CH ₃) ₂ ⁺	3 → 1	35.9	10.4	-25.5
Ni(C ₂ H ₄) ⁺	4 → 2	-39.9	-45.9	5.8
NiCp ⁺	3 → 1	21.2	-9.0	-30.2
Ni-N				
Ni(NH ₂) ₂ ⁺	4 → 2	10.4	-7.3	-17.7
Ni(NH) ₂	3 → 1	-7.2	-29.6	-22.4
Ni ⁺ (η^1 -N ₂)	4 → 2	-37.5	-57.0	-19.5
Ni-O				
NiO	3 → 1	47.1	12.8	-34.3
Ni ₂ O ₂	3 → 1	29.9	1.4	-28.5
NiO ₂	3 → 1	2.7	-14.0	-16.7
Ni(O ₂)	3 → 1	11.9	-10.7	-22.6

^a For instance, 3 → 1 means that the energy of the singlet (low-spin) state relative to the triplet (high-spin) state.

DFTB results is larger for the last three stationary points in Figure 6, but only because these points describe bond breaking and formation directly adjacent to the DFTB layer. The apparent discrepancy can be removed by moving the cut between DFT and DFTB away from the reactive region.

The average absolute deviation for the ONIOM calculation using the medium model is 1.39 kcal/mol, excluding the reference and the last three stationary points. Including the last three points increases the deviation to 2.63 kcal/mol. Corresponding values for the DFTB calculation are 14.4 and 12.9 kcal/mol, respectively.

The present model was selected to show the applicability of the ONIOM(QM:DFTB) method rather than the potential benefits of an electronically active second layer. Still, the DFT:DFTB results are always better than the results from a small DFT model without the DFTB layer. The applicability and benefits of the present transition-metal parameters in ONIOM calculations will be further explored in subsequent work.

5. Summary and Conclusions

From the above given discussions in which geometries and energetics of tiers 3 and 4 molecules were presented for each transition-metal element at both the B3LYP/SDD+6-31G(d) as well as spin-polarized DFTB level of theory, we can make the following summary:

1. Spin-polarized DFTB with the present parameters for transition-metal elements Sc, Ti, Fe, Co, and Ni in combination with H, C, O, N and same-element bonding partners reproduce B3LYP/SDD+6-31G(d) geometries reasonably well, both bond distances (average absolute differences mostly below 0.1 Å) as well as angles (average absolute deviations between 10° and 20°) except for cases where DFTB noticeably prefers linear bond environments, presumably as a consequence of the atomic DFTB parameter evaluation. Problems also occur in bonding situations where particular bond types were not included in tiers 1 and 2 molecule sets, such as metal–nonmetal π -bonds. A remedy of this problem would obviously involve the inclusion of π -complexes in tiers 1 and 2 sets of molecules; however, the overall DFTB performance for geometrical parameters is likely to suffer in such a case.

2. The evaluation of dissociation energies of these small molecules into even small molecular fragments (or sometimes atoms) is very difficult. The dissociation energies are only qualitatively acceptable. Some bonds are overbound and some others underbound. Errors are as large as 4–50 kcal/mol in some cases.

3. For the energy differences between different spin states of tier 3 molecules, spin-polarized DFTB energetic orders qualitatively agree with DFT in most cases. However, for quantitative comparison, there are cases of both over- as well as underbinding by as much as 50 kcal/mol. While DFTB shows a tendency to overestimate the stability of low-spin complexes relative to their corresponding high-spin states, we found several exceptions to this rule. For tier 4 molecules, we also found both over- as well as underbinding situations of the order of tens of kcal/mol, making the use of DFTB predicted energetics only qualitative.

Therefore spin-polarized DFTB parameters for Sc, Ti, Fe, Ni, and Co in connection to H, C, N, O and own elements should be taken as “preliminary”, with a reasonable geometrical performance but with only qualitative or “ballpark” energetic reliability and should be further tested for individual cases.

4. The deficiency in energetic prediction of spin-polarized DFTB with the present transition-metal parameters is expected to be greatly reduced in the ONIOM(QM:QM) scheme adopting DFTB as the low-level method. Since there is virtually no reliable semiempirical method for transition-metal complexes, even “preliminary” spin-polarized DFTB parameters would be useful for ONIOM(QM:QM) calculations. We have demonstrated using the proposed mechanism for the iron enzyme isopenicillin N synthase. The use of QM as the low-level method is in some cases essential; QM methods take into account electronic effects of the environment and are fully polarizable; both of these important effects are neglected completely when standard MM is used as the low-level method. The applicability of the present transition-

metal parameters in the ONIOM(QM:DFTB) will be further explored in subsequent work.

In the present test, we used B3LYP/SDD+6-31G(d) for the calibration of the DFTB parameters. As is well-known, a weak point of the DFT method is the lack of an absolutely reliable functional. In particular, the amount of mixing of the “exact” (Hartree–Fock) exchange functional often controls the relative energies of different spin states. B3LYP has been used in the original parametrization of (H, C, N, O) set and is one of the most popular functionals in chemistry with a “somehow magic” hybrid ratio. Although in many cases such hybrid functionals with a small fraction of the exact exchange are known to give reasonable reaction energies and relative energies of different spin states,^{68–71} there are many exceptions as well.⁷² Therefore, if one tries to fit DFTB parameters to reproduce a different functional, one would result in a different parameter set.

The problems in predicting DFT-like energetics is partly stemming from the current fitting scheme for the diatomic pair repulsive curve and needs further improvement. Another problem of the present scheme of parameter determination is that one has to carefully work on each pair of elements, which is extremely time-consuming; with a few element pairs a year, it will be long before one can cover all the important element pairs. A more systematic method of determining parameters for a set of pairs of elements at the same time will be required. Efforts along these lines are in progress.

Acknowledgment. We would like to acknowledge Dr. David Quinero for his participation in the early stage of this work. This work was supported in part by grants from the U.S. National Science Foundation (CHE-0209660), U.S. Department of Energy (DE-FG02-03ER15461), CREST (Core Research for Evolutional Science and Technology) in the Area of High Performance Computing for Multi-scale and Multi-physics Phenomena from the Japan Science and Technology Agency (J.S.T.), Deutsche Forschungsgemeinschaft (D.F.G.), and University of Paderborn. Computer resources were provided by the Cherry Emerson Center for Scientific Computation.

Supporting Information Available: Table S1–S5 and DFTB optimized Cartesian coordinates (in Å) (Sc, Ti, Fe, Ni, and Co compounds, respectively). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.
- (2) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133.
- (3) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.
- (4) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.
- (5) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4907.
- (6) Nanda, D. N.; Jug, K. *Theor. Chim. Acta* **1980**, *57*, 95.
- (7) Jug, K.; Iert, R.; Schulz, J. *Int. J. Quantum Chem.* **1987**, *32*, 265.
- (8) Dewar, M. J. S.; Zoebisch, E.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

- (9) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.
- (10) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221.
- (11) Dewar, M. J. S.; Jie, C.; Yu, J. *Tetrahedron* **1993**, *49*, 5003.
- (12) Holder, A. J.; Dennington, R. D.; Jie, C. *Tetrahedron* **1994**, *50*, 627.
- (13) Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1992**, *81*, 391.
- (14) Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1996**, *93*, 315.
- (15) SPARTAN. 4.0 ed.; Wavefunction Inc.: Irvine, CA, 1995.
- (16) Voityuk, A. A.; Zerner, M. C.; Rosch, N. *J. Phys. Chem. A* **1999**, *103*, 4553.
- (17) Repasky, M. P. C., J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601.
- (18) Tubert-Brohman, v.; Guimaraes, C. R. W.; Repasky, M. P.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *22*, 138.
- (19) Tubert-Brohman, I.; Guimaraes, C. R. W.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2005**, *1*, 817.
- (20) Sattelmeyer, K. W.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. A* **2006**, *110*, 13551.
- (21) Sattelmeyer, K. W.; Tubert-Brohman, I.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2006**, *2*, 413.
- (22) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, *27*, 1101.
- (23) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260.
- (24) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, *51*, 12947.
- (25) Zheng, G.; Irle, S.; Morokuma, K. *Chem. Phys. Lett.* **2005**, *412*, 210.
- (26) Elstner, M.; Cui, Q.; Muni, P.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Comput. Chem.* **2003**, *24*, 565.
- (27) Koskinen, P.; Häkkinen, H.; Seifert, G.; Sanna, S.; Frauenheim, T.; Moseler, M. *New J. Phys.* **2006**, *8*, 9.
- (28) Zheng, G.; Irle, S.; Elstner, M.; Morokuma, K. *J. Phys. Chem. A* **2004**, *108*, 3128.
- (29) Zheng, G.; Irle, S.; Morokuma, K. *J. Chem. Phys.* **2004**, *122*, 014708/1.
- (30) Krüger, T.; Elstner, M.; Schiffels, P.; Frauenheim, T. *J. Chem. Phys.* **2005**, *122*, 114110.
- (31) Zhou, H.; Tajkhorshid, E.; Frauenheim, T.; Suhai, S.; Elstner, M. *Chem. Phys.* **2002**, *277*, 91.
- (32) Witek, H. A.; Irle, S.; Zheng, G.; Jong, W. A. d.; Morokuma, K. *J. Chem. Phys.* **2006**, *125*, 214706/1.
- (33) Witek, H. A.; Morokuma, K.; Stradomska, A. *J. Theory Comput. Chem.* **2005**, *4*, 639.
- (34) Witek, H. A.; Morokuma, K. *J. Comput. Chem.* **2004**, *25*, 1858.
- (35) Witek, H. A.; Irle, S.; Morokuma, K. *J. Chem. Phys.* **2004**, *121*, 5163.
- (36) Witek, H. A.; Morokuma, K.; Stradomska, A. *J. Chem. Phys.* **2004**, *121*, 5171.
- (37) Malolepsza, E.; Witek, H. A.; Morokuma, K. *Chem. Phys. Lett.* **2005**, *412*, 237.
- (38) Köhler, C.; Seifert, G.; Gerstmann, U.; Elstner, M.; Overhof, H.; Frauenheim, T. *Phys. Chem. Chem. Phys.* **2001**, *3*, 5109.
- (39) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (40) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (41) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (42) Sarkar, P.; Springborg, M.; Seifert, G. *Chem. Phys. Lett.* **2005**, *405*, 103.
- (43) Joswig, J.; Springborg, M.; Seifert, G. *J. Phys. Chem. B* **2000**, *104*, 2617.
- (44) Seifert, G.; Terrones, H.; Terrones, M.; Jungnickel, G.; Frauenheim, T. *Solid State Commun.* **2000**, *114*, 245.
- (45) Seifert, G.; Terrones, H.; Terrones, M.; Frauenheim, T. *Solid State Commun.* **2000**, *115*, 635.
- (46) Seifert, G.; Tamuliene, J.; Gemming, S. *Comput. Mater. Sci.* **2006**, *35*, 316.
- (47) Seifert, G.; Terrones, H.; Terrones, M.; Jungnickel, G.; Frauenheim, T. *Phys. Rev. Lett.* **2000**, *85*, 146.
- (48) Ivanovskaya, V. V.; Heine, T.; Gemming, S.; Seifert, G. *Phys. Status Solidi B* **2006**, *243*, 1757.
- (49) Köhler, C.; Seifert, G.; Frauenheim, T. *Comput. Mater. Sci.* **2006**, *35*, 297.
- (50) Bertram, N.; Cordes, J.; Kim, Y. D.; Ganteför, G.; Gemming, S.; Seifert, G. *Chem. Phys. Lett.* **2006**, *418*, 36.
- (51) Tamuliene, J.; Seifert, G. *Full. Nanot. Carbon Nanostruct.* **2005**, *13*, 279.
- (52) Ivanovskaya, V. V.; Seifert, G.; Ivanovskii, A. L. *Semiconductors* **2005**, *39*, 1058.
- (53) Enyashin, A.; Seifert, G. *Phys. Status Solidi B* **2005**, *242*, 1361.
- (54) Köhler, C.; Seifert, G.; Frauenheim, T. *Chem. Phys.* **2005**, *309*, 23.
- (55) Ivanovskaya, V.; Seifert, G. *Solid State Commun.* **2004**, *130*, 175.
- (56) Krause, M.; Kuzmany, H.; Georgi, P.; Dunsch, L.; Vietze, K.; Seifert, G. *J. Chem. Phys.* **2001**, *115*, 6596.
- (57) Münch, W.; Kreuera, K.-D.; Seifert, G.; Maiera, J. *Solid State Ionics* **2000**, *136*, 183.
- (58) Yang, M.; Jackson, K. A.; Köhler, C.; Frauenheim, T.; Jellinek, J. *J. Chem. Phys.* **2006**, *124*, 024308/1.
- (59) Frauenheim, T.; Seifert, G.; Elstner, M.; Hajnal, Z.; Jungnickel, G.; Porezag, D.; Suhai, S.; Scholz, R. *Phys. Status Solidi B* **2000**, *217*, 41.
- (60) Köhler, C. Berücksichtigung von Spinpolarisationseffekten in einem dichtefunktionalbasierten Ansatz, Ph.D., University of Paderborn, 2004.
- (61) Elstner, M.; Cui, Q.; Muni, P.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Comput. Chem.* **2002**, *24*, 565.
- (62) Eschrig, H. *The Optimized LCAO Method and Electronic Structure of Extended Systems*; Akademieverlag: Berlin, 1988.
- (63) Foulkes, W.; Haydock, R. *Phys. Rev. B* **1989**, *39*, 12520.
- (64) Dolg, M.; Wedig, U.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1987**, *86*, 866.
- (65) Wedig, U.; Dolg, M.; Stoll, H.; Preuss, H. In *Quantum Chemistry: The Challenge of Transition Metals and Coordination Chemistry*; Veillard, A., Ed.; Reidel: Dordrecht, 1986; p 79.

- (66) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; T. V.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian03 Rev. D01+*; 2004.
- (67) Musaev, D. G.; Froese, R. D. J.; Svensson, M.; Morokuma, K. *J. Am. Chem. Soc.* **1997**, *119*, 367.
- (68) Quiñonero, D.; Musaev, G. D. G.; Morokuma, M. *Inorg. Chem.* **2003**, *42*, 8449.
- (69) Blomberg, M. R. A.; Siegbahn, P. E. M.; Svensson, M. *J. Chem. Phys.* **1996**, *104*, 9546.
- (70) Lundberg, M.; Siegbahn, P. E. M. *J. Comput. Chem.* **2005**, *26*, 661.
- (71) Ricca, A.; Bauschlicher, C. W. *J. Phys. Chem. A* **1997**, *101*, 8949.
- (72) Jensen, K. P.; Roos, B. O.; Ryde, U. *J. Chem. Phys.* **2007**, *126*, 014103/1.
- CT600312F

JCTC Journal of Chemical Theory and Computation

Parallel Calculation of Coupled Cluster Singles and Doubles Wave Functions Using Array Files

Tomasz Janowski, Alan R. Ford, and Peter Pulay*

*Department of Chemistry and Biochemistry, Fulbright College of Arts and Sciences,
University of Arkansas, Fayetteville, Arkansas 72701*

Received February 27, 2007

Abstract: A new parallel implementation of the Coupled Cluster Singles and Doubles (CCSD) and related wave functions (e.g. Quadratic Configuration Interaction, QCI, and Coupled Electron Pair, CEPA) is described, based on the Array Files middleware. The program can handle large basis sets, even without utilizing symmetry, on modest distributed memory workstation clusters. High computational efficiency is achieved by formulating all major operations in terms of matrix multiplications. Timings are provided for systems with 50–228 valence electrons and up to 1144 basis functions, with little or no symmetry. Our largest calculation (QCISD/aug-cc-pVQZ for the parallel displaced benzene dimer) uses 1512 basis functions. Calculations on the benzene dimer show that the usual procedure of estimating the effect of basis set enlargement from second-order Møller–Plesset (MP2) calculations is less reliable than previously assumed. Replacing the weak pair amplitudes in CCSD/QCISD calculations by MP2 amplitudes affects the calculated energy only slightly.

I. Introduction

Coupled Cluster¹ (CC) and related techniques are the most accurate routinely applicable electronic structure methods (for reviews see refs 2–4). Tests performed since the first implementations of Coupled Cluster theory with single and double substitutions (CCSD) in 1978^{5–7} have shown that it still has systematic errors (for an excellent treatment of the accuracy of various accurate quantum chemical methods see ref 8). However, adding triple substitutions yields essentially quantitative results for systems with only dynamical electron correlation if large basis sets are employed.⁸ Because of the high cost of triples, they are included in most cases perturbatively. Among the several ways of doing this, the CCSD(T)^{9,10} method became the most popular. The consistently high accuracy of coupled cluster based methods is important in several areas, for instance in accurate thermochemistry or for benchmarking more approximate methods.

The main disadvantage of CC methods is their high computational cost, both in terms of CPU time and random access and disk memory. This makes larger calculation impossible or impractical on the current generation of

computers. Local correlation methods^{11–15} eliminate the steep scaling of traditional CC methods for molecules with well localized electronic structure and allow calculations on large systems. However, local methods perform less well for genuinely delocalized systems or for basis sets augmented with diffuse functions. The latter are essential, e.g., for the description of dispersion forces. One obvious way to extend the applicability of full CC methods is to use parallel computing. Clusters of inexpensive personal computer based workstations, introduced independently by many research groups (including ours) 10 years ago, have excellent price/performance ratios and are available with parallel quantum chemistry programs preinstalled.¹⁶ By using the combined resources of the system, it is possible to overcome limitations in computer time, random access memory, and disk memory. Parallelization of the perturbative triples part of CCSD(T) can be accomplished with relatively little communication between nodes, and thus it is less demanding than CCSD.^{17,18} The latter requires repeated access to large (four index) arrays and is not easy to implement efficiently in parallel. The aim of the present project is a parallel CCSD program that can handle moderately large systems on modest-size (and therefore widely available) clusters with good quality basis

* Corresponding author e-mail: pulay@uark.edu.

sets. This program serves as a starting point for a CCSD(T) program. In our opinion, the ability to handle a system with 10–20 non-hydrogen and 10–20 hydrogen atoms, with basis sets ranging from cc-pVTZ for the largest molecules to aug-cc-pVQZ¹⁹ for the smaller ones, would have the greatest impact at the present stage of computational technology. Such calculations may include up to 2000 basis functions and up to 100 valence electrons (not necessarily simultaneously).

Because of its high computational demand, CCSD appears to be an ideal candidate for parallel implementation. The first distributed-memory parallel CCSD algorithm was described by Rendell et al.²⁰ for the now defunct Intel Hypercube. However, when our first implementation was finished in the spring of 2005, there were only two working parallel implementations of the full CCSD model: in MOLPRO²¹ and in NWChem.²² Neither of these is described in detail in the literature. The general strategy of our program is most similar to MOLPRO which is not surprising, as both MOLPRO and PQS,²³ the vehicle used to develop the present program, trace their roots back to an early ab initio program developed initially (from 1968) by Prof. W. Meyer (Kaiserlautern, Germany) and one of us (P.P.). The parallel CCSD(T) program in NWChem is based on the algorithm of Kobayashi and Rendell.²⁴ In the meantime, parallel CCSD(T) has been implemented in GAMESS-US²⁵ and is under development²⁶ in the Mainz-Austin-Budapest version of the ACES II package.²⁷ Both the NWChem and the GAMESS-US parallel implementations use the massively parallel paradigm (hundreds of processors). Our goal was to be able to perform large calculations without using massive parallelism. Access to such a high number of processors, although improving, is still severely limited.

Both existing parallel CCSD codes, NWChem and MOLPRO, use Global Arrays^{28,29} (GA) as a parallelization tool. GA simulates shared memory programming on distributed memory computer clusters by striping large arrays across nodes. Originally, it envisioned storage in fast memory only. However, in view of the massive amounts of data needed in large CC calculations, such a strategy is prohibitive on moderately sized computer clusters. Later versions of GA allow accessing data on secondary (disk) storage, but they have to be accessed through the GA subsystems. In view of the large primary memories available on the present generation of small workstations, striping individual matrices (with typical sizes of 1–4 MWords, i.e., 8–32 Mbytes) is not necessary any more and may even be counterproductive. Therefore, we have chosen Array Files³⁰ (AF), recently developed in this laboratory, as a parallelization tool. Array Files fits the computational requirements of our matrix-formulated CCSD particularly well.

This paper describes a parallel CCSD/QCISD program for closed shell molecules that can perform large calculations on modest workstation clusters. In addition to CC methods, the program can also calculate various approximations to CCSD, such as versions of the Coupled Electron Pair Approximation³¹ (we have implemented CEPA-0 and CEPA-2), Quadratic Configuration Interaction (QCISD),³² and lower orders of Møller–Plesset perturbation theory (MP2, MP3, and MP4(SDQ)) as well as variational Configuration Interac-

tion with single and double substitutions (CISD). These methods can be viewed as approximations to the full CCSD method. Most calculations were carried out using the QCISD method that is comparable in quality with CCSD but is less expensive.

II. Theory

In this section, we recapitulate the theory and establish the notation. The CCSD energy and wave function are defined by the vanishing of the residuals in eqs 1–3:

$$\langle \Phi^0 | \mathbf{H} - E | \Psi^0 + \exp(\sum_{ia} T_a^i \mathbf{E}_i^a + \sum_{ijab} T_{ij}^{ab} \mathbf{E}_{ij}^{ab}) \Phi^0 \rangle = 0 \quad (1)$$

$$\mathbf{R}_c^k = \langle \tilde{\Psi}_k^c | \mathbf{H} - E | \Psi^0 + \exp(\sum_{ia} T_a^i \mathbf{E}_i^a + \sum_{ijab} T_{ij}^{ab} \mathbf{E}_{ij}^{ab}) \Phi^0 \rangle = 0 \quad (2)$$

$$\mathbf{R}_{cd}^{kl} = \langle \tilde{\Psi}_{kl}^{cd} | \mathbf{H} - E | \Psi^0 + \exp(\sum_{ia} T_a^i \mathbf{E}_i^a + \sum_{ijab} T_{ij}^{ab} \mathbf{E}_{ij}^{ab}) \Phi^0 \rangle = 0 \quad (3)$$

Here Φ^0 is a normalized reference wave function, in our case a closed-shell determinant, and the substitution operators \mathbf{E} acting on Φ^0 transfer one or two electrons from the occupied space (indices i, j, k, l) to the virtual space (a, b, c, d). The bra projectors $\tilde{\Psi}_k^c$, $\tilde{\Psi}_{kl}^{cd}$ span the singly and doubly substituted space of $\mathbf{E}_i^a \Phi^0$ and $\mathbf{E}_{ij}^{ab} \Phi^0$. The tilda notation in eqs 2 and 3 emphasizes that the individual substituted functions Ψ may be different on the right- and left-hand sides; only the spaces spanned by them are identical. In other words, we can use different linear combinations of the substituted (sometimes called by the pedagogically unfortunate name “excited”) wave functions on the two sides. This becomes important in the spin adapted version of the theory. Spin adaptation confers significant computational advantages for closed-shell reference wave functions.

It has been shown that a biorthogonal version of the spin-adapted closed shell coupled cluster doubles (CCD) theory³³ halves the computational effort for the pair coupling terms, compared to formulations employing fully orthogonal spin-adapted functions. We have called this the generator state formulation since the right-hand basis functions are the generator states advocated by Matsen.³⁴ It is worth noting that Čížek’s original formulation¹ used implicitly generator states. Subsequent work concentrated mostly on orthogonal functions. In the generator state form, the singly and doubly substituted configurations are defined as

$$\Psi_i^a = \mathbf{E}_i^a \Phi^0 = (\mathbf{e}_i^a + \bar{\mathbf{e}}_i^a) \Phi^0 = \Phi_i^a + \Phi^{\bar{a}} \quad (4)$$

$$\Psi_{ij}^{ab} = \mathbf{E}_{ij}^{ab} \Phi^0 = \mathbf{E}_i^a \mathbf{E}_j^b \Phi^0 = \Phi_{ij}^{ab} + \Phi^{\bar{j}ab} + \Phi_{ij}^{\bar{a}b} + \Phi_{ij}^{\bar{a}\bar{b}}; \quad i \geq j \quad (5)$$

The spin-summed single substitution operators are the sums of spin-orbit substitutions:

$$\mathbf{E}_i^a = \mathbf{e}_i^a + \bar{\mathbf{e}}_i^a \quad (6)$$

In eqs 4–6, \mathbf{e} is a spin-orbital substitution operator, replacing an occupied spin orbital i or j by virtual orbitals a or b ; indices without overbars refer to α spin and with overbars

β spin. The functions Φ are substituted Slater determinants where the subscript spin-orbitals have been replaced by the superscript spin-orbitals. Interchanging a and b in eq 5 generates a linearly independent doubly substituted function (unless $i=j$ or $a=b$) that is not orthogonal to Ψ_{ij}^{ab} . This pairwise nonorthogonality causes no computational problems if the left-hand (contravariant) projection functions are biorthogonal to the substituted functions on the right-hand side. This is achieved by defining the projection functions as the biorthogonal partners of the expansion functions³³

$$\bar{\Psi}_i^a = \frac{1}{2} \Psi_i^a \quad (7)$$

$$\bar{\Psi}_{ij}^{ab} = \frac{1}{6} (2\Psi_{ij}^{ab} - \Psi_{ij}^{ba}) \quad (8)$$

The present program is based on the elegant matrix formulation of the singles and doubles correlation problem introduced by Meyer³⁵ under the acronym SCEP (Self-Consistent Electron Pair) theory coupled with the generator state spin adaptation. This yields very compact formulas for the CC equations. For reference in the next section, we give the CCD equations³³ explicitly below. Although CCD is seldom used by itself today, it constitutes the computationally most significant part of the CCSD equations.

As usual in SCEP theory, integrals with two internal (occupied) indices are collected in internal Coulomb and exchange matrices:

$$\mathbf{J}_{ab}^{ij} = (ij|ab) \mathbf{K}_{ab}^{ij} = (ai|jb) \quad (9)$$

The CCD amplitudes are calculated by iteratively refining the doubles amplitudes until the doubles residuals vanish:

$$\mathbf{R}^{ij} = \mathbf{K}^{ij} + \mathbf{K}[\mathbf{T}^{ij}] + \mathbf{Q}^{ij} + (\mathbf{Q}^{ij})^\dagger + \mathbf{G}^{ij} + (\mathbf{G}^{ij})^\dagger = \mathbf{0} \quad (10)$$

Here

$$\mathbf{Q}^{ij} = \mathbf{T}^{ij}(\mathbf{F}-\mathbf{A}) + \frac{1}{2} \sum_{\mathbf{k}} [(2\mathbf{T}^{ik} - \mathbf{T}^{ki})\mathbf{Y}^{kj} - \mathbf{T}^{ki}\mathbf{Z}^{kj} - 2(\mathbf{T}^{ki}\mathbf{Z}^{kj})^\dagger - \beta_{ki}\mathbf{T}^{kj}] \quad (11)$$

$$\mathbf{A} = \sum_{\mathbf{kl}} (2\mathbf{K}^{\mathbf{kl}} - \mathbf{K}^{\mathbf{lk}})\mathbf{T}^{\mathbf{kl}} \quad (12)$$

$$\mathbf{Y}^{\mathbf{kj}} = 2\mathbf{K}^{\mathbf{kj}} - \mathbf{J}^{\mathbf{kj}} + \sum_{\mathbf{l}} (2\mathbf{K}^{\mathbf{kl}} - \mathbf{K}^{\mathbf{lk}})(2\mathbf{T}^{\mathbf{lj}} - \mathbf{T}^{\mathbf{jl}}) \quad (13)$$

$$\mathbf{Z}^{\mathbf{kj}} = \mathbf{J}^{\mathbf{kj}} - \sum_{\mathbf{l}} \mathbf{K}^{\mathbf{lk}}\mathbf{T}^{\mathbf{jl}} \quad (14)$$

$$\beta_{ki} = \mathbf{F}_{ki} + \sum_{\mathbf{l}} \text{Tr}[(2\mathbf{K}^{\mathbf{kl}} - \mathbf{K}^{\mathbf{lk}})\mathbf{T}^{\mathbf{li}}] \quad (15)$$

$$\mathbf{G}^{ij} = \sum_{\mathbf{kl}} \alpha_{ij,kl} \mathbf{T}^{\mathbf{kl}} \quad (16)$$

$$\alpha_{ij,kl} = (ik|jl) + \text{Tr}[\mathbf{T}^{ij}\mathbf{K}^{\mathbf{lk}}] \quad (17)$$

In the above formulas, the bold-faced quantities are matrices in the external (virtual) space, \mathbf{F} is the Fock matrix, and the external exchange matrix is defined as

$$\mathbf{K}[\mathbf{T}^{ij}]_{ab} = \sum_{\mathbf{cd}} (ac|bd)(\mathbf{T}^{ij})_{\mathbf{cd}} \quad (18)$$

The quantities $(ik|jl)$ and $(ac|bd)$ are two-electron integrals in the Mulliken notation. The exact notation in eqs 10–17

is that of Hampel et al.³⁶ who have generalized the CCD equations of ref 33 to the CCSD case. Prior to this, Scuseria et al.³⁷ developed a formulation of the CCSD theory that achieves the same computational savings but does not use a matrix form. We consider the matrix/tensor formulation preferable, not only because of its simplicity but also because modern computers are very efficient for matrix manipulations, particularly matrix multiplications.

There are several plausible orbital choices for both the internal (occupied) and the external (virtual) space. The occupied orbitals can be either canonical or localized. The former usually converge slightly faster but are less efficient for the utilization of sparsity; symmetry is also simpler to implement with localized orbitals. The virtual space in the above formulation, just like SCEP,³⁵ can be easily generalized to nonorthogonal atomic basis functions (AOs) or AOs projected against the internal space^{11,12} instead of virtual molecular orbitals (MOs). This makes it particularly suitable for AO-based local correlation theories.^{11,12,38} Our program can use either AOs or canonical MOs in the virtual space; the transformation between the MO and AO representations¹¹ is straightforward. Sparsity can be better exploited in the AO form, but the dimension of the matrices (the number of AOs) is higher than in the MO representation (the number of virtual orbitals). Note that all representations yield strictly identical results (within the limits of numerical precision) if no further approximations are made.

Even if AOs are used for the virtual space, updating the amplitudes is done in an orthogonal MO basis, according to first-order perturbation theory as

$$\Delta \mathbf{T}_{ab}^{ij} = -\mathbf{R}_{ab}^{ij}/(\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j + \delta) \quad (19)$$

where a and b are canonical virtual orbitals, ϵ_a and ϵ_b are their orbital energies, and i and j are either canonical occupied or localized orbitals. In the first case, ϵ_i and ϵ_j are orbital energies; in the second case, they are the Coulson energies of the localized orbitals, e.g., $\epsilon_i = \langle \varphi_i | \mathbf{F} | \varphi_i \rangle$. The quantity δ is a level shift and changes the convergence rate but has no effect on the converged results.

The same program is used to calculate all wave functions available in the program (CEPA-2 and CEPA-0, MP3, MP4-(SDQ), CCD, QCISD, QCID, CISD, and CID), except MP2, since all these many-body methods are computationally simplified versions of CCSD.

III. Algorithm

The parallel CCSD program has been implemented in the academic version of the PQS²³ program package. In its basic architecture, it is similar to the original Self-Consistent Electron Pair program,³⁹ our earlier Local Electron Correlation program,^{11,12} and the implementation in MOLPRO,^{13,40} with special attention paid to parallel performance.

The first step of the algorithm is the calculation and storage of the internal Coulomb and exchange operators, eq 9, and the determination of the MP2 energy and amplitudes. The latter serves as the first approximation to the CC amplitudes and can be substituted for the pair amplitudes of weak pairs¹² in a localized calculation with negligible loss of accuracy

Table 1. Megaflop Ratings for Dense Matrix Multiplication^a

matrix dimensions	method	Mflops/s
1000 × 1000	DGEMM	5000
1000 × 1000	DDOT	487
1000 × 1000	DAXPY	307
1000 × 50	DGEMM	3570
1000 × 25	DGEMM	2600
1000 × 15	DGEMM	1970

^a On a 3.2 GHz Intel Nocona processor, using the Goto BLAS library.⁴⁵

and considerable gain in efficiency (see the Results section). The **J** and **K** matrices are distributed on the aggregate disk storage of the nodes in the cluster. We use the recently developed Array Files³⁰ (AF) for all distributed disk storage. AF allows transparent access to disk records distributed across nodes in a computer cluster. Calculation of the **J** and **K** matrices and the MP2 amplitudes constitutes only a small fraction of the total computational time and follows our efficient canonical MP2⁴¹ and parallel MP2⁴² algorithms. Note, however, that the MP2 algorithms become iterative if localized orbitals are used, as shown in the first full formulation of MP2 with noncanonical orbitals;⁴³ the quadruples contributions in MP4 likewise become iterative.¹¹ An alternative to iterative noncanonical MP2 is the Laplace transform MP2 of Almlöf.⁴⁴

Our main algorithmic goal was, besides minimizing disk access and internode communication, to formulate all computationally significant operations as matrix multiplications. Note that this differs significantly from the “vectorization” strategy of the 1980s and early 1990s. On a typical vector computer, all typical vector operations run at approximately the same speed. On modern CPUs, arithmetic operations are so fast that the main computational bottleneck is fetching data from memory. Matrix multiplication, if implemented in efficient blocked form, allows the reuse of data in cache memory, leading to nearly theoretical efficiency (on many modern microprocessors, the floating point operation rate is twice the clock rate). Table 1 shows that the dot product form of matrix multiplication (DDOT) is 10.5 times slower than a state-of-the-art level 3 BLAS routine⁴⁵ DGEMM for a dense 1000 × 1000 matrix, and the DAXPY (“outer product”) form is 16.7 times slower. The megaflop rating for multiplying two 1000 × 50 matrices is over 70% of the limiting rate, and even 1000 × 15 matrices give nearly 2 Gflops/s (40%) performance.

For large basis sets, the computational effort is usually dominated by the external exchange, the scaling of which is $O(n^2N^4)$ vs $O(n^3N^3)$ for the other sixth-order terms. Here n is the number of correlated occupied orbitals, and N is either the number of AOs (in the AO formulation) or the number of virtual orbitals. The external exchange term is evaluated in an integral-direct manner by transforming the CC amplitudes to AO basis, evaluating eq 18 in AO basis according to

$$\mathbf{K}[\mathbf{T}^{ij}]_{\mu\lambda} = \sum_{\nu\sigma} (\mu\nu|\lambda\sigma) \mathbf{T}_{\nu\sigma} \quad (20)$$

In eq 20, μ , ν , λ , and σ denote AOs and i, j correlated occupied orbitals. The resulting **K** matrices are transformed

to the virtual basis used in the program. Our program can use either canonical virtual orbitals or projected atomic orbitals. This strategy, based on Meyer’s SCEP,³⁵ avoids the storage of integrals with three and four virtual indices and has been adopted by several programs: the Saebo-Pulay local correlation program,^{11,12} MOLPRO,⁴⁰ and the program of Kobayashi and Rendell.²⁴ Without the integral-direct calculation of the external exchange, the storage of integrals with three and four virtual indices becomes a bottleneck, particularly on a single node. E.g. the QCISD calculation with the aug-cc-pVQZ calculation for the benzene dimer, described later, would require about 5.9 Tbytes (5900 Gbytes) for the storage of the all-external ($ab|cd$) integrals alone if symmetry is disregarded. However, in view of rapidly increasing disk capacities and the possibility of distributed storage, storing all transformed integrals on distributed disk memory may be a viable option in the near future.

The computational cost of the external exchange can be reduced by a factor of 2 if symmetric and antisymmetric combinations of the AO integrals are used, according to

$$\Sigma \mathbf{K}^{\pm}[\mathbf{T}^{ij}]_{\mu\lambda} = \sum_{\nu\sigma} (1 + \delta_{\nu\sigma})^{-1} [(\mu\nu|\lambda\sigma) \pm (\mu\sigma|\lambda\nu)] [(\mathbf{T}^{ij})_{\nu\sigma} \pm (\mathbf{T}^{ij})_{\sigma\nu}] \quad \mu \geq \lambda, \nu \geq \sigma \quad (21)$$

$$\mathbf{K}[\mathbf{T}^{ij}]_{\mu\lambda} = 1/2(\mathbf{K}^{+}[\mathbf{T}^{ij}]_{\mu\lambda} + \mathbf{K}^{-}[\mathbf{T}^{ij}]_{\mu\lambda});$$

$$\mathbf{K}[\mathbf{T}^{ij}]_{\lambda\mu} = 1/2(\mathbf{K}^{+}[\mathbf{T}^{ij}]_{\mu\lambda} - \mathbf{K}^{-}[\mathbf{T}^{ij}]_{\mu\lambda}) \quad (22)$$

This algorithm was first explicitly described in ref 11, but it appears that it had been used, at least for symmetrical molecules, in the original SCEP program.³⁵ It has been adopted by Scuseria et al.³⁷ and Kobayashi and Rendell.²⁴ A disadvantage, which it shares with our MP2,⁴¹ is that the number of AO integrals evaluated is approximately four times larger than the minimum necessary if all integral permutation symmetry is utilized. The evaluation of eq 21 requires formally $n^2N^4/2$ floating-point operations, while integral evaluation is CN^4 where C is relatively large and independent of the number of correlated internal orbitals n . Therefore this algorithm is most advantageous for large systems ($n^2 \gg C$).

The matrix formulation used here leads automatically to a highly efficient program for almost all terms. The exception is the external exchange operator. If performed for a single (ij) pair, eq 20 is a scalar product that performs poorly on modern CPUs. To increase its performance, we try to construct **K** matrices simultaneously for as many (ij) pairs as local fast memory permits. Note that it is important to use the size of the actual fast memory and not the virtual memory here. Treating (ij), ($\mu\lambda$), and ($\nu\sigma$) as single indices, eq 20 becomes a matrix operation, although the range of (ij) and ($\mu\lambda$) is much smaller than that of ($\nu\sigma$). A pseudocode of the essential parts of the algorithm is shown in Figure 1. Note that, depending on the size of the shells and the available memory, a large number of ($\mu\lambda$) shells may be processed together. This improves both the CPU timing because the matrix dimensions become larger but has an even bigger effect by reducing the I/O needed to fetch the amplitudes from the distributed disk storage. Although


```

do M=1,NSH      (shells of AOs)
  do Λ=1,M      (shells of AOs)
    Calculate all AO integrals  $(\mu\nu|\lambda\sigma)$ ,  $\mu\in M$ ,  $\lambda\in\Lambda$ 
    Form the matrices  $X_{\mu\lambda,\nu\sigma}^{\pm}=(\mu\nu|\lambda\sigma)\pm(\mu\sigma|\lambda\nu)$   $\nu\geq\sigma$ 
    Accumulate the matrices X in memory until the memory
      set aside for integrals is full
    if the integral memory is full
      do for batches of (ij)
        Read a batch of  $\mathbf{T}^{ij}$  amplitudes and transform
          them to AO basis; put the transformed
          amplitudes in the matrix  $T_{\nu\sigma,ij}$ 
        Calculate  $K_{\mu\lambda,ij}^{\pm}=\sum_{\nu\sigma}X_{\mu\lambda,\nu\sigma}^{\pm}\times T_{\nu\sigma,ij}$  [Eq. (21)]
        Calculate  $K_{\mu\lambda,ij}$  [Eq. (22)]
      end do batches of (ij)
      release integral memory
    end if
  end do Λ
end do M

```

Figure 1. Pseudocode of the external exchange matrix construction, eqs 21 and 22.

MOLPRO uses a different algorithm for building the external exchange operators, it incorporates a similar “shell merging” feature.⁴⁰

Because of the high cost of the external exchange operator, it is important to utilize the natural sparsity of the integral list. However, this is possible only to a limited extent. According to our experience one has to maintain a sharp integral threshold. Aggressive neglect of the integrals in eqs 20–22 can cause the CCSD iteration to diverge, particularly if the basis contains diffuse functions. Integral sparsity is used in the following manner. Integrals $(\mu\nu|\lambda\sigma)$, for all $\nu\sigma$ and as many $\mu\lambda$ as the memory allows, are collected in the fast memory as a matrix with composite row index $\mu\lambda$ and column index $\nu\sigma$. This matrix is divided into horizontal stripes, usually so that all $\mu\lambda$ pairs that come from a common pair of shells constitute a stripe. Each stripe is inspected for columns having all integrals below a threshold and is independently compressed by removing these columns. Integrals with basis functions from the same shells share the sparsity pattern, and therefore most negligible integrals are removed at this stage. An indexing array keeps track of the numbering of the original columns. When the RAM memory is full, the multiplication with the amplitudes is performed, separately for each stripe. To enable the use highly efficient dense matrix multiplication routines, the amplitude matrices have to be also compressed by removing the rows corresponding to columns removed from a stripe. For higher angular momentum functions the stripes are sufficiently wide to guarantee high performance in this step, cf. Table 1. The disadvantage of his algorithm is that the same amplitude matrix has to be compressed separately for each integral stripe. This is an argument for having as big integral stripes as possible, but in this case sparsity deteriorates. Conversely, sparsity is best if each stripe covers only one shell pair. However, this leads to small dimensions for low angular

momentum functions and the corresponding loss of efficiency in the matrix multiplication. In the extreme case of two s type shells, the matrix multiplication becomes a dot product which is ~ 10 times less efficient. The best compromise appears to be to treat larger shell pairs (pd , dd , df , ff) as separate stripes but merge smaller shell pairs to have a minimum dimension of ~ 15 .

The screening algorithm speeds up the calculation by lowering the amount of integrals calculated, decreasing the flop count during matrix multiplication, and it also saves memory because the zero columns are not kept in memory, so more integrals can be stored before performing the multiplication with amplitudes. This allows reducing I/O considerably: the more integrals can be stored in memory the fewer disks reads of amplitudes is needed. For very large calculations (more than 1500 basis functions on a computer with modest memory), disk I/O becomes the dominant part of the EEO’s calculation.

In the present implementation of the CCSD equations the \mathbf{Q} , \mathbf{Y} , and \mathbf{Z} matrices are precalculated and stored on disk before they are used in the residuum construction loop. In order to minimize disk access, the calculation is performed by blocks using a method similar to that employed to speed up matrix multiplication by taking advantage of fast cache memory. Here the RAM memory plays the role of the cache.

This technique is illustrated for the calculation of $\tilde{\mathbf{Q}}^{ij}$, the \mathbf{Y} contributions to \mathbf{Q}^{ij} in eq 11. The $\tilde{\mathbf{Q}}^{ij}$ matrix is considered the ij element of the supermatrix Λ , $2\mathbf{T}^{ij} - \mathbf{T}^{ji}$ is the ij element of a supermatrix Ω , and the matrix \mathbf{Y}^{ij} is an element of the supermatrix Θ . The calculation of $\tilde{\mathbf{Q}}^{ij}$

$$\tilde{\mathbf{Q}}^{ij} = \sum_k (2\mathbf{T}^{ik} - \mathbf{T}^{ki})\mathbf{Y}^{kj} \quad (23)$$

can be written as $\Lambda = \Omega \cdot \Theta$, i.e. as a matrix multiplication where each matrix element is a matrix itself. By dividing Λ

```

Do ii=1, N (N is number of submatrices in a row)
  Do jj=1, M (M is number of submatrices in a column)
    Reserve space for the  $Q^{ij}$  matrices,  $i \in ii, j \in jj$ .
    Initialize  $Q^{ij}$  with zero values.
    Do k=1, n (n is number of correlated orbitals)
      Store all matrices  $T^{ik}$ ,  $i \in ii$  group
      Store all matrices  $K^{kj}$ ,  $j \in jj$  group
      Calculate from them all possible contributions
        to  $Q^{ij}$ ,  $i \in ii, j \in jj$  group.
      Add the calculated contributions to
        current values of  $Q^{ij}$ 
    End do
  Write all  $Q^{ij}$ 
End do
End do
End do

```

Figure 2. Pseudocode of the construction of the Y contribution to the pair coupling terms, eqs 11 and 23.

into smaller square or rectangular submatrices and calculating all elements of a submatrix before proceeding to the next submatrix, the I/O associated with this operation can be substantially reduced because the matrices in fast memory are reused several times. The submatrix sizes are determined by the available memory and the number of nodes working on the calculation, so that each of them gets at least one submatrix to work on. The pseudocode of this algorithm is shown in the Figure 2.

Some terms in eqs 11–17, e.g. $\alpha_{ij,kl}$ in eq 17, are expressed as traces of matrix products, $\text{Tr}(\mathbf{A}\mathbf{B})$; this operation is in effect a dot (scalar) product of two vectors and is not efficient on modern CPUs. The efficiency of this part of the code can be significantly increased by a method similar to the one used for the external exchange operator. Introducing the composite indices ij , kl , and ab for \mathbf{T} and \mathbf{K} transforms eq 17 in a matrix multiplication:

$$\alpha_{ij,kl} = (ik|jl) + \sum_{ab} \mathbf{T}_{ij,ab} \mathbf{K}_{kl,ab} \quad (24)$$

However, the 4-index quantities $\mathbf{T}_{ij,ab}$ and $\mathbf{K}_{kl,ab}$ do not fit into fast memory for larger systems. Therefore the indices ij and kl are subdivided in blocks of appropriate size that allow the storage of these quantities but still give reasonably high efficiency in the matrix multiplications. A similar method is used for the calculation of the \mathbf{G} matrices, eq 16.

The parallelization of the CCSD program with the AF tool was designed as a simple master-slave scheme. All computational tasks in our code are formulated as relatively long loops, typically over pairs of internal (occupied) orbitals or pairs of AO indices. The master assigns the current task, labeled by the loop index, dynamically to the first idle slave. No other programming is needed, because all data can be transparently accessed from each node. Figure 3 shows the algorithm for the distributed computation of the residual matrices, eq 10, parallelized by a pair of occupied indices ij . The Coulomb, exchange, EEO, and three-external integral modules were parallelized similarly, except that the main loop was over μ and ν AO indices instead ij pairs. This is similar to the method used in our parallel MP2 algorithm.³⁰ To minimize I/O, the \mathbf{Q} , \mathbf{Y} , and \mathbf{Z} matrices are calculated in batches of ij 's in both the serial and the parallel program. This makes the parallel code more sensitive to load balancing. We have both a dynamic distribution of these batches

and a static distribution that usually allows better load balancing because the computational task per ij index pair is always the same.

As a message passing software the PVM was used, both for communication with Array Files and master-slave messages. However, both the base PQS code and the AF subsystem also have an MPI version. The major network load comes from nodes to AF traffic (file data reads and writes), the master-slave communication reduces to control messages only.

IV. Results

Table 2 shows representative QCISD timings for some medium-sized molecules with basis sets ranging from small (6-31G*) to large (PC-2⁴⁶ and aug-cc-pVTZ⁴⁷). The number of atoms varies from 20 to 73, and the number of basis functions varies from 282 to 1144. Most calculations were run on a 15-node home-built cluster. The timings demonstrate that CCSD/QCISD calculations can be performed routinely for drug-size molecules with good basis sets and for larger molecules with smaller basis sets. Except for glycine-10, the calculations were performed with an earlier version of the code in which some smaller computational tasks were not yet parallelized. The current code is about 13% faster for 8 nodes and 20% faster for 15 nodes.

Table 3 shows timings for the benzene dimer at the QCISD level using six different basis sets, including very large ones (over 1500 basis functions). The benzene dimer became an important benchmark in testing ab initio techniques for the prediction of dispersion forces, in particular π - π interactions.^{48–52} It is surprisingly difficult to obtain converged results that are correct for the right reason, i.e., without semiempirical adjustment. SCF theory, and most density functional methods, if corrected for basis set superposition error, give a repulsive potential curve. The simplest theoretical level that accounts qualitatively for π - π attraction is MP2. However, MP2, in the basis set limit, overestimates the well depth by as much as a factor of 2 and consequently underestimates the van der Waals distance. CCSD, and the very similar QCISD, overcorrect this defect and lead to an *underestimation* of the well depth (and an overestimation of the distance).

Table 3 demonstrates that, as expected, the external exchange part becomes dominant as the basis set increases

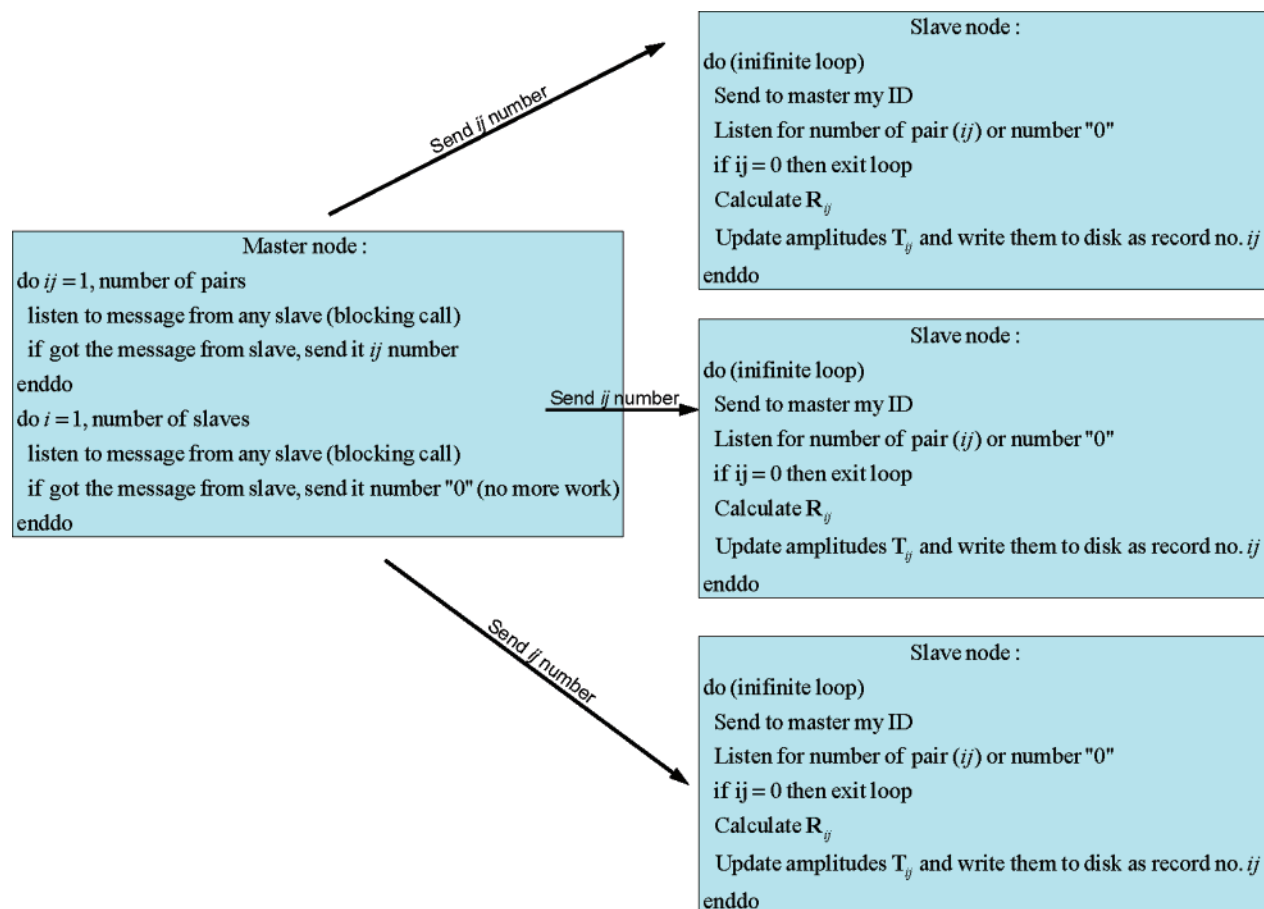


Figure 3. The pseudocode of the parallelization scheme for the main CCSD residuum loop. The same scheme was used for all other quantities calculated.

Table 2. Elapsed Times per QCISD Iteration for Medium-Sized Molecules with a Variety of Basis Sets^a

molecule	empirical formula	basis set	n^b	N^c	nodes	iterations	time/iter (min)
naphthalene+H ₂	C ₁₀ H ₁₀	aug-cc-pVTZ ^d	25	690	14	11	98
aspirin	C ₉ H ₈ O ₄	6-311G**	34	282	6	12	12
sucrose		6-31G*	68	366	11	<i>e</i>	51
sucrose	C ₁₂ H ₂₂ O ₁₁	6-311G**	68	546	11	<i>e</i>	170
yohimbine	C ₂₁ H ₂₆ N ₂ O ₃	PC-2 ^f	69	1144	20	18 ^g	2074
glycine-10 ^h	C ₂₀ H ₃₂ N ₁₀ O ₁₁	6-31G*	114	524	15	12	409

^a On a cluster of 3 GHz dual-core Pentium D processors, except for yohimbine which was run on 20 nodes of the University of Arkansas Red Diamond a 128-node cluster of dual-processor nodes equipped with 3.2 GHz Xeon processors. ^b Number of correlated occupied orbitals. ^c Number of basis functions. ^d Reference 47. ^e These calculations were stopped before full convergence was obtained. ^f Reference 46. ^g The number of iterations is larger than the usual ~12 because the DIIS extrapolation had to be restricted to 4 vectors, instead of the usual 6, due to limited local storage on the Red Diamond cluster. ^h Glycine polypeptide, α helix conformation.

for the same molecule. However, the composition of the basis set also has an effect: diffuse functions reduce the sparsity of the integral list and increase the cost of the external exchange operator. The recalculation of the integrals, although lower scaling in principle, is also quite expensive. The CPU efficiencies in Table 3 are reasonably high, particularly for larger basis sets. The best CPU efficiency is obtained for moderate size systems, i.e., 500–1000 basis functions. For smaller systems the latency associated with message passing becomes significant, as the amount of actual calculation is very small and a parallel synchronization becomes major part of the calculation time. For the systems bigger than 1000 basis functions disk I/O becomes a bottleneck, especially in the external exchange, because multipassing is necessary. The performance of our program

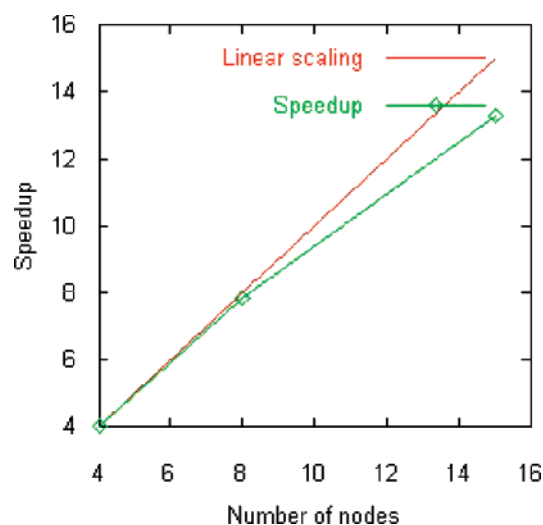
strongly depends on the network throughput and increases dramatically if all nodes are directly connected to the same network switch. Unfortunately, because of the resource limitations we were unable to perform all calculations with the same network configuration.

For big CCSD type calculations memory becomes as important a resource as CPU speed. Since the current algorithm reads the amplitudes many times in order to multiply them by the integrals stored in memory, doubling the amount of memory decreases I/O by the same factor. For large calculations (1500 basis function and more) that are limited by disk I/O, the computational speed almost doubles. Note the timing difference for the aug-cc-pvtz calculation using 32-bit code with 180 MW of memory used and 64-bit code with 380 MW.

Table 3. Timings for Benzene Dimer Using Different Basis Sets on a Cluster of 3.2 GHz Dual-Processor Xeon Machines^f

system/basis	N^a	N^b	memory/slave (MW)	code (bit)	no. of slaves	time/iter (m)	EEO (%) ^c	CPU efficiency ^d (%)
M, cc-pvdz	15	228	180	32	20	1.33	55	45
M, cc-pvtz	15	528	180	32	20	13.5	63	63
M, cc-pvqz	15	1020	180	32	20	160	83	80
M, aug-cc-pvdz	15	384	180	32	20	10.1	79	75
M, aug-cc-pvtz	15	828	180	32	20	117	87	75
M, aug-cc-pvqz	15	1512	400	64	31	858	94	81 ^e
D, cc-pvdz	30	228	180	32	20	2.67	49	40
D, cc-pvtz	30	528	180	32	20	31.7	62	70
D, cc-pvqz	30	1020	180	32	20	682	90	60
D, aug-cc-pvdz	30	384	180	32	20	17.2	69	70
D, aug-cc-pvtz	30	828	180	32	20	275	87	70
D, aug-cc-pvqz	30	828	380	64	20	175	85	80 ^e
D, aug-cc-pvqz	30	1512	470	64	31	1917	95	65 ^e

^a Number of correlated orbitals. ^b Number of atomic basis functions. ^c Calculation of the atomic orbital integrals and their contraction with the amplitudes to form the External Exchange Operator (EEO), eq 18. ^d The CPU efficiency is calculated as the sum of the CPU times on the slaves divided by the product of the elapsed time and the number of slaves. ^e The CPU timings are not reliable for the 64-bit architecture (kernel from Linux 2.4 series) for the multithreaded parts of program. ^f The machines were equipped with 4 GB of RAM memory (1 MW \approx 8 MB), D = dimer in dimer basis Set, M = monomer in dimer basis set. No symmetry was used.

**Figure 4.** Parallel scaling of a calculation on glycine-10 (see Table 2). The efficiency of the 4-node calculation is taken as 4.**Table 4.** Counterpoise Corrected Binding Energies for Benzene Dimer in Different Basis Sets^a

basis set	SCF	MP2	MP3	MP4(SDQ)	QCISD
aug-cc-pvdz	-5.168	4.219	0.765	0.919	0.833
aug-cc-pvtz	-5.159	4.645	1.097	0.976	0.998
aug-cc-pvqz	-5.157	4.790	1.216	0.947	1.047

^a At the dimer geometry of Sinnokrot and Sherrill⁵² (interplane distance = 3.4 Å, lateral shift = 1.6 Å).

Figure 4 shows the scaling of the glycine-10 calculation from 4 to 15 nodes. Scaling is almost linear from 4 to 8 nodes but only 88% efficient in going from 8 to 15 nodes.

Table 4 shows the counterpoise corrected binding energies obtained for the parallel displaced benzene dimer geometry at various levels of theory and for six different basis sets using the geometry of Sinnokrot and Sherrill.⁵² This geometry has been optimized at the MP2/aug-cc-pVQZ* level, and, due to the overestimation of the binding energy, the optimized interplane distance, 3.4 Å, is likely to be too small.

Table 5. Counterpoise Corrected Binding Energies for Benzene Dimer in Different Basis Sets^a

basis set	SCF energy	MP2 energy	QCISD energy
cc-pvdz	-2.554	1.726	0.159
cc-pvtz	-2.414	3.184	1.032
aug-cc-pvdz	-2.392	3.623	1.453
aug-cc-pvtz	-2.359	3.856	1.524

^a The shift and distance optimized at the QCISD/aug-cc-pvtz level (distance = 3.675 Å, shift = 1.870 Å).

The efficiency of our program enabled us to optimize the intermolecular distance and shift of the parallel displaced benzene dimer at the aug-cc-pvtz/QCISD level. The optimum interplane distance we found is $R_1 = 3.675$ Å, and the lateral shift is $R_2 = 1.870$ Å. As QCISD underestimates the binding energy, R_1 is almost certainly too long; the most likely distance in the real molecule is probably bracketed between these two values and is estimated to be around 3.55 Å, close to the estimated value of ref 52. Table 5 shows the calculated binding energies at the aug-cc-pvtz/QCISD geometry.

Tables 4 and 5 allow a critical evaluation of a frequently used procedure, the extrapolation of correlation energy from smaller basis calculations to larger basis sets, using MP2 energy increments. As these tables show, this procedure, although reasonably accurate for total correlation energies, overestimates the basis set effect on the binding energy, just like MP2 overestimates the binding energy. As Tables 4 and 5 show, the change in the binding energy on enlarging the basis set is overestimated by almost a factor of 3 at the MP2 level, relative to QCISD, both for the DZ-TZ and the TZ-QZ transition. Table 5 also shows that neither MP3 nor MP4(SDQ) perform significantly better than MP2. Sinnokrot and Sherrill argue, on the basis of calculations carried out with smaller basis sets at the CCSD(T) level, that the effects of higher substitutions, $\Delta\text{CCSD(T)} = E(\text{CSD(T)}) - E(\text{MP2})$, are not sensitive to the basis sets. However, this would be true only in the unlikely case if the effect of triple substitutions cancels the effect of CCSD/QCISD. We believe that the extrapolated binding energy derived by Sinnokrot

Table 6. CCSD Weak Pairs Amplitudes Approximated by the MP2 Amplitudes^a

threshold	no. of strong pairs	dimerization energy	total elapsed time per iteration	elapsed time for EEO	elapsed time for Q terms
0.0	1225	-1.717	3404	1057	380
1e-5	756	-1.719	3131	796	320
1e-4	514	-1.746	2982	638	300
5e-4	345	-1.625	2771	529	275

^a The accuracy achieved for various thresholds and timings for dimer calculation, 12 nodes. If the weak pair's energy is below the given threshold, then the pair amplitudes are substituted by MP2 amplitudes.

and Sherrill somewhat overestimates the binding energy. Similar results have been obtained in a recent work of Hill et al.⁵³

Table 6 presents the results for the MP2 weak pair approximation. In this approach, the amplitudes of the weak pairs (pairs with MP2 correlation energy below a threshold) are kept fixed at the MP2 level throughout the whole iteration process. The remaining pairs are treated as strong and are fully optimized. The external exchange, **Q** and **G** matrices for weak pairs are not needed and omitted. However, in the first iteration the external exchange is calculated for all pairs because they are needed for the singles residual calculation.

As Table 6 and our other preliminary results show, the interaction energies are well reproduced. However, the overall improvement in timings is disappointing, mainly because the evaluation of the AO integrals imposes a significant overhead. It appears that this approximation works best for a large molecule with a moderate basis set. Another reason for the less-than-expected gain is that the current implementation does not take advantage of the fixed amplitudes in the calculation of the **Y** and **Z** matrices because this would require additional disk storage and would interfere with the blocking algorithm.

Acknowledgment. This work was supported by the National Science Foundation under grant numbers CHE-0219267000 and CHE-0515922 and by the Mildred B. Cooper Chair at the University of Arkansas.

References

- (1) Cizek, J. *J. Chem. Phys.* **1966**, *45*, 4256–4266.
- (2) Bartlett, R. J. In *Modern Electronic Structure Theory*; Yarkony, D. R., Ed.; World Scientific: Singapore, 1995; pp 1047–1131.
- (3) Lee, T. J.; Scuseria, G. E. In *Quantum Mechanical Electronic Structure Calculations*; Langhoff, S. R., Ed.; Kluwer: Dordrecht, 1995; pp 47–108.
- (4) Crawford, T. D.; Schaefer, H. F., III *Rev. Comput. Chem.* **2000**, *14*, 33–136.
- (5) Taylor, P. R.; Bacskay, G. B.; Hush, N. S.; Hurley, A. C. *J. Chem. Phys.* **1978**, *69*, 1971–1979.
- (6) Pople, J. A.; Krishnan, R.; Schlegel, H. B.; Binkley, J. S. *Int. J. Quantum Chem. Symp.* **1978**, *14*, 545–560.
- (7) Bartlett, R. J.; Purvis, G. D., III *Int. J. Quantum Chem. Symp.* **1978**, *14*, 561–581.
- (8) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic Structure Theory*; Wiley: Chichester, 2000; pp 817–833.
- (9) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.
- (10) Bartlett, R. J.; Watts, J. D.; Kucharski, S. A.; Noga, J. *Chem. Phys. Lett.* **1990**, *165*, 513–522.
- (11) Saebo, S.; Pulay, P. *J. Chem. Phys.* **1987**, *86*, 914–922.
- (12) Saebo, S.; Pulay, P. *J. Chem. Phys.* **1988**, *88*, 1884–1890.
- (13) Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1996**, *104*, 6286–6297.
- (14) Schütz, M.; Werner, H.-J. *J. Chem. Phys.* **2001**, *114*, 661–681.
- (15) Subotnik, J. E.; Sodt, A.; Head-Gordon, M. *J. Chem. Phys.* **2006**, *125*, 074116/1–12.
- (16) See: www.pqs-chem.com.
- (17) Baker, D. J.; Moncrieff, D.; Saunders, V. R.; Wilson, S. *Comput. Phys. Commun.* **1990**, *62*, 25–41.
- (18) Rendell, A. P.; Lee, T. J.; Komornicki, A. *Chem. Phys. Lett.* **1991**, *178*, 462–470.
- (19) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (20) Rendell, A. P.; Lee, T. J.; Lindh, R. *Chem. Phys. Lett.* **1992**, *194*, 84–94.
- (21) MOLPRO, a package of ab initio programs designed by H.-J. Werner and P. J. Knowles. Amos, R. D.; Bernhardsson, A.; Berning, A.; Celani, P.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Knowles, P. J.; Korona, T.; Lindh, R.; Lloyd, A.W.; McNicholas, S. J.; Manby, F. R.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Rauhut, G.; Schütz, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Werner, H.-J. *Version 2002.1*.
- (22) Kendall, R. A.; Apra, E.; Bernholdt, D. E.; Bylaska, E. J.; Dupuis, M.; Fann, G. I.; Harrison, R. J.; Ju, J.; Nichols, J. A.; Nieplocha, J.; Straatsma, T. P.; Windus, T. L.; Wong, A. T. *Comput. Phys. Comm.* **2000**, *128*, 260–283.
- (23) *PQS version 3.2*; Parallel Quantum Solutions, 2013 Green Acres Road, Fayetteville, AR 72703; 2005. See: www.pqs-chem.com.
- (24) Kobayashi, R.; Rendell, A. P. *Chem. Phys. Lett.* **1997**, *265*, 1–11.
- (25) <http://www.msg.ameslab.gov/GAMESS/changes.html> (accessed September 2006).
- (26) Szalay, P.; Gauss, J. private communication, 2006.
- (27) Stanton, J. F.; Gauss, J.; Watts, J. D.; Szalay, P. G.; Bartlett, R. J. with contributions from Auer, A. A.; Bernholdt, D. B.; Christiansen, O.; Harding, M. E.; Heckert, M.; Heun, O.; Huber, C.; Jonsson, D.; Jusélius, J.; Lauderdale, W. J.; Metzroth, T.; Ruud, K. and the integral packages: MOL-ECULE (J. Almlöf and P. R. Taylor), PROPS (P. R. Taylor), and ABACUS (T. Helgaker, H. J. Aa. Jensen, P. Jørgensen, and J. Olsen).
- (28) Nieplocha, J.; Harrison, R. J.; Littlefield, R. *Proc. Supercomputing 1994*; IEEE Computer Society Press: Washington, D.C., 1994; pp 340–346.
- (29) Nieplocha, J.; Palmer, B.; Tipparaju, V.; Manojkumar, K.; Trease, H.; Apra, E. *Int. J. High Perform. Comput. Appl.* **2006**, *20*, 203–231.
- (30) Ford, A. R.; Janowski, T.; Pulay, P. *J. Comput. Chem.* **2007**, *28*, xxxx–xxxx.
- (31) Meyer, W. *J. Chem. Phys.* **1973**, *58*, 1017–1035.

- (32) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063–1079.
- (33) Pulay, P.; Saebo, S.; Meyer, W. *J. Chem. Phys.* **1984**, *81*, 1901–1905.
- (34) Matsen, F. A. *Int. J. Quantum Chem. Symp.* **1981**, *15*, 163–175.
- (35) Meyer, W. *J. Chem. Phys.* **1976**, *64*, 2901–2907.
- (36) Hampel, C.; Peterson, K. A.; Werner, H.-J. *Chem. Phys. Lett.* **1992**, *190*, 1–12.
- (37) Scuseria, G. E.; Janssen, C. L.; Schaefer, H. F., III *J. Chem. Phys.* **1988**, *89*, 7382–7387.
- (38) Schütz, M.; Werner, H.-J. *J. Chem. Phys.* **2001**, *114*, 661–681.
- (39) Dykstra, C.; Schaefer, H. F., III; Meyer, W. *J. Chem. Phys.* **1976**, *65*, 2740–2750.
- (40) Schütz, M.; Lindh, R.; Werner, H.-J. *Mol. Phys.* **1999**, *96*, 719–733.
- (41) Pulay, P.; Saebo, S.; Wolinski, K. *Chem. Phys. Lett.* **2001**, *344*, 543–552.
- (42) Baker, J.; Pulay, P. *J. Comput. Chem.* **2002**, *23*, 1150–1156.
- (43) Pulay, P.; Saebo, S. *Theor. Chim. Acta* **1986**, *69*, 357–368.
- (44) Almlöf, J. *Chem. Phys. Lett.* **1991**, *176*, 319–320.
- (45) Goto, K.; van de Geijn, R. *ACM Trans. Math. Software* Submitted for publication.
- (46) Jensen, F. *J. Chem. Phys.* **2002**, *116*, 7372–7379.
- (47) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.
- (48) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Am. Chem. Soc.* **1994**, *116*, 3500–3506.
- (49) Jaffe, R. L.; Smith, G. D. *J. Chem. Phys.* **1996**, *105*, 2780–2788.
- (50) Tsuzuki, S.; Uchimaru, T.; Matsumura, K.; Mikami, M.; Tanabe, K. *Chem. Phys. Lett.* **2000**, *319*, 547–554.
- (51) Sinnocrot, M. O.; Valeev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 10887–10207.
- (52) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200–10207.
- (53) Hill, J. G.; Platts, J. A.; Werner, H.-J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4072–4078.

CT700048U

Self-Consistent Polarization of the Boundary in the Redistributed Charge and Dipole Scheme for Combined Quantum-Mechanical and Molecular-Mechanical Calculations

Yan Zhang,[†] Hai Lin,^{*,†} and Donald G. Truhlar[‡]

Chemistry Department, University of Colorado at Denver and Health Sciences Center, Denver, Colorado 80217-3364, and Chemistry Department and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455-0431

Received January 10, 2007

Abstract: The recently developed redistributed charge (RC) and redistributed charge and dipole (RCD) schemes are electrostatic-embedding schemes to treat a quantum-mechanical/molecular-mechanical (QM/MM) boundary that passes through covalent bonds. In the RC and RCD schemes, the QM subsystem is polarized by the MM subsystem, but the MM subsystem is not polarized by the QM one; this results in an unbalanced treatment of the electrostatic interactions. In the work reported here, we developed improved schemes, namely, the polarized-boundary RC scheme (PBRC) and the polarized-boundary RCD (PBRCD) scheme, by adding self-consistent mutual polarization of the boundary region of the MM subsystem to the previous schemes. The mutual polarizations are accounted for in the polarized-boundary calculations by adjusting the boundary-region MM point charges according to the principles of electronegativity equalization and charge conservation until the charge distributions in both the QM subsystem and the polarizable region of the MM subsystem converge. In particular, we implemented three literature parametrizations of electronegativity equalization: the original electronegativity equalization method (EEM) by Mortier and co-workers, the charge equalization (QEq) method proposed by Rappé and Goddard, and a modified version of the QEq method by Bakowies and Thiel. The PBRC and PBRCD schemes were tested by calculating proton affinities for small organic compounds and capped amino acids. As compared to full-QM calculations, the PBRC and PBRCD schemes produced more accurate proton affinities, on average, than the original RC and RCD methods; the mean unsigned error in proton affinities is reduced from about 5 kcal/mol to 3 kcal/mol with little change in geometry. The improvement is encouraging and illustrates the importance of mutual polarization of the QM and MM subsystems in treating reactions where noticeable charge transfer occurs in the QM subsystem.

I. Introduction

Combined quantum-mechanical and molecular-mechanical (QM/MM) calculations^{1–120} have become very popular in the past decade. The basic idea of QM/MM is to partition

an entire system (ES), e.g., an enzyme–substrate complex in a solvent or reagents bound to a heterogeneous catalyst into two subsystems: a localized primary system (PS) where the bond-breaking, bond-forming, and/or electron excitation processes take place and a secondary system (SS) that interacts with the PS. The PS is treated at a quantum mechanical (QM) level of theory, whereas the SS is described at the molecular mechanics (MM) level. Therefore, the PS

* Corresponding author e-mail: hai.lin@cudenver.edu.

[†] University of Colorado at Denver and Health Sciences Center.

[‡] University of Minnesota.

is also called the QM subsystem, and the SS is known as the MM subsystem. Because the precise partitioning of a system into a QM and an MM subsystem does not have a basis in experiment or quantum theory, it is somewhat arbitrary, but the concept of a partition can be justified theoretically by the observation that in many reactions the electronic structure of only a small number of atoms changes significantly. The QM/MM energy for the entire system (ES) can be formally defined as the sum of the QM energy of the PS, the MM energy of the SS, and the QM/MM interaction energy between them.

$$E(\text{QM/MM};\text{ES}) = E(\text{QM};\text{PS}) + E(\text{MM};\text{SS}) + E(\text{QM/MM};\text{PS}|\text{SS}) \quad (1)$$

In comparison with isolated QM calculations on model systems, the inclusion in eq 1 of the interactions between the PS and its surroundings (the SS) provides a more realistic description of the reactive system. For large systems, in principle, the QM/MM description combines the accuracy of a quantum mechanical description with the low computational cost of molecular mechanics, making computations for large-size reactive systems feasible. However, in order to achieve this ideal situation, the QM/MM interaction must not be oversimplified.

The interactions between the PS and the SS include valence interactions, van der Waals (VDW) interactions, and electrostatic interactions. The valence interactions require special procedures (e.g., link atoms^{3,5,11,24,40–42,50,72,76,107,111,119} or localized orbitals^{6,9,15,33,51,77,78,85,87,102,109,113}), and the van der Waals interactions are typically evaluated at the MM level. The electrostatic interactions are treated differently in various QM/MM schemes, and they are the main focus of the present article. Bakowies and Thiel¹⁶ have classified the treatments of these electrostatic interactions into two general kinds of approach. In mechanical-embedding schemes, the electrostatic interactions between the PS and SS are computed at the MM level, e.g., by Coulomb's law employing atomic charges assigned to both the PS and SS atoms, and the QM calculations for the PS are performed in the gas phase. The second and more advanced kind of treatment is called electrostatic-embedding and involves QM computations for the PS that are carried out by including in the QM Hamiltonian the operators that describe the electrostatic interaction between the nuclei and electrons of the PS and the MM partial atomic charges of the SS. In such a case, the QM/MM energy can be defined by¹⁰⁷

$$E(\text{QM/MM};\text{ES}) = E(\text{MM};\text{ES}) - E(\text{MM};\text{PS}^*) + E(\text{QM};\text{PS}^{**}) \quad (2)$$

Here the asterisk (*) denotes that the PS is embedded in the electrostatic field of the SS, and the double asterisks (**) denote such an embedding in an appropriately modified electrostatic field of the SS. (Examples of appropriate modifications are discussed below.) Usually the charge models developed for full-MM calculations are employed to represent the SS in the effective QM Hamiltonian of $E(\text{QM};\text{PS}^{**})$. The use of the MM partial atomic charges is convenient for two reasons: First, most MM force fields,

like AMBER,¹²¹ CHARMM,¹²² and OPLS-AA,^{123–128} already contain parameters or protocols for generating the needed partial atomic point charges for calculating electrostatic interactions at the MM level. Second, many electronic-structure programs (e.g., Gaussian03,¹²⁹ TURBOMOLE,⁹² and ORCA¹³⁰) have the functionality of carrying out QM calculations with background point charges, and one does not need to modify the QM codes. More sophisticated representations of the SS charge density include distributed multipoles and the effective fragment potential (EFP) developed by Gordon and co-workers,¹⁹ but multipole expansions suffer from the strong dependence of the parameters on geometries,^{131–133} which limits the transferability of the parameters.¹³⁴ By including the electrostatic field due to the charge distribution of the SS into the embedded-QM calculations, the electrostatic-embedding schemes allow the PS to be polarized by the SS. The polarization perturbs the electronic structure of the PS, and it changes the energy profile of the reaction. It can even change the electronic ground state of the reagents, especially for those systems having two or more low-energy electronic states.⁸³

The polarization in the electrostatic-embedding scheme described above is unbalanced, because the PS is polarized by the SS, but the SS is not polarized by the PS. In principle, the PS and SS will polarize each other until their charge distributions are self-consistent. Schemes that allow such mutual polarization are called self-consistent mutual-polarized-embedding schemes or polarized-embedding schemes for short.¹⁶ The polarized-embedding schemes should be more accurate than electrostatic-embedding schemes, but the price that one has to pay for them is a more expensive computational cost. Complete development of the polarized-embedding QM/MM method requires a polarizable MM force field^{135–150} that has the flexibility to respond to a perturbation by an external electric field; such flexibility is not available yet in today's most popular MM force fields. Nevertheless, attempts have been made to develop polarized-embedding QM/MM schemes by combining unpolarizable MM potentials (such as AMBER,¹²¹ CHARMM,¹²² and OPLS-AA,^{123–128}) with classical polarization models.^{17,137,140,151–161} In practice these procedures are equivalent to employing polarizable force fields, but they do not assume that a generally parametrized polarizable force field is already available. The basic idea is similar to reaction field theory,^{162–164} although the response is now given by a discrete model incorporating the atomic polarizability of individual SS atoms instead of by a continuum. The classical polarization models that can be used can largely be divided into three categories: the first kind of models is based on induced dipoles (or induced multipoles),^{137,158–161} the second kind of models is based on the principle of electronegativity equalization,^{17,140,151–157} and the third is the shell model of Dick and Overhauser.¹³⁵

The first kind of polarized-embedding QM/MM scheme adds induced dipoles to the SS atomic centers^{14,16,28,114} so that the SS can respond to the electric field generated by the PS. One might wish to follow the more general theory of Stone¹⁶⁰ and assign polarizabilities of various orders (rather

than just dipole polarizabilities) to a variety of regions of the system (rather than just atomic centers). However, including only induced dipoles at atomic centers is often the most practical compromise of affordability and accuracy. Thole's empirical¹⁵⁹ treatment has been employed to model the damped behavior of short-range polarization.¹⁶ The induced dipoles are typically re-expressed as pairs of point charges in the vicinity of the SS atomic centers so as to take advantage of the ability of QM codes to carry out embedded-QM calculations with background point charges. Fitting the induced dipoles to charges on selected atomic centers is also possible.^{161,165}

The second kind of polarized-embedding QM/MM scheme allows the MM point charges in the SS to adjust in the presence of the PS^{17,27,32,154–157} according to the principle of electronegativity equalization and the principle of charge conservation; this kind of method has been labeled in a number of different ways including charge equilibration and chemical potential equalization. Since MM point charges include the contributions due to higher-order multipoles implicitly, i.e., the higher-order contributions are folded into the parameters, the adjustment of MM point charges by electronegativity equalization is a powerful way to include the response of the PS charge distribution to an external electric field and thus to account for the polarization effects. The advantage of using only point charges instead of the dipoles and higher-order multipoles is higher computational efficiency, since the computational effort is roughly proportional to the square of the number of charge sites but nine times the square of the number of dipole sites.¹⁴³ However, the polarization response of a point-charge-only model is limited in certain cases.¹⁴⁰ For example, there is no out-of-plane response for a planar molecule like benzene or water. That is, the point-charge-only model cannot represent the polarization of a benzene molecule if the external electric field is perpendicular to the molecular plane. By adopting a point-charge-only polarization model, one makes a compromise between accuracy and efficiency. For applications to large biomolecules, the errors due to the neglect of higher-order terms in the polarization treatment are often not significant, although they may be of central importance for cation- π interactions and stacking. In many cases, though they are likely to be smaller than the errors due to other approximations in the QM/MM methodology. It is also possible to improve the point-charge-only polarization model by adding a small numbers of polarizable dipoles, as has been done in the combined fluctuating charge-dipole model by Stern et al.¹⁴³

The third kind of polarized QM/MM implementation^{67,69,166} uses the ion-shell model¹³⁵ and is mainly applied in the studies of solid-state materials such as metals, metal oxides, and surface-adsorbate systems. A notable difference between crystalline materials and liquids is the periodicity of the lattices of the crystals. Usually, in QM/MM simulations of crystalline materials,^{2,20,31,43,54,64,66,67,69,73,95,166–169} the PS is treated by a cluster model embedded in a finite number of point charges (and higher-order multipoles) that mimic the infinite and periodic charge distribution of the environment (the SS). The finite number of point charges, which model

the charge-distribution of the SS, can be obtained by minimizing the difference between the electrostatic potential that is generated by the point charges and that generated by the infinite and periodic charge distribution at a set of sampling points in the active site. By doing so, one truncates the infinite and periodic system to a finite embedded cluster model, which is now much easier to handle. The polarization effects on the SS are taken into account by using the shell model,¹³⁵ which represents an ion by two charges (a positive core and a negative shell) connected by a harmonic potential. In response to the external electric field, the positions of the charges are adjusted to achieve the lowest energy.

A general conclusion that emerges from the polarizable-embedding QM/MM calculations^{10,16,27,28,32,114} is that the polarization of the SS by the PS is most significant when the PS is charged and generates large electric fields. The polarization of the SS by the reactant PS may be similar to that by the product PS so that there is some cancellation when computing relative energies (such as the energy difference between the reactants and the products), especially if the charge distribution of the PS does not change significantly during the reaction. In such a case, the effect of polarization of the SS may be less important. However, if there is significant charge transfer in the PS during a reaction, one cannot expect cancellation of errors.

The interactions between the PS and the SS are sometimes (unavoidably) complicated by making the QM/MM boundary pass through a covalent bond, leaving dangling bonds at the frontier atoms of the PS. Special care is needed to handle such a case. Treatments of QM/MM boundaries that pass through covalent bonds can be largely divided into two categories. The first category contains the so-called link-atom schemes that use H atoms^{5,11,24,41,42,72,80,82} or parametrized one-free-valence atoms^{40,50,76,111} to saturate the dangling bonds. The second category can be called the local-orbital schemes, because it contains schemes that use localized orbitals to provide a quantum mechanical description of the charge distribution near the QM/MM boundary. Two examples of the local-orbital schemes are the so-called local self-consistent field (LSCF) scheme^{6,9,15,77,113,170} and the generalized hybrid orbitals (GHO) scheme.^{33,51,78,85,87,102,103,109} The link-atom schemes are easier to implement, and thus they are widely used. The local-orbital schemes are theoretically more fundamental^{78,102} but more complicated. Extensive calculations have demonstrated that, if used carefully, both the link-atom and the local-orbital schemes can give reasonably good accuracy.

In a recent paper, we¹⁰⁷ developed two new electrostatic-embedding schemes to treat the QM/MM boundary by combining features of the local-orbital treatment with the link-atom treatment. Our schemes, which are called the redistributed charge (RC) scheme and the redistributed charge and dipole (RCD) scheme, use redistributed charges and dipoles as molecular mechanical mimics of the localized auxiliary hybrid orbitals in the GHO theory. As described in ref 107, we find it convenient to label the atoms in "tiers", i.e., the QM frontier atom as the Q1 atom, the MM boundary atom to which it is bonded as the M1 atom, the MM atoms directly connected to the M1 atom as M2 atoms, the MM

atoms directly connected to the M2 atoms as M3 atoms, and so on. In our treatment, a hydrogen link atom replaces the boundary atom and active hybrid orbital of GHO theory. This link atom, which is denoted HL, is placed on the Q1–M1 bond. The Q1–HL distance $R(\text{Q1–HL})$ depends on the Q1–M1 distance $R(\text{Q1–M1})$ by a scaling factor C_{HL} :

$$R(\text{Q1–HL}) = C_{\text{HL}}R(\text{Q1–M1}) \quad (3)$$

$$C_{\text{HL}} = R_0(\text{Q1–H})/R_0(\text{Q1–M1}) \quad (4)$$

Here $R_0(\text{Q1–H})$ and $R_0(\text{Q1–M1})$ are the MM bond distance parameters for the Q1–H and Q1–M1 stretches in the MM force field, respectively. The PS, which is now capped by the HL atoms, is called the capped PS, or CPS. Consequently, the terms $E(\text{MM};\text{PS}^*)$ and $E(\text{QM};\text{PS}^{**})$ in eq 2 are replaced by $E(\text{MM};\text{CPS}^*)$ and $E(\text{QM};\text{CPS}^{**})$, respectively, giving rise to

$$E(\text{QM/MM};\text{ES}) = E(\text{MM};\text{ES}) - E(\text{MM};\text{CPS}^*) + E(\text{QM};\text{CPS}^{**}) \quad (5)$$

In the RC scheme, one evenly distributes the MM charge of the M1 atom to the midpoints of the M1–M2 bonds. The magnitude of the bond-midpoint charges is

$$q_0^{\text{RC}} = q_{\text{M1}}/n \quad (6)$$

where n is the number of M2 atoms. In the RCD scheme, the redistributed charge q_0 and the charge on the M2 atoms (q_{M2}) are further modified in order to preserve the M1–M2 bond dipoles

$$q_0^{\text{RCD}} = 2q_0^{\text{RC}} \quad (7)$$

$$q_{\text{M2},k}^{\text{RCD}} = q_{\text{M2},k} - q_0^{\text{RC}} \quad (8)$$

where $k = 1, 2, \dots, n$. The RC and RCD schemes can be viewed as providing classical analogs to the GHO quantum descriptions of the charge distribution around the QM/MM boundary. Test calculations for geometries, proton affinities, and reaction energy profiles showed that the RC and RCD schemes provide quite good accuracy in comparison with full-QM calculations.¹⁰⁷ In particular, the mean unsigned errors (MUEs) for the proton affinities for a set of seven small organic molecules are within 3–10 kcal/mol, depending on the set of partial charges used for the embedded-QM calculations. The redistribution of the charge q_{M2} also helps in alleviating the “overpolarization” of the Q1–HL bond by q_{M2} , a problem seen in many electrostatic-embedding schemes that use link atoms to cap the PS. The overall performances of the RC and RCD schemes are roughly the same as the performance of the shifted-charge scheme¹⁷¹ (denoted Shift), which also conserves the features of the charge distribution around the QM/MM boundary, and they are significantly better than the performances of those schemes that do not conserve such features.¹⁰⁷

The RC and RCD schemes are electrostatic-embedding schemes and do not to treat the polarization of the SS by

the CPS. In this study, we improve the RC and RCD schemes by including polarization of a portion of the SS by the CPS. We use the approach of electronegativity equalization in treating the MM polarization of the SS. We note that Field²⁷ had explored the same idea of electronegativity equalization under the name of fluctuating charge in his polarized-embedding QM/MM scheme. Field studied cases where the PS (solute) and the SS (solvent) are not covalently bonded: methane and formaldehyde in water. In this article, we focus on the more complicated situations where the PS and SS are covalently bonded to each other.

The new schemes introduced in this paper will be called the PBRC scheme and the PBRCD scheme, where PB indicates that they are polarized-embedding schemes that polarize (P) the boundary (B) region of the SS, as described in detail below. The theory is given in section II, where we begin with general descriptions and proceed to implementation details. Test calculations are described in section III, and the results are presented in section IV. Section V discusses the performance of the PBRC and PBRCD schemes, and a summary and conclusions are presented in section VI.

II. Theory

All equations in this section are written in atomic units.

II.A. General Description of the QM/MM Polarization Treatments. In the PBRC and PBRCD calculations, the SS charges enter the QM Hamiltonian of the CPS as one-electron terms; this accounts for the polarization of the CPS due to the SS. The polarization of the SS due to the CPS is realized by the adjustment of the SS boundary-region charges in the presence of the electric field generated by the CPS. The variation of the SS charges is determined by the principle of electronegativity equalization, which includes the constraint of charge conservation.

The procedure for incorporating self-consistent polarization is as follows: First, MM charges are assigned to the SS atoms, and the self-consistent-field (SCF) iterations of the embedded-QM calculation are performed with fixed partial atomic charges on the SS atoms. Second, the electric field generated by the CPS (nuclei and electronic wavefunctions) and the unpolarized part (if any) of the SS are computed and imposed on the polarized part of the SS, and a new set of charges is determined for the polarized part of the SS by electronegativity equalization and charge conservation. The new set of SS charges replaces the old set of SS charges, and a new embedded-QM SCF calculation is performed with the updated SS charges. Iterations continue until the variations in the charges are smaller than preset thresholds. Although this iterative algorithm is acceptable for the present test calculations, it is inefficient, and for production work the polarization of the SS should be recomputed at every step of the regular CPS self-consistent-field iterations. Implemented in this way, the increase in cost as compared to the RC and RCD calculations should be negligible.

There are two general issues to be considered here. The first consideration is that we adopt a prescription that is in the spirit of the “intramolecular-charge-transfer” treatment in the fluctuating-charge model by Berne and co-workers.¹⁴⁰

In their intramolecular-charge-transfer prescription, charge transfer was allowed within each molecule but prohibited between molecules. In our prescription, which can be called an intragroup-charge-transfer prescription, the (SS) atoms in the boundary region are treated as a group, and charge transfer is allowed within this group only. In many MM force fields like CHARMM¹²² and OPLS-AA,^{123–128} the total charge on each functional group is constrained to zero (or an integer) during the parametrization, so that charges can be easily transferred to other molecules with similar chemical groupings.¹⁴⁷ Simple examples of a functional group are the CH₂ group in an *n*-butane molecule or a whole water molecule. A group can also be constructed by selectively putting together a number of atoms that are connected to each other via covalent bonds but do not belong to the same functional group, and we will use this more general approach, as explained in the next paragraph. We note that Field²⁷ also adopted the intramolecular-charge-transfer prescription in his polarized-embedding QM/MM scheme. Field studied cases where the SS is water as solvent. Because each water molecule formed a group, the intramolecular-charge-transfer prescription and the intra-group-charge-transfer prescription would be equivalent in Field's calculations.

In the intragroup-charge-transfer prescription, a given group is in the presence of the electric field generated by the CPS and the electric fields generated by the other SS groups; those electric fields are combined into one electric field, which is referred to as the "external electric field" in this study, meaning that this combined electric field is external to the given group, which responds by adjusting its charge distribution. In the PBRC and PBRCD schemes, the only SS group that we allow to polarize is called the boundary group, since it consists of all the atoms in tiers M2 and M3, even though these atoms need not belong to the same functional group, or is called the polarizable group, since there is only one polarizable group in the PBRC and PBRCD schemes. The boundary group is the only SS group within which charge redistribution is permitted. All the other charges on the SS atoms as well as the redistributed charges q_0 at the M1–M2 midpoints retain their values, and they are put into a second group called the unpolarized group. We made such an arrangement because the polarization effect due to the bond to the CPS should be most pronounced in the QM/MM boundary region where the M2 and M3 atoms reside. The goal of the charge equalization in this study is to optimize the SS charges that appear in the QM Hamiltonian for the embedded-QM calculations rather than to fully account for the polarization of the SS at the MM level, which is a task requiring the development of polarized force fields. Although the redistributed charges q_0 are very close to the QM/MM boundary, we did not allow them to change values for two reasons: First, the redistributed charges are classical mimics of the auxiliary hybrid orbitals in the GHO theory, in which each auxiliary orbital retains its electron occupation in the QM/MM calculations.^{33,51} Second, the redistributed-charge sites are very close to the M2 atom, and we found that the electronegativity equalization calculations in which q_0 were variable gave unrealistically large values for both q_0 and q_{M2} . Therefore, we treated q_0 in the same way as we

treated the SS atoms distant from the boundary, namely by fixing their charges.

The second consideration is that the polarized SS charges are used only in the embedded-QM calculations and do not effect the MM calculations of $E(\text{MM};\text{ES})$ and $E(\text{MM};\text{CPS}^*)$ in eq 5. Although it is a common practice (and is convenient) to use MM charge parameters as partial atomic SS charges in electrostatically embedded QM calculations, we should keep in mind that these parameters are not designed for this purpose. The MM partial charges are part of an MM force field that also includes, for example, van der Waals parameters that are cross correlated with the charge parameters, and the force field is parametrized to be used as a whole for calculating MM energies, not for polarizing a quantum calculation. For the same reason, it is not appropriate to use charges from the polarized-QM calculation in the MM force field. Anyway, charges need to be consistent with the rest of the formalism and cannot be transferred between formalisms without validation.

We have implemented three literature methods that are based on the principle of electronegativity equalization, in particular the charge equalization (QEq) method employing a shielded Coulomb term (SCT) by Rappé and Goddard,¹⁵³ the modified QEq method by Bakowies and Thiel (BT),¹⁷ and the original electronegativity equalization method (EEM) of Mortier and co-workers.¹⁵¹ These methods will be abbreviated as SCT, BT, and EEM, respectively. We begin with brief descriptions of the QEq and EEM methodologies in sections II.B and II.C, respectively, and these sections also include our modifications to the original formulas by inclusion of external electric fields. Full details of the QEq and EEM methods, beyond the brief descriptions given here, can be found in the literature and will not be repeated here.

II.B. Treatments Based on the QEq Method and Its Variants. In the QEq approach of Rappé and Goddard,¹⁵³ the total electrostatic energy of a molecule of N atoms is written as the sum of the energy of all atoms in the molecule and the interatomic electrostatic energy. If we apply this to the polarizable group, we obtain

$$E(Q_1 \cdots Q_N) = \sum_A \left(E_{A0} + \chi_A^0 Q_A + \frac{1}{2} J_{AA}^0 Q_A^2 \right) + \sum_{A < B} J_{AB} Q_A Q_B \quad (9)$$

where Q_A or Q_B (A or $B = 1, 2, \dots, N$) is the charge at atomic center A or B in the polarizable SS group, E_{A0} is the energy of an isolated neutral atom A , χ_A^0 is the electronegativity of this isolated atom, J_{AA}^0 is the Coulomb repulsion integral of two electrons residing at the same isolated atom, J_{AB} is the Coulomb interaction integral between unit charges on atomic centers A and B , and the last term of eq 9 is the interatomic internal electrostatic energy of this group. The chemical potential at atomic center A is given by

$$\chi_A(Q_1 \cdots Q_N) = \partial E / \partial Q_A = \chi_A^0 + J_{AA}^0 Q_A + \sum_{B \neq A} J_{AB} Q_B \quad (10)$$

Since we have applied eq 9 to a group rather than to a whole system, we must modify it so that it includes the external electric field U_{ext} , which is the summation of the field due to the CPS ($U_{\text{A,CPS}}$) and the field due to the unpolarized group charges, the latter including the contribution by the redistributed charge q_0 at the M1–M2 midpoint

$$U_{\text{A,ext}} = U_{\text{A,CPS}} + \sum_{\text{C}} J_{\text{AC}} Q_{\text{C}} \quad (11)$$

where C denotes an SS center not in the polarized group, and J_{AC} is the Coulombic interaction integral between atoms A and C. The SS centers include the SS atoms and the M1–M2 bond midpoint. The energy of atom A in the external field is written as

$$E_{\text{A,ext}} = U_{\text{A,CPS}} Q_{\text{A}} + Q_{\text{A}} \sum_{\text{C}} J_{\text{AC}} Q_{\text{C}} \quad (12)$$

Consequently, the total energy $E(Q_1 \cdots Q_N)$ of the boundary-group atoms is given by

$$E(Q_1 \cdots Q_N) = \sum_{\text{A}} E_{\text{A}} + \sum_{\text{A}<\text{B}} E_{\text{AB}} + \sum_{\text{A}} E_{\text{A,ext}} \quad (13)$$

where E_{A} is the energy of atom A (the sum of three terms in the parentheses in eq 9), and E_{AB} is the interatomic electrostatic of interaction of atoms A and B (the last term in eq 9). The atomic chemical potential at atom A is then rewritten as

$$\chi_{\text{A}}(Q_1 \cdots Q_N) = \chi_{\text{A}}^0 + J_{\text{AA}}^0 Q_{\text{A}} + U_{\text{A,CPS}} + \sum_{\text{C}} J_{\text{AC}} Q_{\text{C}} + \sum_{\text{A} \neq \text{B}} J_{\text{AB}} Q_{\text{B}} \quad (14)$$

According to the principle of electronegativity equalization, the chemical potential should be equal at all atomic centers within the molecule. This leads to N equations

$$\bar{\chi} = \chi_1 = \cdots = \chi_N \quad (15)$$

where $\bar{\chi}$ is the common value. One additional equation comes from the principle of charge conservation, which imposes a constraint on the total charge

$$Q_{\text{tot}} = \sum_{i=1}^N Q_i \quad (16)$$

Substituting eq 14 into eq 15 and using eq 16, one obtains a total of $N + 1$ equations given in matrix form as

$$\begin{bmatrix} J_{11}^0 & J_{12} & \cdots & J_{1N} & -1 \\ J_{21} & J_{22}^0 & \cdots & J_{2N} & -1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ J_{N1} & J_{N2} & \cdots & J_{NN}^0 & -1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_N \\ \bar{\chi} \end{bmatrix} = \begin{bmatrix} -\chi_1^0 - U_{1,\text{CPS}} - \sum_{\text{C}} J_{1\text{C}} Q_{\text{C}} \\ -\chi_2^0 - U_{2,\text{CPS}} - \sum_{\text{C}} J_{2\text{C}} Q_{\text{C}} \\ \vdots \\ -\chi_N^0 - U_{N,\text{CPS}} - \sum_{\text{C}} J_{N\text{C}} Q_{\text{C}} \\ Q_{\text{tot}} \end{bmatrix} \quad (17)$$

The QEq charges in the presence of an external field are obtained by solving eq 17.

The Coulomb interactions are evaluated approximately by using an empirical function. One function tested by Rappé and Goddard¹⁵³ is

$$J_{\text{AB}} = \frac{1}{\epsilon R_{\text{AB}}} \quad (18)$$

where ϵ is the dielectric constant taken to be 2, and R_{AB} is the distance between atoms A and B. Although eq 18 does not give a correct description of J_{AB} when $R \rightarrow 0$, it was shown that with $\epsilon = 2$, eq 18 produced reasonable charges for a set of molecules at their equilibrium geometries.¹⁵³ It is interesting to see whether such an empirical treatment can produce reasonable charges in the present study. For this reason, we make use of eq 18 and denote the corresponding calculations as SCT. The SCT results should be interpreted with care by keeping in mind that eq 18 is approximate and is not valid if the atoms are close to each other.

Bakowies and Thiel¹⁷ proposed a modified QEq method, where the Klopman–Ohno function^{172,173} is employed to compute the Coulomb interactions because it is more realistic than eq 18 for small R_{AB} . Their formula is

$$J_{\text{AB}} = \frac{1}{\sqrt{R_{\text{AB}}^2 + \left(\frac{1}{2J_{\text{AA}}^0} + \frac{1}{2J_{\text{BB}}^0} \right)^2}} \quad (19)$$

Employing eq 19, Bakowies and Thiel¹⁷ optimized a set of parameters for the elements H, C, N, and O for the QM/MM calculations where the QM methods are AM1¹⁷⁴ and MNDO¹⁷⁵ and the force field is MM3.^{176–178} In this article, computations employing eq 19 and the set of parameters optimized by Bakowies and Thiel are denoted BT.

II.C. Treatments Based on the EEM Model. The EEM model is also based on the principle of electronegativity equalization.¹⁵² In the EEM model, the atomic electronegativity in a molecule can be written as

$$\chi_{\text{A}} = \chi_{\text{A}}^* + 2\eta_{\text{A}}^* Q_{\text{A}} + \sum_{\text{B} \neq \text{A}} \frac{Q_{\text{B}}}{R_{\text{AB}}} \quad (20)$$

where Q_A and Q_B are charges on atoms A and B, respectively, and χ_A^* and η_A^* are defined by

$$\chi_A^* = \chi_A^0 + \Delta\chi \quad (21)$$

$$\eta_A^* = \eta_A^0 + \Delta\eta \quad (22)$$

Here, χ_A^0 and η_A^0 are the electronegativity and hardness of the isolated atom, respectively, and $\Delta\chi$ and $\Delta\eta$ are correction terms due to the fact that an atom is now in a molecule instead of being isolated, thus changing its size and shape.

Again, just as for the QEq method, we modified the original EEM equations so that they apply to a group in the presence of an external electric field U_{ext} . In this work, the external field at atomic center A in the boundary group is the summation of the field due to the CPS ($U_{A,\text{CPS}}$) and the field due to the unpolarized-group charges

$$U_{A,\text{ext}} = U_{A,\text{CPS}} + \sum_C \frac{Q_C}{R_{AC}} \quad (23)$$

where C denotes a center in the unpolarized SS group, and R_{AC} is the distance between atoms A and C. In the presence of the external field, the total energy of the boundary-group atoms can be written as

$$E = \sum_A (E_A^{\text{atm}} + E_A^{\text{int}} + E_{A,\text{ext}}) \quad (24)$$

where E_A^{atm} is the energy of atom A in the absence of external electric field, E_A^{int} is the interatomic interactions attributed to atom A, and $E_{A,\text{ext}}$ is the potential energy of atom A due to the external field. The electronegativity of atom A in the presence of the external potential is thus given by

$$\chi_A = \chi_A^* + U_{A,\text{CPS}} + \sum_C \frac{Q_C}{R_{AC}} + 2\eta_A^* Q_A + \sum_{B \neq A} \frac{Q_B}{R_{AB}} \quad (25)$$

In comparison with the expression in original EEM model (eq 20), here one has two additional terms accounting for the external field.

Upon substituting eq 25 in eqs 15 and 16, one obtains a set of $N + 1$ coupled equations. The equations are solved for the electronegativity $\bar{\chi}$ and the atomic charges. Bultinck et al. had listed in Table 1 of ref 154 several sets of EEM parameters, and we found that the set of parameters developed by Mortier and co-workers showed the best agreements between the EEM- and QM-calculated dipole moments for small organic molecules in our test calculations. Thus, we adopt in the present study the EEM parameters by Mortier and co-workers, and the corresponding QM/MM calculations are denoted EEM.

III. Computations

The new polarized QM/MM schemes, PBRC and the PBRCD, were tested by calculating geometries and proton affinities for seven small organic molecules and four capped amino acids, one of which is protonated in two different places. The proton affinity is defined as the energy difference

between a chemical species ($X^- + H^+$ or $X + H^+$) and its protonated form (XH or XH^+), each at its optimized geometry. The proton affinity is a challenging test for the QM/MM methodology, because the protonation causes significant changes in the charge distribution of the PS and proton affinities are very sensitive to the treatment of electrostatic interactions between the PS and SS. This is especially the case if the protonation site is very close to the QM/MM boundary (as it is in our test cases) and if significantly polar SS functional groups are nearby.

The small organic molecules in the test suite are $\text{CH}_3\text{-CH}_2\text{OH}$, $\text{CH}_3\text{-CH}_2\text{SH}$, $\text{CH}_3\text{-CH}_2\text{NH}_3^+$, $\text{CH}_3\text{-CH}_2\text{COOH}$, $\text{CF}_3\text{-CH}_2\text{OH}$, $\text{CH}_2\text{OH-CH}_2\text{OH}$, and $\text{CH}_2\text{OH-CH}_2\text{SH}$, where only the protonated form is listed, and the dash indicates the boundary between the SS on the left and the PS on the right. These molecules have been employed in the recent study¹⁰⁷ of the original RC and RCD schemes. The capped amino acids in the test suite are histidine, glutamic acid, lysine, and tyrosine, for which the N-terminals are capped by an acetyl (Ace) group and the C-terminals are capped by an N-methylamide (NMe) group. Figure 1a–1f shows the models of histidine (Ace-His⁺-NMe), histidine deprotonated at the δ position (Ace-His $^\delta$ -NMe), histidine deprotonated at the ϵ position (Ace-His $^\epsilon$ -NMe), glutamine acid (Ace-Glu-NMe), lysine (Ace-Lys⁺-NMe), and tyrosine (Ace-Tyr-NMe). The side chains are treated at the QM level, while the backbones are described by an MM force field.

For the present study, the Gaussian03¹²⁹ program is employed for QM calculations, TINKER¹⁷⁹ is used for MM calculations, and QMMM¹⁸⁰ is utilized for QM/MM calculations. In both the PBRC and PBRCD computations, treatments based on the QEq (including the SCT and BT treatments) and EEM schemes were tested. For comparison, we also carried out calculations employing the electrostatic-embedding RC and RCD schemes, where no polarization of the SS is allowed. Full-QM calculations were performed and were used as benchmarks for assessment of the QM/MM methods.

The QM level for the examples in this article is Hartree–Fock theory¹⁸¹ with the MIDI¹⁸² basis set, which was used in our previous paper¹⁰⁷ on the RC and RCD schemes. The OPLS-AA force field^{123–128} implemented in TINKER¹⁷⁹ was employed for the MM descriptions. For some molecules, several force field parameters were missing, and we solved the problem by using parameters for similar atom types; the parameters are given in the Supporting Information. Although the protonated and deprotonated species have different sets of atom types, we used only one set of MM parameters throughout the calculations, in particular the set of parameters for the species that has one more atom. As discussed previously,¹⁰⁷ such a selection of MM parameters is not perfect, but is the only straightforward option for reaction path calculations. Thus we are testing the methods under the conditions that would be used in actual applications to chemical reactions.

A question that arises is how the accuracy would change if we used more popular methods of electronic structure theory. To explore this issue, we carried out additional calculations with the B3LYP and MP2 levels of theory and

Table 1. Atomic Charges Derived from the Full-QM and QM/MM Calculations for the Ace-His⁺-NME System^a

atom	QM	RC	PBRC			RCD	PBRCD		
			SCT ^b	BT ^c	EEM		SCT ^b	BT ^c	EEM
				PS					
C13	-0.376	-0.514	-0.544	-0.570	-0.541	-0.475	-0.512	-0.564	-0.505
C14	0.162	0.284	0.314	0.327	0.306	0.277	0.309	0.325	0.300
N15	-0.130	-0.269	-0.288	-0.279	-0.274	-0.267	-0.286	-0.277	-0.272
C16	-0.207	-0.281	-0.277	-0.303	-0.288	-0.277	-0.275	-0.298	-0.287
C17	0.074	0.075	0.104	0.076	0.080	0.076	0.104	0.078	0.081
N18	-0.241	-0.155	-0.187	-0.159	-0.169	-0.154	-0.185	-0.159	-0.166
H19	0.168	0.186	0.179	0.179	0.181	0.186	0.180	0.190	0.182
H20	0.151	0.226	0.219	0.229	0.215	0.224	0.218	0.236	0.212
H21	0.321	0.451	0.462	0.452	0.454	0.453	0.464	0.454	0.455
H22	0.250	0.268	0.261	0.272	0.265	0.269	0.262	0.272	0.266
H23	0.229	0.237	0.227	0.234	0.234	0.238	0.228	0.234	0.235
H24	0.385	0.369	0.372	0.370	0.371	0.369	0.373	0.370	0.371
				SS					
N7	-0.582	-0.500	-0.409	-0.937	-0.444	-0.547	-0.451	-0.967	-0.482
C9	0.664	0.500	0.217	0.663	0.184	0.453	0.167	0.615	0.117
H12	0.047	0.060	-0.009	0.006	0.003	0.013	-0.057	-0.031	-0.049
C2	0.762	0.500	0.361	0.702	0.451	0.500	0.359	0.689	0.450
H11	0.303	0.300	0.283	0.526	0.231	0.300	0.285	0.521	0.232
O10	-0.515	-0.500	-0.290	-0.441	-0.172	-0.500	-0.289	-0.442	-0.162
N25	-0.535	-0.500	-0.294	-0.660	-0.393	-0.500	-0.293	-0.665	-0.387
redistributed charge (q_0)	n/a	0.047	0.047	0.047	0.047	0.093	0.093	0.093	0.093
link atom (HL)	n/a	0.122	0.157	0.172	0.165	0.082	0.120	0.139	0.129

^a The side chain is the PS, and the backbone is the SS (see also Figure 1). The QM level is HF/MIDI!, and the MM force field is OPLS-AA. The iterative procedure for polarization was initiated by using the OPLS-AA charges for the embedded-QM calculations. ^b QEq model with SCT of Rappé and Goddard. ^c QEq model of Bakowies and Thiel.

the 6-31+G* basis set,^{183–186} where the B3LYP denotes the Becke 3-parameter Lee–Yang–Parr density functional method,^{187,188} and MP2 denotes Møller–Plesset second-order perturbation theory.¹⁸⁹

In the self-consistent polarization treatment, the M2 and M3 atoms in the SS constitute the boundary group, whereas the other SS atoms and the redistributed charge q_0 at the M1–M2 bond midpoint constitute the unpolarized SS group. The iterative procedure was initiated by assigning MM partial charges to the SS atoms; these may be called the initial or unpolarized charges. These unpolarized charges remain constant through the calculation for the unpolarized SS group but are just initial charges for the polarized boundary group. It is of interest to see the sensitivity of the calculations to the initial SS charges. Therefore we tried two choices for the unpolarized charges in the cases of small organic molecules, where, in addition to OPLS-AA charges, we also used charges obtained by fitting the electrostatic potential (ESP) using the Merz–Kollman algorithm;^{190,191} those ESP charges had been computed in our previous study of the RC and RCD schemes.¹⁰⁷ In that study, we found that the ESP charges worked the best among several charge models examined for electrostatically embedded QM/MM calculations of the protonation of small organic molecules in the gas phase. For the capped amino acids, we only employed the OPLS-AA charges as partial atomic charges on the SS atoms in the embedded-QM calculations in the present paper. We note that the ESP charges can be problematic for large molecules due to the large uncertainty of the charges on buried atoms derived from the fitting procedure to electro-

static potentials. Nevertheless, when they can be computed stably, the ESP charges are worthy of consideration.

The convergence thresholds of the self-consistent polarization calculations were set to 0.005 e for the maximum change and 0.002 e for the root-mean-square change in the QEq and EEM charges for the SS atoms.

Geometry optimization for large molecules is generally difficult because of multiple local minima. When comparing QM/MM and full-QM calculations, we wanted to make the comparisons for a given molecule such that the QM and QM/MM calculations have the same conformations. To accomplish this we followed a standard procedure so that the optimized geometries resemble each other in a systematic way. In particular, for a given molecule, we began by optimizing the geometry for the protonated species (XH or XH⁺) at the full-QM level; this served as a starting point both for the full-QM geometry optimization for the deprotonated species (X⁻ or X) and for the QM/MM geometry optimization for the protonated geometry. Finally, starting from the optimized QM/MM protonated geometry, we optimized the QM/MM deprotonated geometry. Visualization of the superimposed geometries ensured that the full-QM and QM/MM conformations are approximately similar to each other. For the capped amino acids, we also made comparisons of single-point energies at fixed geometries, as discussed below.

IV. Results

Figure 2 illustrates the convergence of the embedded-CPS energy in a PBRC single-point energy calculation using the

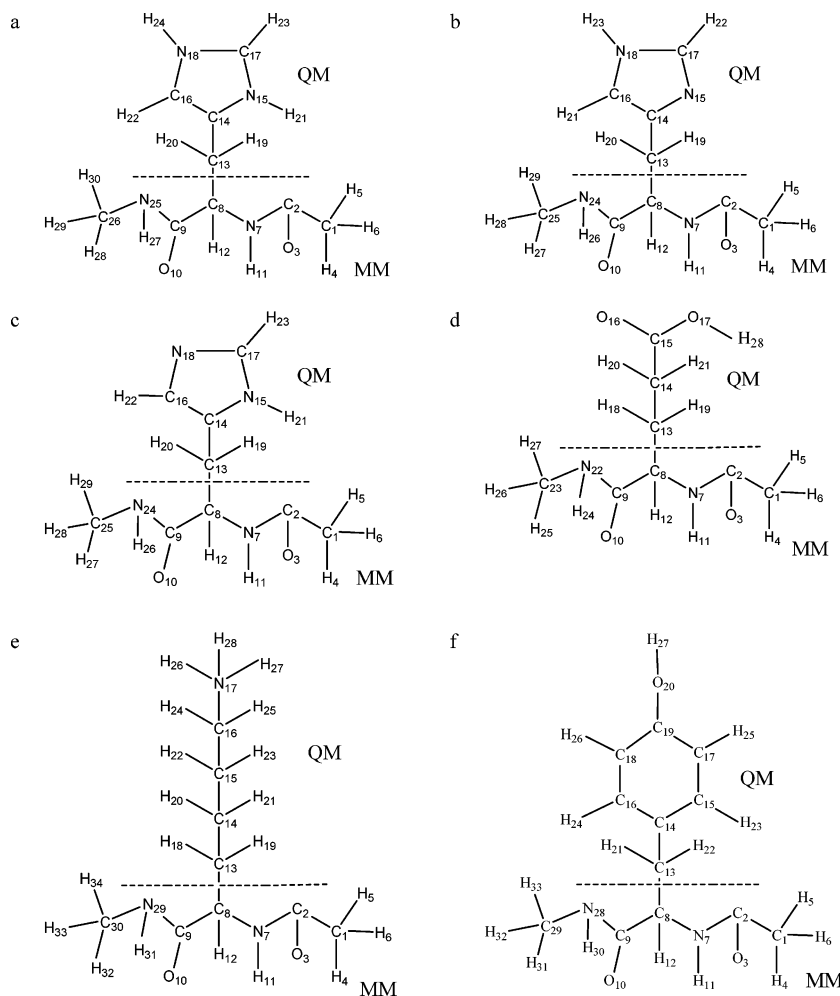


Figure 1. The capped amino acids in the test suite: (a) Ace-His⁺-NMe, (b) Ace-His^δ-NMe, (c) Ace-His^ε-NMe, (d) Ace-Glu-NMe, (e) Ace-Lys⁺-NMe, and (f) Ace-Tyr-NMe.

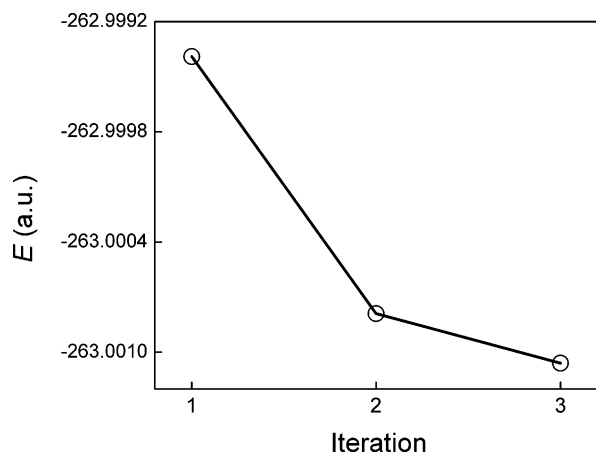


Figure 2. The convergence of the embedded-QM energy for the CPS in the Ace-His⁺-NMe system in the PBRC calculation employing the SCT scheme.

SCT parametrization for the Ace-His⁺-NMe system. The embedded-CPS energy formally includes the QM energy of the CPS, Coulombic interaction energy of the SS charges, and electrostatic interaction energy between the CPS and the SS charges. Figures 3 and 4 display for the same calculation the convergence of the ESP charges for the PS atoms and the convergence of the SCT charges for the SS atoms,

respectively. Only the heavy atoms whose charges changed by more than 0.001 e are shown in Figures 3 and 4.

In Table 1, we compare the atomic charges of Ace-His⁺-NMe derived from the full-QM and QM/MM computations; the geometries were optimized at the given levels of theory. In the case of full-QM calculation, we list ESP charges. For QM/MM calculations, we give the ESP charges for the CPS atoms including the hydrogen link atom, we give OPLS-AA charges for the SS atoms when the RC and RCD methods are used, and we give electronegativity-equalized or charge-equilibrated charges for the polarized SS atoms when the PBRC and PBRCD schemes are used (the unpolarized SS atoms of the PBRC and PBRCD methods are not shown in Table 1). The redistributed charge q_0 was fixed to $q_{M1}/3$ in the RC and PBRC calculations and to $2q_{M1}/3$ in the RCD and PBRCD calculations, where q_{M1} is the OPLS-AA charge on the M1 atom in the SS.

The proton affinities of the seven small organic molecules are listed in Table 2, where results are shown both for calculations using the OPLS-AA charges to initiate the self-consistent polarization procedure and for those using ESP charges for initialization. Table 3 tabulates the QM/MM optimized Q1–M1 bond distances for both the neutral and charged species involved in the calculations of Table 2. The QM/MM geometries and energetics are compared with full-

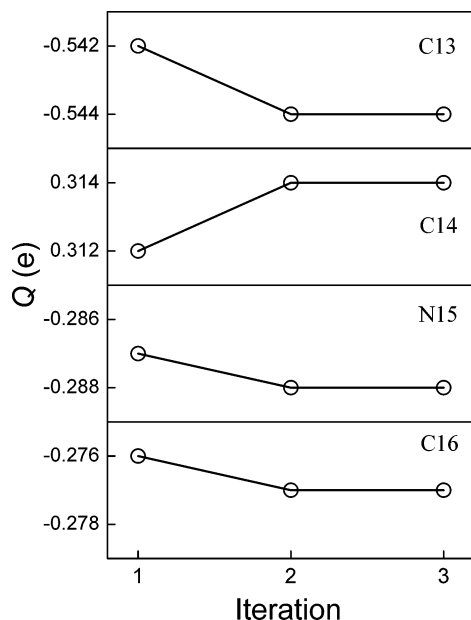


Figure 3. The convergence of the ESP charges for the CPS heavy atoms in the Ace-His⁺-NMe system in the PBRC calculation employing the SCT scheme. Two atoms C17 and N18, which are distant from the QM/MM boundary, underwent negligible (<0.001 e) changes of the charges over the iterations and are therefore not shown.

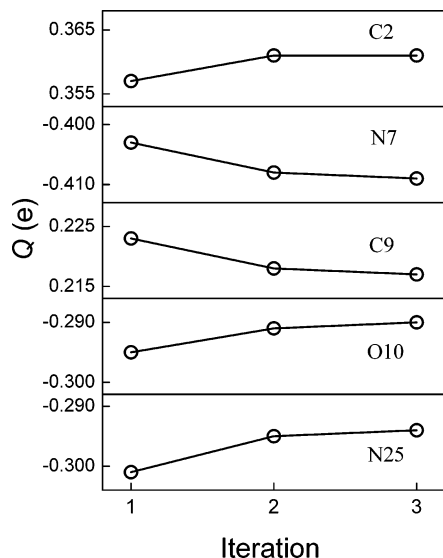


Figure 4. The convergence of the SCT charges for the SS heavy atoms in the Ace-His⁺-NMe system in the PBRC calculation employing the SCT scheme.

QM calculations, and mean unsigned errors (MUEs) and mean signed errors (MSEs) were computed in order to give an assessment of the performance for the QM/MM treatments; in particular, the MUE and MSE were calculated for each QM/MM treatment by averaging the differences between the full-QM and the QM/MM results over all involved molecules.

Because parameters are not available for the fluorine atom in the BT and the EEM parameter sets, the CF₃CH₂OH molecule was excluded from the BT and EEM calculations as well as from their MUE and MSE evaluations. Consequently, to make the comparison on the same grounds, we

computed for each of the other QM/MM treatments two sets of MUE (and MSE), one of which includes CF₃CH₂OH and the other does not. The results including CF₃CH₂OH are especially interesting in that they can be compared to our previous paper.¹⁰⁷

The proton affinities for the capped amino acids are presented in Table 4. In addition to the PBRC and PBRC schemes where the M2 and M3 charges in the SS are polarized, we also examined (for the SCT parametrization only) a more general scheme where all SS atoms were put into a single polarizable group. Table 5 displays the QM/MM optimized Q1–M1 bond distances of the molecules involved in the calculations involved in Table 4. The MUE and MSE were computed for each QM/MM treatment by averaging over all capped amino acids and using the full-QM results as the standard reference data.

To help in examining the influence of geometry on the proton affinities of the capped amino acids, we computed the QM/MM proton affinities using the same geometries, namely, the optimized full-QM geometries, for all QM/MM schemes. These are given in Table 6, and they may be denoted as QM/MM//full-QM proton affinities, where A/B means a single-point energy calculation at the A level using a geometry optimized at the B level.

Finally, overall mean unsigned errors for proton affinities are computed in Table 7 by averaging over the results for organic molecules and capped amino acids.

V. Discussion

The discussion will focus on the calculations with QM = HF/MIDI! level of theory except for section V.E, where additional calculations with two QM levels (B3LYP/6-31+G* and MP2/6-31+G*) of theory are analyzed, and for section V.F, where an overall assessment of the new boundary treatment for all three QM levels are given.

V.A. Convergence of the Self-Consistent Polarization Procedure. There are two kinds of convergence one might consider. The first is convergence with respect to allowing more and more atoms to be polarized; the second is convergence of the self-consistent iterations for a given number of polarized atoms. The first type of convergence, however, is not one of the goals of the present paper, and the results in Table 4 show that, in practice, poor results were produced for proton affinities when all SS atoms were put into a polarizable group and were permitted to vary their charges in the self-consistent polarization procedure. Most currently available molecular mechanics force fields are parametrized to give the correct results in the absence of explicit polarization. To polarize the entire MM system would therefore require a new parametrization. However, the MM force fields were parametrized for use in a fully MM calculation, and the QM/MM boundary can introduce inconsistent electrostatic interactions that can lead to unstable or unphysical polarization. Our polarized boundary treatment is primarily directed to eliminating this problem, and so the goal is to parametrize the region that is most likely to suffer from this inconsistency (and thereby alleviate it) rather than the regions that are most polarizable. Therefore the rest of

Table 2. Proton Affinities (kcal/mol) for Small Organic Molecules^a

molecule (SS–PS)	QM ^b	initial charge	RC ^b	PBRC			PBRCD			
				SCT ^c	BT ^d	EEM	SCT ^c	BT ^d	EEM	
CH ₃ –CH ₂ OH	416.8	OPLS-AA	427.2	426.8	426.6	427.0	431.5	431.1	430.9	431.3
			ESP	423.9	423.6	423.4	423.8	425.4	425.0	424.8
CH ₃ –CH ₂ SH	381.5	OPLS-AA	386.7	385.9	385.5	386.4	389.5	388.6	388.2	389.1
			ESP	384.7	383.9	383.6	384.3	395.6	384.8	384.4
CH ₃ –CH ₂ NH ₃ ⁺	232.8	OPLS-AA	233.3	233.7	234.0	233.5	236.5	236.9	237.1	236.6
			ESP	231.0	231.4	231.7	231.2	232.0	232.4	232.7
CH ₃ –CH ₂ COOH	375.3	OPLS-AA	379.9	379.2	378.9	379.6	382.2	381.6	381.2	381.9
			ESP	378.1	377.5	377.2	377.9	378.9	378.3	378.0
CF ₃ –CH ₂ OH	396.8	OPLS-AA	415.5	415.3	n/a	n/a	398.7	398.4	n/a	n/a
			ESP	408.8	408.5	n/a	n/a	395.4	395.2	n/a
CH ₃ OH–CH ₂ OH	413.2	OPLS-AA	424.2	415.8	416.6	418.3	420.0	411.4	411.6	414.2
			ESP	422.0	413.6	414.1	416.2	415.3	406.6	406.1
CH ₂ OH–CH ₂ SH	376.5	OPLS-AA	383.4	378.1	379.0	380.2	380.8	375.5	375.9	377.8
			ESP	382.4	376.8	377.4	379.0	378.3	372.7	372.6
MUE (seven) ^e		OPLS-AA	6.4/8.2	3.9/6.0	4.1	4.8	7.4/6.6	5.8/5.2	5.5	5.8
			ESP	4.9/5.9	2.3/3.6	2.3	3.3	5.1/4.6	4.2/3.8	4.1
MSE (seven) ^f		OPLS-AA	6.4/8.2	3.9/6.0	4.1	4.8	7.4/6.6	4.8/4.4	4.8	5.8
			ESP	4.3/5.4	1.8/3.2	1.9	2.7	4.9/4.0	0.6/0.3	0.4

^a The QM level is HF/MIDI!, and the MM force field is OPLS-AA. The iterative procedure for polarization was initiated by using the OPLS-AA or ESP charges for the embedded-QM calculations. ^b Reference 107. The mean errors in ref 107 differ slightly from those in this paper because they were computed from unrounded data. In this paper, all mean errors were computed from data rounded to the nearest 0.1 kcal/mol. ^c QEq model with SCT of Rappé and Goddard. ^d QEq model of Bakowies and Thiel. ^e Mean unsigned error excluding/including CF₃–CH₂OH. ^f Mean signed error excluding/including CF₃–CH₂OH.

this subsection is devoted to the other convergence issue, namely convergence of the iterative steps.

Although, as pointed out above, the polarized boundary step should be included in each QM SCF step for efficiency, the present algorithm is well suited for studying convergence. For the calculations in the present work, we found rapid convergence of the self-consistent polarization iterations; typically, the convergence was achieved within three sets of iterations. This is illustrated here for the Ace-His⁺-NMe system as an example. The Ace-His⁺-NMe system was selected for demonstration because the charged CPS can have strong polarization effects on the SS and because the electron delocalization over the imidazole ring in the CPS makes the CPS unusually polarizable by the SS. For brevity, we only discuss the PBRC and RC calculations here, but we note that the PBRCD and RCD calculations are similar.

Figure 2 shows that for the PBRC calculation employing the SCT scheme, the energy in the embedded-QM calculations drops by 0.88 kcal/mol and 0.17 kcal/mol in the second and the third iterations, respectively. The convergence thresholds (maximum change <0.005 e and root-mean-square change <0.002 e for the charges on the SS atoms in the boundary region) in our tests are modest, because tighter thresholds do not lead to systematically higher accuracy in the QM/MM calculations. The convergence thresholds we used in this study should be adequate for most QM/MM applications where only the energy is important (for example, Monte Carlo calculations¹⁹²), but tighter thresholds may be needed for molecular dynamics calculations requiring gradients.

Figures 3 and 4 illustrate the rapid convergence of the atomic charges for the heavy atoms in the CPS and in the SS in the PBRC calculations employing the SCT scheme.

The charge variations after the second iteration are almost negligible (≤ 0.001 e). As expected, the charge variations in the iterative procedure are more prominent for the CPS atoms close to the QM/MM boundary than for the CPS atoms distant to the QM/MM boundary. The charges on C17 and N18 changed negligibly (<0.001 e) during the whole self-consistent procedure due to shielding effects.

V.B. Case Study of the Atomic Charges. In this subsection, we examine the converged atomic charges resulting from the RC and PBRC calculations. Again we use the Ace-His⁺-NMe system as the example. Although we focus on the RC and PBRC schemes, the conclusions are also applicable to the RCD and PBRCD calculations.

Table 1 shows only small differences between the RC and PBRC calculations of the charges on the CPS atoms. Furthermore, the three treatments (SCT, BT, and EEM) yielded quite similar ESP charges on the CPS atoms in the PBRC calculations. For example, the ESP charge on C13 (which is the Q1 atom) is -0.51 e in the RC calculation, and the charge increases in the PBRC calculations to -0.54 e when employing the SCT treatment or the EEM treatment to -0.57 e when employing the BT treatment. These changes are in the range of 0.03–0.06 e. For the CPS atoms that are distant from the QM/MM boundary, the changes in the charges are even smaller. The small differences in the ESP charges on the CPS atoms between the RC and PBRC calculations and between the three different polarization treatments in the PBRC calculations imply that, for this molecule, the polarization on the CPS by the SS is treated reasonably even in the electrostatic-embedding schemes.

Generally speaking, the ESP charges on the PS atoms in the QM/MM computations agree qualitatively with those in

Table 3. QM/MM Optimized Q1–M1 Bond Distances (Å) in Comparison with Full-QM Results for Small Organic Molecules^a

molecule (SS–PS)	QM ^b	initial charge	RC ^b	PBRC			RCD ^b	PBRCD		
				SCT ^c	BT ^d	EEM		SCT ^c	BT ^d	EEM
CH ₃ –CH ₂ OH	1.527	OPLS-AA	1.519	1.519	1.519	1.519	1.514	1.513	1.513	1.514
		ESP	1.523	1.523	1.523	1.523	1.522	1.521	1.521	1.521
CH ₃ –CH ₂ SH	1.539	OPLS-AA	1.519	1.519	1.519	1.519	1.514	1.514	1.514	1.514
		ESP	1.523	1.523	1.523	1.523	1.521	1.521	1.521	1.521
CH ₃ –CH ₂ NH ₃ ⁺	1.528	OPLS-AA	1.517	1.517	1.517	1.517	1.513	1.513	1.513	1.513
		ESP	1.520	1.520	1.520	1.520	1.519	1.519	1.518	1.519
CH ₃ –CH ₂ COOH	1.533	OPLS-AA	1.521	1.521	1.521	1.521	1.516	1.516	1.516	1.516
		ESP	1.524	1.524	1.524	1.524	1.523	1.523	1.523	1.523
CF ₃ –CH ₂ OH	1.498	OPLS-AA	1.537	1.537	n/a	n/a	1.562	1.562	n/a	n/a
		ESP	1.537	1.537	n/a	n/a	1.558	1.557	n/a	n/a
CH ₂ OH–CH ₂ OH	1.521	OPLS-AA	1.523	1.526	1.524	1.527	1.527	1.529	1.527	1.531
		ESP	1.526	1.527	1.526	1.529	1.531	1.533	1.531	1.534
CH ₂ OH–CH ₂ SH	1.529	OPLS-AA	1.525	1.526	1.525	1.528	1.528	1.530	1.528	1.531
		ESP	1.527	1.528	1.527	1.529	1.532	1.533	1.531	1.534
CH ₃ –CH ₂ O [–]	1.594	OPLS-AA	1.562	1.561	1.561	1.561	1.548	1.548	1.547	1.548
		ESP	1.571	1.571	1.570	1.571	1.567	1.566	1.566	1.567
CH ₃ –CH ₂ S [–]	1.550	OPLS-AA	1.525	1.524	1.524	1.525	1.517	1.516	1.516	1.517
		ESP	1.531	1.530	1.530	1.530	1.528	1.528	1.527	1.528
CH ₃ –CH ₂ NH ₂	1.543	OPLS-AA	1.526	1.526	1.526	1.526	1.519	1.519	1.519	1.519
		ESP	1.530	1.530	1.530	1.530	1.528	1.528	1.528	1.528
CH ₃ –CH ₂ COO [–]	1.533	OPLS-AA	1.527	1.527	1.527	1.527	1.521	1.521	1.521	1.521
		ESP	1.531	1.531	1.531	1.531	1.529	1.529	1.529	1.529
CF ₃ –CH ₂ O [–]	1.521	OPLS-AA	1.617	1.616	n/a	n/a	1.677	1.676	n/a	n/a
		ESP	1.613	1.613	n/a	n/a	1.661	1.659	n/a	n/a
CH ₂ OH–CH ₂ O [–]	1.561	OPLS-AA	1.580	1.587	1.585	1.587	1.591	1.600	1.598	1.599
		ESP	1.587	1.594	1.592	1.593	1.605	1.614	1.613	1.612
CH ₂ OH–CH ₂ S [–]	1.525	OPLS-AA	1.533	1.535	1.533	1.537	1.539	1.541	1.539	1.543
		ESP	1.537	1.538	1.536	1.540	1.546	1.548	1.547	1.550
MUE ^e		OPLS-AA	0.014/0.021	0.015/0.022	0.014	0.015	0.020/0.033	0.021/0.034	0.021	0.021
		ESP	0.012/0.019	0.012/0.020	0.012	0.013	0.016/0.028 ^e	0.017/0.029	0.017	0.017
MSE ^f		OPLS-AA	–0.009/0.002	–0.008/0.003	–0.009	–0.007	–0.11/0.006	–0.010/0.007	–0.011	–0.010
		ESP	–0.004/0.006	–0.004/0.006	–0.004	–0.003	–0.003/0.012	–0.002/0.013	–0.002	–0.001

^a The QM level is HF/MIDI!, and the MM force field is OPLS-AA. The iterative procedure for polarization was initiated by using the OPLS-AA or ESP charges for the embedded-QM calculations. The mean unsigned error (MUE) and mean signed error (MSE) were averaged over the molecules for each QM/MM treatment using the full-QM calculations as standard reference values. ^b Reference 107. ^c QEq model with SCT of Rappé and Goddard. ^d QEq model of Bakowies and Thiel. ^e Mean unsigned error excluding/including CF₃–CH₂OH. ^f Mean signed error excluding/including CF₃–CH₂OH.

Table 4. Proton Affinities (kcal/mol) for Capped Amino Acids^a

molecule	QM	RC	PBRC			PBRCD ^b		PBRCD			PBRCD ^b
			SCT ^c	BT ^d	EEM	SCT ^c	RCD	SCT ^c	BT ^d	EEM	SCT ^c
Ace-Lys-NMe	236.7	234.6	235.7	235.7	235.2	240.3	233.7	234.6	234.4	234.3	238.8
Ace-His ^δ -NMe	251.9	247.3	254.6	242.2	254.5	271.8	246.1	253.2	244.0	252.9	270.5
Ace-His ^ε -NMe	254.7	245.6	251.4	250.5	250.9	269.1	244.3	249.9	248.4	249.8	267.8
Ace-Tyr-NMe	366.8	368.4	363.0	368.3	368.6	348.2	367.2	361.7	366.4	367.6	348.4
Ace-Glu-NMe	358.3	358.0	356.4	357.5	361.2	358.0	356.2	354.4	354.6	359.5	355.3
MUE ^e		3.5	2.5	3.4	2.5	11.4	4.3	3.4	4.1	2.1	11.0
MSE ^f		–2.9	–1.5	–2.8	0.4	3.8	–4.2	–2.9	–4.1	–0.9	2.5

^a The side chain is the PS, and the backbone is the SS (see also Figure 1). The QM level is set to HF/MIDI!, and the MM force field is OPLS-AA. The iterative procedure for polarization was initiated by using the OPLS-AA charges for the embedded-QM calculations. The mean unsigned error (MUE) and the mean signed error (MSE) were averaged over the molecules for each QM/MM treatment using the full-QM calculations as standard reference values. ^b This model allows all SS atoms to change charges. ^c QEq model with SCT of Rappé and Goddard. ^d QEq model of Bakowies and Thiel. ^e Mean unsigned error. ^f Mean signed error.

the full-QM calculations. Bigger discrepancies are seen for the PS atoms close to the QM/MM boundary than for the PS atoms distant from the boundary. But the QM/MM and full-QM calculations yield very different electronic structures for the PS atoms that are close to the QM/MM boundary.

Turning to the SS atoms, we found that in the PBRC calculations the M2 and M3 charges depend strongly on the method of polarization treatment as well as on the parameters used. The BT treatment tends to yield larger charges than the SCT and EEM do. The SCT and EEM schemes gave

Table 5. QM/MM Optimized Q1–M1 Bond Distances (Å) in Comparison with Full-QM Results for Amino Acids^a

molecule	QM	RC	PBRC			RCD	PBRCD		
			SCT ^b	BT ^c	EEM		SCT ^b	BT ^c	EEM
XH or XH ⁺									
Ace-Lys-NMe	1.533	1.542	1.541	1.539	1.540	1.544	1.543	1.541	1.541
Ace-His-NMe	1.558	1.550	1.548	1.546	1.548	1.552	1.550	1.548	1.549
Ace-Tyr-NMe	1.561	1.546	1.547	1.543	1.545	1.549	1.549	1.546	1.548
Ace-Glu-NMe	1.551	1.541	1.540	1.538	1.539	1.543	1.543	1.540	1.541
X ⁻ or X									
Ace-Lys-NMe	1.531	1.543	1.543	1.541	1.542	1.546	1.545	1.544	1.544
Ace-His ^δ -NMe	1.537	1.546	1.545	1.553	1.543	1.549	1.548	1.553	1.543
Ace-His ^ε -NMe	1.559	1.551	1.551	1.549	1.550	1.554	1.554	1.552	1.552
Ace-Tyr-NMe	1.569	1.555	1.556	1.553	1.554	1.560	1.561	1.558	1.558
Ace-Glu-NMe	1.550	1.548	1.549	1.546	1.547	1.551	1.553	1.551	1.550
MUE ^d		0.010	0.009	0.012	0.010	0.009	0.009	0.010	0.009
MSE ^e		-0.003	-0.003	-0.005	-0.005	-0.001	-0.003	-0.002	-0.003

^a The side chain is the PS, and the backbone is the SS (see also Figure 1). The QM level is HF/MIDI!, and the MM force field is OPLS-AA. The iterative procedure for polarization was initiated by using the OPLS-AA charges for the embedded-QM calculations. If not otherwise indicated, only the M2 and M3 atoms in the SS were allowed to change charges. The mean unsigned error (MUE) and the mean signed error (MSE) were averaged over the molecules for each QM/MM treatment using the full-QM calculations as standard reference values. ^b QEq model with SCT of Rappé and Goddard. ^c QEq model of Bakowies and Thiel. ^d Mean unsigned error. ^e Mean signed error.

Table 6. Proton Affinities (kcal/mol) for Amino Acids at the Full-QM Optimized Geometries^a

molecule	QM	RC	PBRC			RCD	PBRCD		
			SCT ^b	BT ^c	EEM		SCT ^b	BT ^c	EEM
Ace-Lys-NMe	236.7	234.0	234.9	235.0	234.1	233.1	233.9	233.7	233.2
Ace-His ^δ -NMe	251.9	241.3	247.6	243.1	246.7	240.1	246.2	240.9	245.6
Ace-His ^ε -NMe	254.7	251.2	256.3	253.6	254.4	250.0	254.9	251.7	253.3
Ace-Tyr-NMe	366.8	368.6	364.0	367.8	368.8	367.3	362.7	365.8	367.7
Ace-Glu-NMe	358.3	363.7	362.1	362.6	366.6	362.0	360.1	359.8	364.9
MUE		4.8	2.9	3.4	3.7	4.9	2.9	3.9	3.7
MSE		-1.9	-0.7	-1.3	0.4	-3.2	-2.1	-3.3	-0.7

^a Single-point QM/MM calculations were carried out at the geometries optimized by full QM. The side chain is the PS, and the backbone is the SS (see also Figure 1). The QM level is HF/MIDI!, and the MM force field is OPLS-AA. The iterative procedure for polarization was initiated by using the OPLS-AA charge parameters as partial atomic charges of SS atoms in the embedded-QM calculations. The mean unsigned error (MUE) and the mean signed error (MSE) were averaged over the molecules for each QM/MM treatment using the full-QM calculations as standard reference data. ^b QEq model with SCT of Rappé and Goddard. ^c QEq model of Bakowies and Thiel.

Table 7. Overall Mean Unsigned Errors (kcal/mol) for Organic Molecules and Capped Amino Acids^a

	PBRC				PBRCD			
	RC	SCT	BT	EEM	RCD	SCT	BT	EEM
HF/MIDI! ^a	4.9	2.9	3.3	3.6	5.4	4.1	4.4	3.8
B3LYP/6-31+G ^{*a,b}	4.4	2.1	2.9	3.2	4.5	2.9	3.5	3.3
MP2/6-31+G ^{*a,c}	4.4	2.2	3.0	3.5	4.5	2.7	3.4	3.4
overall ^d	4.6	2.4	3.1	3.4	4.8	3.2	3.8	3.5

^a First the results are averaged over the two choices of unpolarized partial atomic SS charges for the six small organic molecules (excluding 1,1,1-trifluoroethanol) and averaged over the five capped amino acids for the two sets of geometries. Then the results for the small organic molecules and the capped amino acids were each weighted 0.5 for a final average, which is given in the table. ^b See Table S9 in the Supporting Information for proton affinities for small organic molecules and Tables S13 and S15 for proton affinities for capped amino acids. ^c See Table S10 in the Supporting Information for proton affinities for small organic molecules and Tables S16 and S17 for proton affinities for capped amino acids. ^d Average over all three QM levels.

quite similar charges. For example, the charge on the N7 atom was predicted to be -0.41e in the SCT treatment and -0.44 e in the EEM treatment but was predicted to be -0.94 e in the BT treatment. The large charge of -0.94 e seems

unrealistic and may result in an overestimation of the electrostatic interaction between the N7 atom and other atoms.

Although the charges on the SS atoms produced in the PBRC calculations by all the three polarization treatments are in qualitative agreement with the full-QM ESP charges, it is difficult to tell which set of QM/MM charges are more realistic just by comparing them with the full-QM charges. As discussed before, the SS charges enter the QM Hamiltonian of the CPS as one-electron operators. Thus the SS charges are parameters of the effective QM Hamiltonian, and they are not strictly comparable with the full-QM ESP charges or the partial charges in the MM force field, although all the QM (ESP), MM, and QM/MM charges describe the interatomic interactions in their own theoretical frameworks. Which set of parameters are the best parameters depends on the problem one is studying and can only be answered after comparing the calculated molecular properties (e.g., energies and geometries) with standard reference data such as reliable experimental results or highly accurate theoretical calculations.

V.C. Energies and Geometries of Small Organic Molecules. For the optimized Q1–M1 bond distances of the

small organic molecules, including protonated and deprotonated forms (see Table 3), the RC and PBRC calculations yielded almost identical results for the first five molecules. For the last two molecules, $\text{CH}_2\text{OH}-\text{CH}_2\text{OH}$ and $\text{CH}_2\text{OH}-\text{CH}_2\text{SH}$, the differences between the RC and PBRC results are slightly larger: up to 0.004 Å for the neutral species and up to 0.007 Å for the charged species. In comparison with their neutral partners, the charged species were affected more, as expected, by the stronger polarization of the SS due to the charged PS. Different polarization treatments (SCT, BT, and EEM) in the PBRC computations showed negligible differences (typically no more than 0.002 Å) between each other for the Q1–M1 distances. The comparisons between the RCD and PBRCD calculations lead to the same conclusions.

Overall, the polarization of the SS due to the CPS does not change the QM/MM optimized Q1–M1 bond distances significantly. Indeed, the MUE and MSE of Q1–M1 bond distances for the RC (or RCD) calculations are almost the same as those for the PBRC (or PBRCD) calculations with differences of only 0.002 Å or less. Again, we found that the polarization procedure initiated with the ESP charges provided better results for the Q1–M1 bond distances than the polarization procedure initiated with the OPLS-AA charges, just as we saw in the proton affinities.

As indicated in Table 2, the PBRC calculations for the small organic molecules yield noticeable overall improvements in the proton affinities, for which the MUEs (and MSEs) are about 25–60% smaller than the MUE (and MSE) yielded by the RC calculations. The performances by the SCT, BT, and EEM schemes are rather similar, with the MUEs (and MSEs) usually agreeing within 1 kcal/mol. The same trend emerges from the comparisons between the PBRCD and RCD calculations.

The first five molecules in the test suite have no M3 atoms in the SS, and thus for those molecules there is not much freedom for charge variation among the atoms during the polarization procedure. The PBRC and RC proton affinities are very similar to each other with differences usually less than 1 kcal/mol, as are the PBRCD and RCD results. For the last two molecules, the presence of both the M2 and M3 atoms in the SS allows more variational freedom during charge equilibration, and one observes more pronounced improvements in the range of 3–9 kcal/mol.

Interestingly, the polarization treatments initiated with the ESP charges produced better results (with MUEs and MSEs typically 2 kcal/mol smaller) than those initiated with the OPLA-AA charges. The use of the ESP charges, which are generally smaller than the OPLS-AA charges, imposes a smaller Q_{tot} in the charge conservation constraint (eq 12) for the boundary group. The observation of better performance by the ESP charges in this study is consistent with our previous conclusion¹⁰⁷ that the OPLS-AA charges or other charges (e.g., CHARMM or AMBER charge parameters) designed for use in condensed-phase simulations are not very suitable for QM/MM calculations of small molecules in the gas phase.

In order to put the results in perspective it is useful to compare the results in Table 2 to those in ref 107, which

considered the same seven proton affinities (including $\text{CF}_3-\text{CH}_2\text{OH}$). Reference 107 contains results for six treatments not included in Table 2, and they yielded MUEs of 7.1, 5.5, 6.2, 27.4, 14.5, and 8.2 kcal/mol for these seven proton affinities, which may be compared to 3.6 kcal/mol for the PBRC-ESP method and 3.8 kcal/mol for the PBRCD-ESP method. This shows that these new methods are relatively successful, considering the difficulty of the tests. The encouraging improvement demonstrated by the PBRC and PBRCD schemes in comparison with the RC and RCD schemes and the other treatments of ref 107 does not mean that one can use the polarized-embedding schemes without care. The polarized-embedding schemes must be used with caution, as there is no guarantee that the PBRC or PBRCD schemes will give results superior to those by the RC and RCD schemes for all systems. The results depend on the adequacy of the electronegativity equalization scheme and its parameters. Thus, it is recommended that users validate the schemes with a particular parameter set on similar small model systems before applying them to the systems of ultimate interest. In section V.D, we will consider such validation for capped amino acids.

V.D. Energies and Geometries of Capped Amino Acids.

First, we note that, as revealed in Table 4, poor results were produced for proton affinities when all SS atoms were put into a polarizable group and were permitted to vary their charges in the self-consistent polarization procedure. Such an arrangement of the SS atoms in the polarization treatment yielded MUEs of 11 kcal/mol (in both the PBRC and PBRCD calculations); such MUEs are about 3 times as large as the MUEs when only the M2 and M3 atoms were put into the polarizable SS group. Therefore, the scheme of modifying the charges of all SS atoms will not be considered further, and our discussion in the rest of this work will concentrate on the PBRC and PBRCD calculations where only the M2 and M3 atoms were allowed to change their charges.

Table 4 shows that the PBRC and PBRCD calculations using the SCT and EEM parametrizations yield MUEs that are 1–2 kcal/mol smaller than for the RC and RCD calculations, but the PBRC and PBRCD calculations employing the BT parametrization offer only a tiny improvement (0.1 kcal/mol) over the RC and RCD method. Among the three polarization options, the EEM option works best with an MUE of only 2.5 kcal/mol and an MSE of 0.4 kcal/mol in the PBRC calculations and an MUE of 2.1 kcal/mol and an MSE of –0.9 kcal/mol in the PBRCD calculations. The less satisfactory performance of the BT scheme as compared to the SCT and EEM schemes is largely due to the big discrepancies for the histidine deprotonation at the delta position. The proton affinities of Ace-His^δ-NME calculated by the BT treatment are lower than those calculated by the SCT and EEM treatments by ca. 12 kcal/mol in the PBRC calculations and by ca. 9 kcal/mol in the PBRCD calculations. This is quite unusual, as in the other cases, the SCT, BT, and EEM calculations yield rather similar proton affinities (always <6 kcal/mol and mostly <2 kcal/mol). In particular, the BT proton affinities agree with the SCT and EEM results within 2 kcal/mol for the

Ace-His^ε-NMe system, i.e., the histidine deprotonation at the epsilon position, a closely related system.

The large discrepancies in the proton affinities at the delta position for histidine between BT and the other two (SCT and EEM) polarization treatments are due to substantial differences in their optimized geometries. We found that upon the deprotonation at the delta position, the imidazole group bends toward the backbone due to the interaction between the negatively charged N15 atom in the imidazole group and the positively charged H11 atom in the backbone. The BT treatment produces much larger charges on these atoms in Ace-His^δ-NMe than the BT and EEM treatment does. For example, in PBRC calculations, BT yields $Q_{H11} = 0.665$ e and $Q_{N15} = -0.776$ e, while SCT gives $Q_{H11} = 0.269$ e and $Q_{N15} = -0.594$ e, and EEM predicts $Q_{H11} = 0.221$ e and $Q_{N15} = -0.583$ e. The charges in the PBRC calculations are similar to the charges in the PBRC calculations. The larger charges lead to stronger interactions between N15 and H11. Consequently, the imidazole group is more significantly bent toward the backbone in the BT treatment than in the SCT and EEM treatments. In the BT calculations, the distances between the H11 and N15 are about 1.8 Å in the PBRC calculations and 1.7 Å in the PBRC calculations, respectively. These distances are much shorter than those given by the SCT and EEM treatments, which were 2.5 Å or longer. The significant differences in geometries contribute to the difference in proton affinities.

The story is different for the histidine deprotonation at the epsilon position. The epsilon position is far away from the backbone atoms, and the N18 atom does not interact closely with the backbone atoms. Moreover, the charge on the H21 atom that is bonded to the N15 atom is also positive; the H21 atom is present in the Ace-His^ε-NMe and the Ace-His⁺-NMe systems but absent in the Ace-His^δ-NMe system. The repulsion between H21 and H11 in the Ace-His^ε-NMe model did not induce the same geometric changes (the bending of the imidazole group) as one observed in the Ace-His^δ-NMe model. Actually the distance between H11 and H21 barely change upon deprotonation at the epsilon position: they increased from about 3.9–4.0 Å in Ace-His⁺-NMe to about 4.1–4.2 Å in the Ace-His^ε-NMe. As expected, the conformations for all three polarization treatments are very close to each other, and the proton affinities were calculated to be quite similar.

The above explanations are confirmed by the QM/MM//full-QM proton affinities given in Table 6. The use of the same full-QM geometries eliminates the discrepancies due to different geometries, and we indeed found relatively small (<5 kcal/mol) variations between proton affinities calculated by all three treatments. In comparison with the QM/MM proton affinities, the overall variations in the QM/MM//full-QM proton affinities are reduced by about 50%.

For the QM/MM optimized Q1–M1 bond distances, as shown in Table 5, the performances of all three polarization treatments resemble their performance for proton affinities. They all give MUEs of about 0.01 Å and MSEs no larger than 0.005 Å, which are acceptably small for a variety of applications.

V.E. Other QM Methods. Additional calculations using the B3LYP/6-31+G* and MP2/6-31+G* levels of theory are given in the Supporting Information Tables S8–S19.

Comparing the proton affinities for small organic molecules listed in Table 2 (QM = HF/MIDI!), in Table S9 (QM = B3LYP/6-31+G*), and in Table S11 (QM = MP2/6-31+G*), there are no large changes in the accuracy in energy when different QM methods are employed for a given (polarizable) boundary treatment. Generally speaking, B3LYP/6-31+G* produces the best results, and HF/MIDI! shows the largest errors. Inspection of the Q1–M1 bond length in Tables 3, S10, and S12 gives similar conclusions. Turn to the capped amino acids, all three QM levels gave quite comparable results for proton affinities (Tables S13 and S15 for QM = DFT and Tables S16 and S17 for QM = MP2 calculations), and it is hard to tell which is superior to the others. Thus, all these additional calculations support our conclusions: Inclusion of the mutual polarization yields improves the QM/MM proton affinities for all the three QM levels in our test, and the extents of improvement are rather similar for all three levels. Inclusion of the mutual polarization leads to little change in the geometry. Interestingly, the ESP charges seem to always perform better than OPLS-AA charges for small organic compound.

V.F. Overall Assessment and Future Work. Table 7 presents an overall energetic assessment of the performance of the new polarized-boundary methods and the unpolarized boundaries from which they were evolved. This table is based on the proton affinities for the 11 cases (six organic molecules and five capped amino acids) where parameters are available for all methods examined. In general, in the ultimate applications, if some atoms close to the QM/MM boundary are found to undergo significant polarization effects, they should, if possible, be included into the QM subsystems; however, for testing a method to see if it is robust, it is instructive to push the envelope. That is why we included a few difficult cases in our test suite; for example, in CF₃–CH₂OH the bond breaking is occurring at a place very close to the QM/MM boundary, and MM atoms carrying large partial charges are located near the boundary. This particular molecule is included in Tables 2 and 3 but not in Table 7. Nevertheless, we see that the mean unsigned error of the most successful method without boundary polarization is ca. 5 kcal/mol, but that this is reduced to ca. 3 kcal/mol by four of the six methods with boundary polarization. On one hand, this shows the importance of boundary polarization, and the reduction of the error by 40% is encouraging. On the other hand, even though these are very challenging tests (adding a whole charge very close to the boundary), one might have hoped that the mean unsigned error would be reduced even further. One remaining source of error is that the present treatment assumes no charge transfer between the PS and the SS, even though the PS and SS are covalently bonded.

In principle, the interactions between the PS and SS can be modeled more realistically if one allows fractional (or whole) charges to be transferred between the PS and SS. Such a treatment could be called flexible-boundary QM/MM (FB-QM/MM, FBRC, or FBRC). For FB-QM/MM calcu-

lations one needs an algorithm that describes the electronic structure of a system with fractional electrons and provides a prescription for how much charge should be transferred. Gogonea and Merz^{193,194} have proposed a combined quantum mechanical-Poisson-Boltzmann equation approach to study the charge transfer between ions and a solvent medium treated as a dielectric continuum. In their treatment, the charge being transferred is represented by a surface charge density at the dielectric interface, which modifies the boundary condition for which the Poisson-Boltzmann equation is solved. The ions are described by an effective QM Hamiltonian that resembles Dewar's half-electron method^{195,196} but with subtle differences in handling the electron-electron repulsion term. The self-consistent QM calculations are carried out in terms of the density matrix by adding electron density to the LUMO (in the case of charge transferred to ions) or by subtracting electron density from the HOMO (in the case of charge transferred to solvent). The amount of charge being transferred is determined variationally subject to the criterion of the free energy including the environment. Sprik and co-workers¹⁹⁷ proposed another scheme that can potentially be used to handle fractional charge transfer between the PS and SS in the QM/MM calculations. They model the exchange of electrons between molecule and a reservoir of fixed chemical potential by a modification of the Car-Parrinello¹⁹⁸ method allowing for fluctuating numbers of electrons under constraints of fixed electronic chemical potential. They adopted an approach involving multiple diabatic potential energy surfaces where each surface corresponds to a system with a strictly integer number of electrons, e.g., a surface for the reduced state whose charge is 0 and a surface for the oxidized state whose charge is +1 e. Thermochemical properties in a molecular dynamics run were computed by a weighted average of the partition functions for the two oxidation states; in other words, one avoids treating a fractional number of electrons by moving the system on an effective (adiabatic) potential that is a weighted average of diabatic potential surfaces corresponding to integer numbers of electrons. The weights are determined by the chemical potential and the mole fraction of the cations. This provides a more justifiable treatment of electron exchange, but it has been criticized because of the need for a uniform background charge.¹⁹⁹ Further research on FB-QM/MM would be valuable.

VI. Summary and Conclusions

In this work, we developed two polarized boundary embedding schemes, the PBRC and PBRCD schemes, to allow a more accurate treatment of QM/MM boundaries. The newly developed schemes combine the electrostatic-embedding RC and RCD schemes,¹⁰⁷ where the CPS is polarized by the SS, with an electronegativity-equalization or charge-equilibration scheme to describe the polarization of the boundary region of the SS by the CPS. More specifically, in the PBRC and PBRCD schemes, the polarization of the SS by the CPS is realized by adjusting the point charges at the SS atomic sites in the embedded-QM calculations. The calculations were carried out with a new version of the QMMM computer program¹⁸⁰ that is general enough to handle cases where the

QM-MM boundary passes between molecules (as in the earlier work of Field²⁷) or when it passes through a covalent bond. For either type of boundary, the QMMM computer program is also general enough to polarize only one group of the SS, to polarize two or more defined groups, or to polarize the whole SS as a single group. We focus in this paper though on the PBRC and PBRCD schemes where the boundary passes through a covalent bond and only one group, in particular the boundary group, of the SS is polarized.

In this work, the implementation polarization of the SS by the CPS is based on the principle of electronegativity equalization or charge equilibration. The advantage of this scheme is that it is simple and easy to implement. Moreover, the polarization effects expressed in the charge redistribution are easy to interpret. Although variation of atomic charges does not account for all polarization effects,¹⁴⁰ in most applications, the variation of charges based on electronegativity equalization is probably adequate to account for the dominant polarization effect on the boundary region of the SS.

For the determination of the charges on the SS atoms, we implemented both the QEq¹⁵³ and EEM¹⁵¹ methods with modifications to take into account the external electric field generated by the CPS and by the unpolarized part of the SS. In the QEq calculations we employed the empirical functions and parameter sets of both Rappé and Goddard¹⁵³ and Bakowies and Thiel¹⁷ to compute interatomic electrostatic interactions. Self-consistency in the mutual polarization of the CPS and the SS is accomplished by an iterative procedure; usually convergence is achieved within three sets of iterations.

If there is no significant charge transfer involved, due to the cancellation of errors in the reactant state and the errors in the product state, good energetics might be obtained by using the electrostatic-embedding schemes where only the CPS is polarized by the SS or even by using the mechanical-embedding scheme where the electrostatic interactions between the CPS and the SS are handled at the MM level. The electronic structures of the CPS will be different in the polarized-embedding schemes, in the electrostatic-embedding schemes, and in the mechanical-embedding schemes. However, as indicated by the ESP charges for the CPS in section V.B., the difference in the electronic structures of the CPS is often small between the polarized-embedding calculations and the electrostatic-embedding calculations. Therefore, in many applications, it is probably sufficient to use the electrostatic-embedding methods. However, the mutual polarization of the CPS and SS is expected to be important in situations where significant charge-transfer takes place in the CPS during a reaction, e.g., the protonation reactions we investigated in the present study. Therefore the PBRC and PBRCD schemes were tested by calculating proton affinities for small organic molecules and capped amino acids. The proton affinity is a critical test for QM/MM methods because of the significant changes in the charges of the CPS upon protonation; thus proton affinities are very sensitive to the treatment of electrostatic interactions between the CPS and SS and are likely to show prominent polarization effects in the SS. We found that there is no significant difference in

the Q1–M1 bond distances between the calculations with and without polarization of the boundary region of the SS, but encouraging improvement in the computed proton affinities was obtained by the new methods in comparison with the errors of RC and RCD calculations that did not consider the polarization of the SS by the CPS (and also in comparison with even larger errors obtained in other treatments considered in a previous paper). These findings suggest the importance of the mutual polarization of the CPS and SS in QM/MM calculations where charge transfers occur in the CPS, and the success of the new polarization treatment implemented here in handling the polarization of the SS is gratifying.

Acknowledgment. This work is supported by the Research Corporation through the award no. CC6725 to H. Lin and by the Office of Naval Research under award no. N00014-05-1-0538 to D. G. Truhlar. We thank the Advanced Biomedical Computing Center and the Minnesota Supercomputing Institute for providing CPU time and access to the *Gaussian03* program.

Supporting Information Available: QEq and EEM parameters in the present study taken from the literature (Table S1), nonstandard MM parameters in this work in the TINKER parameter file format (Table S2), atom types, full-QM optimized coordinates, and connectivities of the capped amino acids in the test suite (in the tinker.xyz file format) (Table S3), atomic charges derived from the full-QM and selected QM/MM calculations for the capped amino acids (Table S4), convergence of the atomic charges of the heavy atoms in the PS and SS and of embedded-QM energies for the CPS in the PBRC calculations employing the SCT parameters for the Ace-His⁺-NMe system and its deprotonated forms (Table S5), comparison of the charges on H11, H21, and N15 and distances between H11 and H21 and between H11 and N15 for the Ace-His⁺-NMe system and its deprotonated forms (Table S6), atomic charges derived from the full-QM and selected QM/MM calculations for the capped amino acids at the full-QM optimized geometries (Table S7), and effect of changing the quantum mechanical level, as discussed in section V.E (Tables S8–S19). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.
- Bacalis, N. C.; Kunz, A. B. *Phys. Rev. B: Condens. Matter* **1985**, *32*, 4857.
- Singh, U. C.; Kollmann, P. A. *J. Comput. Chem.* **1986**, *7*, 718.
- Barandiaran, Z.; Seijo, L. *J. Chem. Phys.* **1988**, *89*, 5739.
- Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700.
- Ferenczy, G. G.; Rivail, J.-L.; Surjan, P. R.; Naray-Szabo, G. *J. Comput. Chem.* **1992**, *13*, 830.
- Gao, J.; Xia, X. *Science* **1992**, *258*, 631.
- Åqvist, J.; Warshel, A. *Chem. Rev.* **1993**, *93*, 2523.
- Thery, V.; Rinaldi, D.; Rivail, J.-L.; Maigret, B.; Ferenczy, G. G. *J. Comput. Chem.* **1994**, *15*, 269.
- Thompson, M. A.; Glendening, E. D.; Feller, D. *J. Phys. Chem.* **1994**, *98*, 10465.
- Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170.
- Stanton, R. V.; Hartsough, D. S.; Merz, K. M., Jr. *J. Comput. Chem.* **1995**, *16*, 113.
- Thompson, M. A. *J. Phys. Chem.* **1995**, *99*, 4794.
- Thompson, M. A.; Schenter, G. K. *J. Phys. Chem.* **1995**, *99*, 6374.
- Assfeld, X.; Rivail, J.-L. *Chem. Phys. Lett.* **1996**, *263*, 100.
- Bakowies, D.; Thiel, W. *J. Phys. Chem.* **1996**, *100*, 10580.
- Bakowies, D.; Thiel, W. *J. Comput. Chem.* **1996**, *17*, 87.
- Barnes, J. A.; Williams, I. H. *Biochem. Soc. Trans.* **1996**, *24*, 263.
- Day, P. N.; Jensen, J. H.; Gordon, M. S.; Webb, S. P.; Stevens, W. J.; Krauss, M.; Garmer, D.; Basch, H.; Cohen, D. *J. Chem. Phys.* **1996**, *105*, 1968.
- Eichler, U.; Kölmel, C. M.; Sauer, J. *J. Comput. Chem.* **1996**, *18*, 463.
- Eurenius, K. P.; Chatfield, D. C.; Brooks, B. R.; Hodoscek, M. *Int. J. Quantum Chem.* **1996**, *60*, 1189.
- Gao, J. *Rev. Comput. Chem.* **1996**, *7*, 119.
- Kerdcharoen, T.; Liedl, K. R.; Rode, B. M. *Chem. Phys.* **1996**, *211*, 313.
- Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357.
- Bersuker, I. B.; Leong, M. K.; Boggs, J. E.; Pearlman, R. S. *Int. J. Quantum Chem.* **1997**, *63*, 1051.
- Cummins, P. L.; Gready, J. E. *J. Comput. Chem.* **1997**, *18*, 1496.
- Field, M. J. *Mol. Phys.* **1997**, *91*, 835.
- Gao, J. *J. Comput. Chem.* **1997**, *18*, 1061.
- Gao, J.; Alhambra, C. *J. Chem. Phys.* **1997**, *107*, 1212.
- Pascual, J.; Pettersson, L. G. M. *Chem. Phys. Lett.* **1997**, *270*, 351.
- Sherwood, P.; De Vries, A. H.; Collins, S. J.; Greatbanks, S. P.; Burton, N. A.; Vincent, M. A.; Hillier, I. H. *Faraday Discuss.* **1997**, *106*, 79.
- Bryce, R. A.; Vincent, M. A.; Malcolm, N. O. J.; Hillier, I. H.; Burton, N. A. *J. Chem. Phys.* **1998**, *109*, 3077.
- Gao, J.; Amara, P.; Alhambra, C.; Field, M. J. *J. Phys. Chem. A* **1998**, *102*, 4714.
- Kaminski, G. A.; Jorgensen, W. L. *J. Phys. Chem. B* **1998**, *102*, 1787.
- Mordasini, T.; Thiel, W. *Chimia* **1998**, *52*, 288.
- Sinclair, P. E.; de Vries, A.; Sherwood, P.; Catlow, C. R. A.; van Santen, R. A. *J. Chem. Soc., Faraday Trans.* **1998**, *94*, 3401.
- Stefanovich, E. V.; Truong, T. N. *J. Phys. Chem. B* **1998**, *102*, 3018.
- Tongraar, A.; Liedl, K. R.; Rode, B. M. *J. Phys. Chem. A* **1998**, *102*, 10340.

- (39) Woo, T. K.; Cavallo, L.; Ziegler, T. *Theor. Chem. Acc.* **1998**, *100*, 307.
- (40) Antes, I.; Thiel, W. *J. Phys. Chem. A* **1999**, *103*, 9290.
- (41) Dapprich, S.; Komiroimi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *THEOCHEM* **1999**, 461–462, 1.
- (42) Eichinger, M.; Tavan, P.; Hutter, J.; Parrinello, M. *J. Chem. Phys.* **1999**, *110*, 10452.
- (43) Hillier, I. H. *THEOCHEM* **1999**, 463, 45.
- (44) Lyne, P. D.; Hodoscek, M.; Karplus, M. *J. Phys. Chem. A* **1999**, *103*, 3462.
- (45) Monard, G.; Merz, K. M., Jr. *Acc. Chem. Res.* **1999**, *32*, 904.
- (46) Philipp, D. M.; Friesner, R. A. *J. Comput. Chem.* **1999**, *20*, 1468.
- (47) Pitarch, J.; Pascual-Ahuir, J. L.; Silla, E.; Tunon, I.; Ruiz-Lopez, M. F. *J. Comput. Chem.* **1999**, *20*, 1401.
- (48) Shoemaker, J. R.; Burggraf, L. W.; Gordon, M. S. *J. Phys. Chem. A* **1999**, *103*, 3245.
- (49) Turner, A. J.; Moliner, V.; Williams, I. H. *Phys. Chem. Chem. Phys.* **1999**, *1*, 1323.
- (50) Zhang, Y.; Lee, T.-S.; Yang, W. *J. Chem. Phys.* **1999**, *110*, 46.
- (51) Amara, P.; Field, M. J.; Alhambra, C.; Gao, J. *Theor. Chem. Acc.* **2000**, *104*, 336.
- (52) Cui, Q.; Karplus, M. *J. Chem. Phys.* **2000**, *112*, 1133.
- (53) Cui, Q.; Karplus, M. *J. Phys. Chem. B* **2000**, *104*, 3721.
- (54) Derenzo, S. E.; Klintonberg, M. K.; Weber, M. J. *J. Chem. Phys.* **2000**, *112*, 2074.
- (55) Gogonea, V.; Westerhoff, L. M.; Merz, K. M., Jr. *J. Chem. Phys.* **2000**, *113*, 5604.
- (56) Hall, R. J.; Hindle, S. A.; Burton, N. A.; Hillier, I. H. *J. Comput. Chem.* **2000**, *21*, 1433.
- (57) Hammes-Schiffer, S. *Acc. Chem. Res.* **2000**, *34*, 273.
- (58) Kairys, V.; Jensen, J. H. *J. Phys. Chem. A* **2000**, *104*, 6656.
- (59) Luque, F. J.; Reuter, N.; Cartier, A.; Ruiz-López, M. F. *J. Phys. Chem. A* **2000**, *104*, 10923.
- (60) Murphy, R. B.; Philipp, D. M.; Friesner, R. A. *J. Comput. Chem.* **2000**, *21*, 1442.
- (61) Murphy, R. B.; Philipp, D. M.; Friesner, R. A. *Chem. Phys. Lett.* **2000**, *321*, 113.
- (62) Reuter, N.; Dejaegere, A.; Maignet, B.; Karplus, M. *J. Phys. Chem. A* **2000**, *104*, 1720.
- (63) Röthlisberger, U.; Carloni, P.; Doclo, K.; Parrinello, M. *J. Bio. Inorg. Chem.* **2000**, *5*, 236.
- (64) Sauer, J.; Sierka, M. *J. Comput. Chem.* **2000**, *21*, 1470.
- (65) Sherwood, P. In *Modern Methods and Algorithms of Quantum Chemistry*; Grotendorst, J., Ed.; NIC-Directors: Princeton, 2000; Vol. 3, pp 285.
- (66) Sierka, M.; Sauer, J. *J. Chem. Phys.* **2000**, *112*, 6983.
- (67) Sushko, P. V.; Shluger, A. L.; Catlow, C. R. A. *Surf. Sci.* **2000**, *450*, 153.
- (68) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 569.
- (69) Nasluzov, V. A.; Rivanenkov, V. V.; Gordienko, A. B.; Neyman, K. M.; Birkenheuer, U.; Rösch, N. *J. Chem. Phys.* **2001**, *115*, 8157.
- (70) Nicoll, R. M.; Hindle, S. A.; MacKenzie, G.; Hillier, I. H.; Burton, N. A. *Theor. Chem. Acc.* **2001**, *106*, 105.
- (71) Poteau, R.; Ortega, I.; Alary, F.; Solis, A. R.; Barthelat, J. C.; Daudey, J. P. *J. Phys. Chem. A* **2001**, *105*, 198.
- (72) Vreven, T.; Mennucci, B.; da Silva, C. O.; Morokuma, K.; Tomasi, J. *J. Chem. Phys.* **2001**, *115*, 62.
- (73) Batista, E. R.; Friesner, R. A. *J. Phys. Chem. B* **2002**, *106*, 8136.
- (74) Colombo, M. C.; Guidoni, L.; Laio, A.; Magistrato, A.; Maurer, P.; Piana, S.; Rohrig, U.; Spiegel, K.; Sulpizi, M.; VandeVondele, J.; Zumstein, M.; Röthlisberger, U. *Chimia* **2002**, *56*, 13.
- (75) Das, D.; Eurenium, K. P.; Billings, E. M.; Sherwood, P.; Chatfield, D. C.; Hodoscek, M.; Brooks, B. R. *J. Chem. Phys.* **2002**, *117*, 10534.
- (76) DiLabio, G. A.; Hurley, M. M.; Christiansen, P. A. *J. Chem. Phys.* **2002**, *116*, 9578.
- (77) Ferré, N.; Assfeld, X.; Rivail, J.-L. *J. Comput. Chem.* **2002**, *23*, 610.
- (78) Gao, J.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467.
- (79) Gogonea, V. *Internet Electron. J. Mol. Des.* **2002**, *1*, 173.
- (80) Kerdcharoen, T.; Morokuma, K. *Chem. Phys. Lett.* **2002**, 355, 257.
- (81) Laio, A.; VandeVondele, J.; Röthlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941.
- (82) Morokuma, K. *Philos. Trans. R. Soc. London, Ser. A* **2002**, *360*, 1149.
- (83) Schöneboom, J. C.; Lin, H.; Reuter, N.; Thiel, W.; Cohen, S.; Ogliaro, F.; Shaik, S. *J. Am. Chem. Soc.* **2002**, *124*, 8142.
- (84) Titmuss, S. J.; Cummins, P. L.; Rendell, A. P.; Bliznyuk, A. A.; Gready, J. E. *J. Comput. Chem.* **2002**, *23*, 1314.
- (85) Truhlar, D. G.; Gao, J.; Alhambra, C.; Garcia-Viloca, M.; Corchado, J.; Sanchez, M. L.; Villa, J. *Acc. Chem. Res.* **2002**, *35*, 341.
- (86) Amara, P.; Field, M. J. *Theor. Chem. Acc.* **2003**, *109*, 43.
- (87) Devi-Kesavan, L. S.; Garcia-Viloca, M.; Gao, J. *Theor. Chem. Acc.* **2003**, *109*, 133.
- (88) Dinner, A. R.; Lopez, X.; Karplus, M. *Theor. Chem. Acc.* **2003**, *109*, 118.
- (89) Hu, H.; Elstner, M.; Hermans, J. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 451.
- (90) Kongsted, J.; Osted, A.; Mikkelsen, K. V.; Christiansen, O. *J. Phys. Chem. A* **2003**, *107*, 2578.
- (91) Li, G.; Zhang, X.; Cui, Q. *J. Phys. Chem. B* **2003**, *107*, 8643.
- (92) Loferer, M. J.; Loeffler, H. H.; Liedl, K. R. *J. Comput. Chem.* **2003**, *24*, 1240.
- (93) Molina, P. A.; Sikorski, R. S.; Jensen, J. H. *Theor. Chem. Acc.* **2003**, *109*, 100.
- (94) Mordasini, T.; Curioni, A.; Andreoni, W. *J. Biol. Chem.* **2003**, *278*, 4381.

- (95) Sherwood, P.; de Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; Billeter, S.; Terstegen, F.; Thiel, S.; Kendrick, J.; Rogers, S. C.; Casci, J.; Watson, M.; King, F.; Karlsen, E.; Sjøvoll, M.; Fahmi, A.; Schafer, A.; Lennartz, C. *THEOCHEM* **2003**, 632, 1.
- (96) Sulpizi, M.; Laio, A.; VandeVondele, J.; Cattaneo, A.; Röthlisberger, U.; Carloni, P. *Proteins: Struct., Funct., Genet.* **2003**, 52, 212.
- (97) Swart, M. *Int. J. Quantum Chem.* **2003**, 91, 177.
- (98) Tresadern, G.; Faulder, P. F.; Gleeson, M. P.; Tai, Z.; MacKenzie, G.; Burton, N. A.; Hillier, I. H. *Theor. Chem. Acc.* **2003**, 109, 108.
- (99) Vreven, T.; Morokuma, K.; Farkas, O.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, 24, 760.
- (100) Yang, W.; Drueckhammer, D. G. *J. Phys. Chem. B* **2003**, 107, 5986.
- (101) Laio, A.; Gervasio, F. L.; VandeVondele, J.; Sulpizi, M.; Röthlisberger, U. *J. Phys. Chem. B* **2004**, 108, 7963.
- (102) Pu, J.; Gao, J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, 108, 632.
- (103) Pu, J.; Gao, J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, 108, 5454.
- (104) Riccardi, D.; Li, G. H.; Cui, Q. *J. Phys. Chem. B* **2004**, 108, 6467.
- (105) Gregersen, B. A.; York, D. M. *J. Phys. Chem. B* **2005**, 109, 536.
- (106) König, P. H.; Hoffmann, M.; Frauenheim, T.; Cui, Q. *J. Phys. Chem. B* **2005**, 109, 9082.
- (107) Lin, H.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, 109, 3991.
- (108) Nam, K.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2005**, 1, 2.
- (109) Pu, J.; Gao, J.; Truhlar, D. G. *ChemPhysChem* **2005**, 6, 1853.
- (110) Wanko, M.; Hoffmann, M.; Strodel, P.; Koslowski, A.; Thiel, W.; Neese, F.; Frauenheim, T.; Elstner, M. *J. Phys. Chem. B* **2005**, 109, 3606.
- (111) Zhang, Y. *J. Chem. Phys.* **2005**, 122, 024114.
- (112) Cisneros, G. A.; Piquemal, J.-P.; Darden, T. A. *J. Phys. Chem. B* **2006**, 110, 13682.
- (113) Fornili, A.; Loos, P.-F.; Sironi, M.; Assfeld, X. *Chem. Phys. Lett.* **2006**, 427, 236.
- (114) Illingworth, C. J. R.; Gooding, S. R.; Winn, P. J.; Jones, G. A.; Ferenczy, G. G.; Reynolds, C. A. *J. Phys. Chem. A* **2006**, 110, 6487.
- (115) Mallik, A.; Taylor, D. E.; Runge, K.; Dufty, J. W.; Cheng, H.-P. *J. Comput.-Aided Mater. Des.* **2006**, 13, 45.
- (116) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; König, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, 110, 6458.
- (117) (a) Vreven, T.; Frisch, M. J.; Kudin, K. N.; Schlegel, H. B.; Morokuma, K. *Mol. Phys.* **2006**, 104, 701. (b) Vreven, T.; Byun, K. S.; Komromi, I.; Dapprich, S.; Montgomery, J. A., Jr.; Morokuma, K.; Frisch, M. J. *J. Chem. Theory Comput.* **2006**, 2, 815.
- (118) (a) Raff, L. M. *J. Chem. Phys.* **1974**, 60, 2220. (b) Joseph, T.; Steckler, R.; Truhlar, D. G. *J. Chem. Phys.* **1987**, 87, 7036. (c) Chakraborty, A.; Zhao, Y.; Lin, H.; Truhlar, D. G. *J. Chem. Phys.* **2006**, 124, 44315.
- (119) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, 117, 185.
- (120) Heyden, A.; Lin, H.; Truhlar, D. G. *J. Phys. Chem. B* **2007**, 111, 2231.
- (121) 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. *J. Am. Chem. Soc.* **1995**, 117, 5179.
- (122) 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., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, 102, 3586.
- (123) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, 118, 11225.
- (124) Jorgensen, W. L.; McDonald, N. A. *THEOCHEM* **1998**, 424, 145.
- (125) McDonald, N. A.; Jorgensen, W. L. *J. Phys. Chem. B* **1998**, 102, 8049.
- (126) Rizzo, R. C.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1999**, 121, 4827.
- (127) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, 105, 6474.
- (128) Kahn, K.; Bruice, T. C. *J. Comput. Chem.* **2002**, 23, 977.
- (129) 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. *Gaussian03*; Gaussian, Inc.: Pittsburgh, PA, 2003.
- (130) Neese, F. *ORCA, Version 2.4 ed.*; Max Planck Institute for Bioinorganic Chemistry: Muelheim/Ruhr, 2006.
- (131) Price, S. L.; Stone, A. J. *J. Chem. Soc., Faraday Trans.* **1992**, 88, 1755.
- (132) Koch, U.; Popelier, P. L. A.; Stone, A. J. *Chem. Phys. Lett.* **1995**, 238, 253.
- (133) Matta, C. F.; Bader, R. F. W. *Proteins: Struct., Funct., Genet.* **2000**, 40, 310.
- (134) Minikis, R. M.; Kairys, V.; Jensen, J. H. *J. Phys. Chem. A* **2001**, 105, 3829.
- (135) Dick, B. G., Jr.; Overhauser, A. W. *Phys. Rev.* **1958**, 112, 90.
- (136) Stillinger, F. H.; David, C. W. *J. Chem. Phys.* **1978**, 69, 1473.

- (137) Sprik, M.; Klein, M. L. *J. Chem. Phys.* **1988**, *89*, 7556.
- (138) Dang, L. X.; Rice, J. E.; Caldwell, J.; Kollman, P. A. *J. Am. Chem. Soc.* **1991**, *113*, 2481.
- (139) Dang, L. X. *J. Chem. Phys.* **1992**, *97*, 2659.
- (140) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141.
- (141) Kowall, T.; Foglia, F.; Helm, L.; Merbach, A. E. *J. Am. Chem. Soc.* **1995**, *117*, 3790.
- (142) Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **1999**, *110*, 741.
- (143) Stern, H. A.; Kaminski, G. A.; Banks, J. L.; Zhou, R.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. B* **1999**, *103*, 4730.
- (144) Martinez, J. M.; Hernandez-Cobos, J.; Saint-Martin, H.; Pappalardo, R. R.; Ortega-Blake, I.; Marcos, E. S. *J. Chem. Phys.* **2000**, *112*, 2339.
- (145) Rick, S. W.; Stuart, S. J. *Rev. Comput. Chem.* **2002**, *18*, 89.
- (146) Lamoureux, G.; MacKerell, A. D.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185.
- (147) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27.
- (148) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621.
- (149) Patel, S.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1.
- (150) Patel, S.; Mackerell, A. D., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1504.
- (151) Mortier, W. J.; Van Genechten, K.; Gasteiger, J. *J. Am. Chem. Soc.* **1985**, *107*, 829.
- (152) Mortier, W. J.; Ghosh, S. K.; Shankar, S. *J. Am. Chem. Soc.* **1986**, *108*, 4315.
- (153) Rappé, A. K.; Goddard, W. A. *J. Phys. Chem.* **1991**, *95*, 3358.
- (154) Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Profit, F.; Geerings, P.; Waroquier, M.; Tollenaere, J. P. *J. Phys. Chem. A* **2002**, *106*, 7887.
- (155) York, D. M.; Yang, W. *J. Chem. Phys.* **1996**, *104*, 159.
- (156) Itskowitz, P.; Berkowitz, M. L. *J. Phys. Chem. A* **1997**, *101*, 5687.
- (157) Yang, Z.-Z.; Wang, C.-S. *J. Phys. Chem. A* **1997**, *101*, 6315.
- (158) Applequist, J.; Carl, J. R.; Fung, K.-K. *J. Am. Chem. Soc.* **1972**, *94*, 2952.
- (159) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341.
- (160) Stone, A. J. *Mol. Phys.* **1985**, *56*, 1065.
- (161) Winn, P. J.; Ferenczy, G. G.; Reynolds, C. A. *J. Comput. Chem.* **1999**, *20*, 704.
- (162) Rinaldi, D.; Rivail, J. L. *Theor. Chim. Acta* **1973**, *32*, 57.
- (163) Tapia, O.; Goscinski, O. *Mol. Phys.* **1975**, *29*, 1653.
- (164) Cramer, C. J.; Truhlar, D. G. Continuum solvation models. In *Solvents Effects and Chemical Reactivity*; Tapia, O., Bertran, J., Eds.; Kluwer: Dordrecht, 1996; p 1.
- (165) Ferenczy, G. G.; Reynolds, C. A. *J. Phys. Chem. A* **2001**, *105*, 11470.
- (166) French, S. A.; Sokol, A. A.; Bromley, S. T.; Catlow, C. R. A.; Rogers, S. C.; King, F.; Sherwood, P. *Angew. Chem.* **2001**, *113*, 4569.
- (167) Khaliullin, R. Z.; Bell, A. T.; Kazansky, V. B. *J. Phys. Chem. A* **2001**, *105*, 10454.
- (168) Herschend, B.; Baudin, M.; Hermansson, K. *J. Chem. Phys.* **2004**, *120*, 4939.
- (169) Bludsky, O.; Silhan, M.; Nachtigall, P.; Bucko, T.; Benco, L.; Hafner, J. *J. Phys. Chem. B* **2005**, *109*, 9631.
- (170) Monard, G.; Loos, M.; Thery, V.; Baka, K.; Rivail, J.-L. *Int. J. Quantum Chem.* **1996**, *58*, 153.
- (171) de Vries, A. H.; Sherwood, P.; Collins, S. J.; Rigby, A. M.; Rigutto, M.; Kramer, G. J. *J. Phys. Chem. B* **1999**, *103*, 6133.
- (172) Klopman, G. *J. Am. Chem. Soc.* **1964**, *86*, 4550.
- (173) Ohno, K. *Theor. Chim. Acta* **1964**, *2*, 219.
- (174) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- (175) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.
- (176) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8551.
- (177) Lii, J. H.; Allinger, N. L. *J. Am. Chem. Soc.* **1989**, *111*, 8566.
- (178) Lii, J. H.; Allinger, N. L. *J. Am. Chem. Soc.* **1989**, *111*, 8576.
- (179) Ponder, J. W. *TINKER, Version 4.2 ed.*; Washington University: St. Louis, MO, 2004.
- (180) Lin, H.; Zhang, Y.; Truhlar, D. G. *QMMM, Version 1.3 ed.*; University of Minnesota: Minneapolis, 2007.
- (181) Roothaan, C. C. J. *Rev. Mod. Phys.* **1951**, *23*, 69.
- (182) Easton, R. E.; Giesen, D. J.; Welch, A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **1996**, *93*, 281.
- (183) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1971**, *54*, 724.
- (184) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294.
- (185) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265.
- (186) Rassolov, V. A.; Ratner, M. A.; Pople, J. A.; Redfern, P. C.; Curtiss, L. A. *J. Comput. Chem.* **2001**, *22*, 976.
- (187) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter* **1988**, *37*, 785.
- (188) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (189) Møller, C. M. S.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.
- (190) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129.
- (191) Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431.
- (192) Siepmann, J. I. *Adv. Chem. Phys.* **1999**, *105*, 1.
- (193) Gogonea, V.; Merz, K. M., Jr. *J. Chem. Phys.* **2000**, *112*, 3227.
- (194) Gogonea, V.; Merz, K. M., Jr. *J. Phys. Chem. B* **2000**, *104*, 2117.
- (195) Dewar, M. J. S.; Hashmall, J. A.; Venier, C. G. *J. Am. Chem. Soc.* **1968**, *90*, 1953.

- (196) Dewar, M. J. S.; Trinajstić, N. *J. Chem. Soc., Chem. Commun.* **1970**, 646.
- (197) Tavernelli, I.; Vuilleumier, R.; Sprik, M. *Phys. Rev. Lett.* **2002**, 88, 213002.
- (198) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, 55, 2471.
- (199) Jaque, P.; Marenich, A.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. C* **2007**, asap article: 10.1021/jp066765w; Web released on 28 Mar. 2007.

CT7000107

What Is the Limit of Atom Encapsulation for Icosahedral Carboranes?

Vanesa Manero and Josep M. Oliva*

Instituto de Química-Física Rocasolano (CSIC), ES-28006 Madrid, Spain

Luis Serrano-Andrés

Instituto de Ciencia Molecular, Universidad de Valencia, ES-46100 Valencia, Spain

Douglas J. Klein

Texas A&M University at Galveston, Galveston, Texas 77553-1675

Received February 19, 2007

Abstract: The stability of endohedral carboranes $X@ \{1,n\text{-C}_2\text{B}_{10}\text{H}_{12}\}$ ($X = \text{Li}^+, \text{Be}^{2+}; n = 2, 7, 12$) and $X@ \{\text{CB}_{11}\text{H}_{12}^-\}$ ($X = \text{Li}^+, \text{Be}^{2+}$) is studied using electronic structure calculations with the B3LYP/6-311+G(d,p) model. Our calculations suggest that all endohedral compounds are local energy minima; for the exohedral complexes $X \cdots \text{cage}$, the global energy minimum always corresponds to the X atom above a triangular face of the icosahedron. In the latter the X atom is furthest apart from the carbon atoms of the cage. As opposite to exohedral $\{\text{Be}^{2+} \cdots \text{cage}\}$ complexes, no global energy minima were found for exohedral complexes $\{\text{Li}^+ \cdots \text{cage}\}$ whereby a carbon atom is present in the triangular face of the icosahedron below the Li^+ cation.

1. Introduction

To our knowledge, no reports on the syntheses of endohedral carboranes have appeared in the bibliography to date.¹ The potential applications of these complexes within such research fields as nanotechnology or biology are still to be investigated: While fullerene-derived endohedral complexes have been thoroughly studied since the first detection of C_{60} ,² with prediction of insertion and ejection mechanisms for the endohedral atoms,^{3,4} this is not the case with carborane-derived endohedral compounds. Atom-filled carbon nanotubes and fullerene boxes have been proposed as superconductors, drug-delivery agents, and 3D atom carriers under different control forms.^{5–7} In a recent work, a nanoencapsulation of two o-carborane molecules has been carried out through $\text{BC}-\text{H} \cdots \pi$ hydrogen bonds in a ball-and-socket structure.⁸ What should we then expect for endohedral complexes derived from carboranes? In a recent work, an ejection mechanism of the endohedral atom viable for the $\text{Li}^+@ \text{CB}_{11}\text{H}_{12}^-$ system was proposed on theoretical grounds.⁹ On the other hand, related guest systems derived from the monoanion $[\text{CB}_{11}\text{H}_{12}]^-$, such as the stable weakly nucleo-

philic anion $[\text{CB}_{11}\text{Me}_{12}]^-$, have been synthesized with an inner negative charge prone to accept cationic species.¹⁰ Predictions on novel stuffed polyhedral boranes ($X@ \text{B}_{12}\text{H}_{12}$)ⁿ have been previously published,^{1,11,12} but we are not aware of similar reports considering carborane clusters as guest systems as the ones we propose in the current report. In this work we present a computational study on the stabilities and geometries of exohedral ($X \cdots \text{cage}$) and endohedral ($X@ \text{cage}$) icosahedral carboranes derived from ortho-carborane, meta-carborane, para-carborane, the monoanion $\text{CB}_{11}\text{H}_{12}^-$, and the cations $X = \{\text{Li}^+, \text{Be}^{2+}\}$.

2. Computational Method

All the calculations in this work were performed at the B3LYP/6-31G(d) and B3LYP/6-311+G(d, p) level of theory with the suite of programs Gaussian03.¹³ Energy minima were characterized by computing second derivatives and harmonic vibrational frequencies used to obtain zero point vibrational energy (ZPE) corrections, at the same level of theory. The global energy minima for the exo complexes $X \cdots \text{cage}$ were found after positioning the X atom above

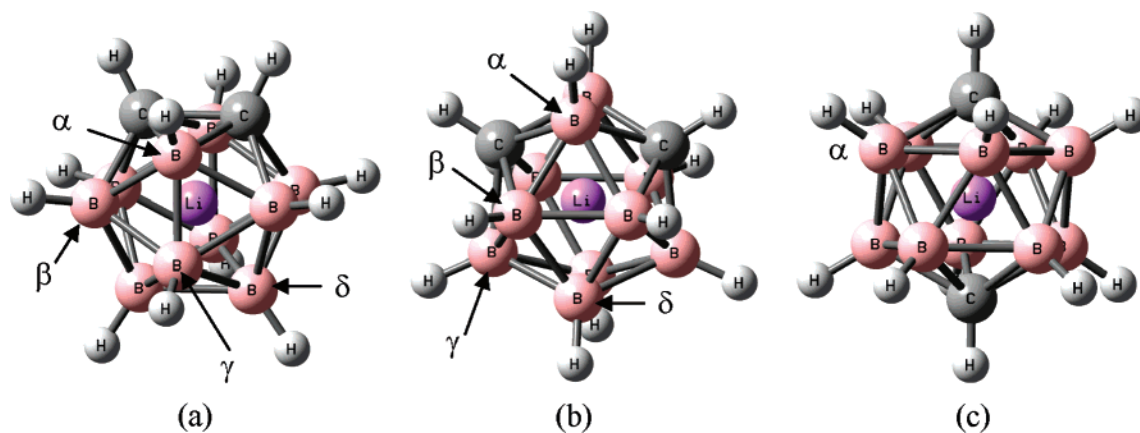


Figure 1. Optimized geometries of (a) Li^+ @o-carborane (**1a**), (b) Li^+ @m-carborane (**1b**), and (c) Li^+ @p-carborane (**1c**). All calculations are at the B3LYP/6-311+G(d, p) level of theory. All geometries correspond to energy minima.

Table 1. Selected Distances (Å), Energies (au), Zero-Point Energy (ZPE) Corrections (kcal/mol), and Strain Energies E^c (kcal/mol) in Compounds **1a–c** and **3a–c** Computed at the B3LYP/6-311+G(d, p) Level of Theory

compound	C···C	Li···C	Li···B $_{\alpha}$	Li···B $_{\beta}$	Li···B $_{\gamma}$	Li···B $_{\delta}$	E	ZPE	E^c
1a	1.712	1.712	1.836	1.789	1.750	1.727	−339.218643	108.7	183.1
1b		1.673	1.831 ^a	1.787 ^a	1.756 ^a	1.741 ^a	−339.240692	109.7	181.1
1c		1.644 ^a	1.783 ^a				−339.241376	109.9	177.2
compound	C···C	Be···C	Be···B $_{\alpha}$	Be···B $_{\beta}$	Be···B $_{\gamma}$	Be···B $_{\delta}$	E	ZPE	E^c
3a	1.701	1.778	1.901	1.804	1.729	1.687	−345.989878	105.6	135.8
3b		1.692	1.898 ^a	1.802 ^a	1.751 ^a	1.713 ^a	−346.009676	105.8	128.7
3c		1.637 ^a	1.800 ^a				−346.007415	105.6	118.8

^a Average of Li/Be···B/C distances.

all nonequivalent triangular faces of the respective icosahedron and checking energies and frequencies of all optimizations; the global energy minimum of the X···cage structure corresponds to the system with lowest energy and all positive frequencies. Only results from the B3LYP/6-311+G(d, p) calculations are included in this work since the conclusions reached are identical for both basis sets.

3. Results and Discussion

3.1. Endohedral Systems X@carborane. *3.1.1. Li^+ @{o-carborane}, Li^+ @{m-carborane}, and Li^+ @{p-carborane}.* The optimized geometries of the endohedral carboranes derived from Li^+ and o-carborane (**1a**), m-carborane (**1b**), and p-carborane (**1c**) are shown in parts a–c, respectively, of Figure 1. The optimized structures depicted in Figure 1 all correspond to energy minima.

Table 1 gathers selected geometrical parameters and energies of the endo compounds displayed in Figure 1 (the optimized geometries of all systems in this work are included in the Supporting Information).

A comprehensive computational study of the dependence of C···C distances in o-carboranes as a function of the substituents on the C's in the cage was recently published.¹⁴ The computed C–C bond distance in o-carborane is $R_{\text{CC}} = 1.627$ Å–B3LYP/6-311+G(d, p) calculations—the experimental value is $R_{\text{CC}} = 1.629$ Å.¹⁵ As shown in Table 1, when a Li^+ atom is introduced in the o-carborane cage, the C···C distance increases by ~ 0.09 Å. In **1a** the two carbon atoms and the Li atom form an equilateral triangle. In **1b** and **1c** the Li···C distance decreases as compared to **1a** by ~ 0.04

Å and ~ 0.07 Å, respectively. The most noticeable change in the remaining parameters included in Table 1 are the Li···B $_{\alpha}$ distances, the latter decreasing for **1c** as compared to **1a** in ~ 0.05 Å. An enhanced stability is thus evidenced from o-carborane to p-carborane when a Li^+ atom is introduced in the cage, as in the case of simple o-carborane, m-carborane, and p-carborane, where the energy order is $E(\text{o-carborane}) < E(\text{m-carborane}) < E(\text{p-carborane})$.

3.1.2. Be^{2+} @{o-carborane}, Be^{2+} @{m-carborane}, and Be^{2+} @{p-carborane}. The optimized geometries of the endohedral carboranes derived from Be^{2+} and o-carborane (**3a**), m-carborane (**3b**), and p-carborane (**3c**) are shown in parts a–c, respectively, of Figure 2. All optimized structures depicted in Figure 2 correspond to energy minima.

As shown in Figure 2 and Table 1, the endohedral complexes derived from Be^{2+} and o-carborane, m-carborane, and p-carborane are stable structures from the energetical point of view (all correspond to local energy minima). The C···C distance in **3a** is even smaller than in the Li analogue (**1a**), by ~ 0.01 Å. However, for ortho and meta derivatives the cages show larger X···B parameters for α and β boron atoms with X = Be^{2+} . This behavior is opposite for γ and δ boron atoms as shown in Table 1. Note that the energetic order in **3a–c** is now $E(\mathbf{3b}) < E(\mathbf{3c}) < E(\mathbf{3a})$, as opposed to the neutral and endohedral complexes derived from Li^+ , where the sequence is $E(\mathbf{3c}) < E(\mathbf{3b}) < E(\mathbf{3a})$.

3.2. Exohedral X···Cage Systems. *3.2.1. Global Minima in Li^+ ···{Carborane} Complexes.* In section 2.1 we showed that the endohedral compounds X@carborane—with X = { Li^+ , Be^{2+} } and carborane = {o-carborane, m-carborane,

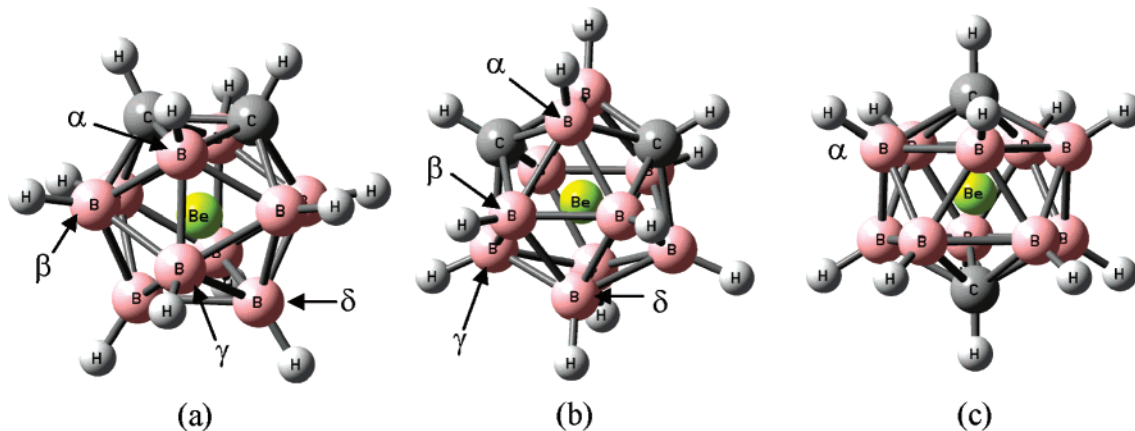


Figure 2. Optimized geometries of (a) Be²⁺@o-carborane (**3a**), (b) Be²⁺@m-carborane (**3b**), and (c) Be²⁺@p-carborane (**3c**). All calculations are at the B3LYP/6-31G* level of theory. All geometries correspond to energy minima.

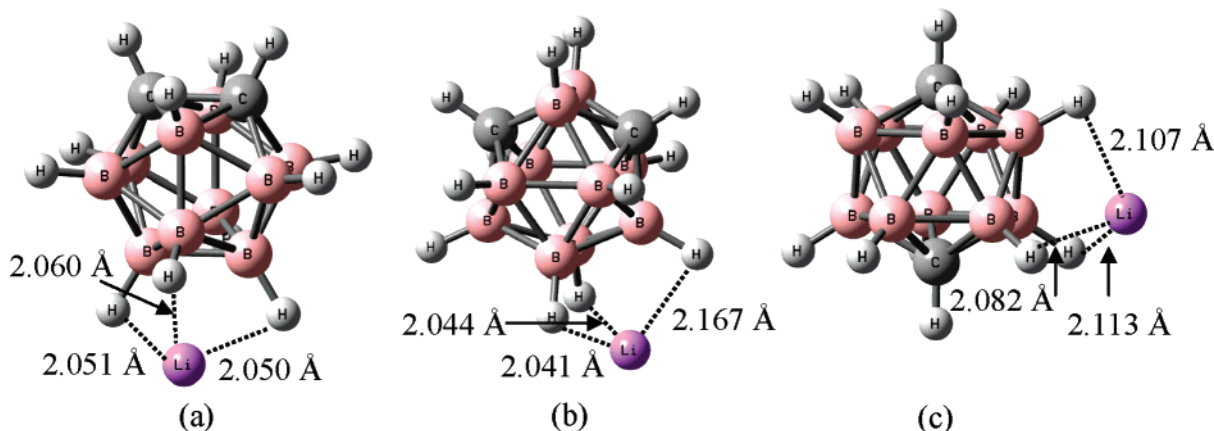


Figure 3. Optimized geometries of the exohedral complex Li⁺...cage (a) Li⁺...{o-carborane} (**2a**), (b) Li⁺...{m-carborane} (**2b**), and (c) Li⁺...{p-carborane} (**2c**). All calculations are at the B3LYP/6-311+G(d, p) level of theory. All geometries correspond to energy minima. Also displayed are the closest Li...H distances.

p-carborane}—corresponded to local energy minima. Similar endohedral compounds derived from fullerenes have been observed and synthesized^{16,17} since the first detection of C₆₀ more than 20 years ago.² A comprehensive search of energies on the surface of o-carborane, m-carborane, and p-carborane and the Li⁺ cation showed that the global energy minima of the complex Li⁺...cage corresponds to the structures displayed in Figure 3.

Other local minima were found above the triangular faces of the boron cage, but they were always higher in energy as compared to the structures displayed in Figure 3: the global minimum is always in the triangular face of the icosahedron furthest apart from any of the carbon atoms in the cage, regardless of the carborane isomer. No triangular faces in complexes **2a** and **2c** containing at least one carbon atom correspond to energy minima. However, this is not the case for the meta complex **2b**. Table 2 gathers the energetics and zero-point energy corrections for the complexes displayed in Figure 3.

The lowest energy complex corresponds to the meta isomer: $E(\mathbf{2b}) < E(\mathbf{2c}) < E(\mathbf{2a})$. A B3LYP/6-311+G(d, p) calculation on Li–H results in a distance $R(\text{Li–H}) = 1.592$ Å; in the radical cation (Li–H)^{•+}, the computed distance is $R = 2.188$ Å. Therefore the situation displayed in Figure 3, with regards to Li...H distances, is closer to the radical cation

Table 2. Energies (au) and Zero-Point Energy (ZPE) Corrections (kcal/mol) for Exo Complexes **2a–c** and **4a–c** Computed at the B3LYP/6-311+G(d, p) Level of Theory

compound	E	ZPE
2a	–339.510368	111.8
2b	–339.529334	111.9
2c	–339.523684	111.8
4a	–346.206304	111.6
4b	–346.214700	111.5
4c	–346.196739	111.1

(Li–H)^{•+} rather than to the neutral Li–H system; we should also take into account the multiple coordination of the Li⁺ cation around the carborane cage, as usual for this cation.¹⁸

3.2.2. Global Minima in Be²⁺...{Carborane} Complexes. Turning now to the Be²⁺ cation, we also performed a comprehensive search for energy minima around the carborane cage surface and this cation. Figure 4 shows the global minima obtained from the interaction of Be²⁺ and o-carborane, m-carborane, and p-carborane leading to complexes **4a–c**, respectively.

A B3LYP/6-311G+(d, p) geometry optimization on BeH₂ gives $R(\text{Be–H}) = 1.327$ Å. For the radical cation (BeH₂)^{•+}, the same calculation gives $R(\text{Be–H}) = 1.413$ Å. As for the dication, the optimization results in a C_{2v} Be²⁺...{(H₂)

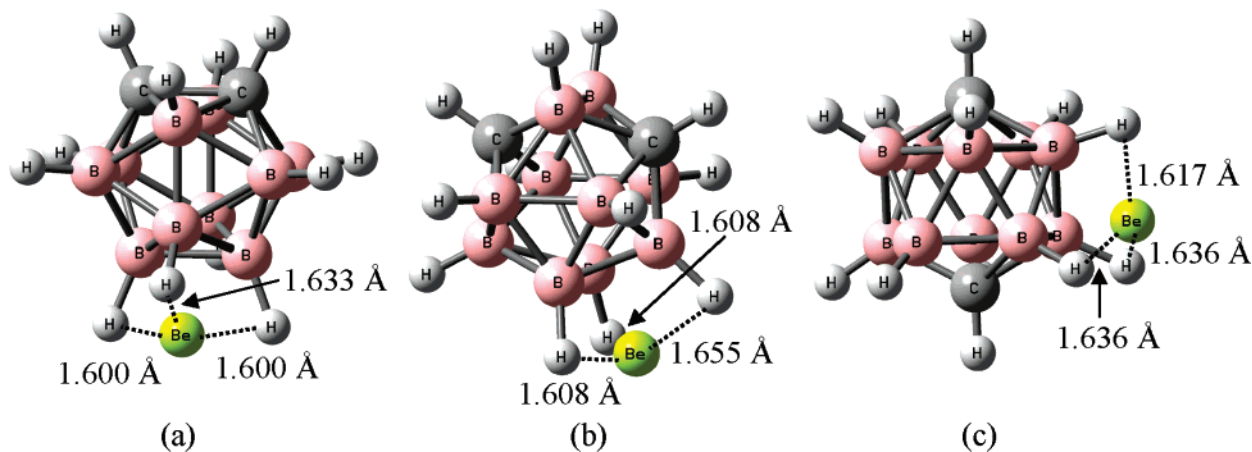


Figure 4. Optimized geometries of the exohedral complex $\text{Be}^{2+}\cdots\{\text{o-carborane}\}$ (**4a**), (b) $\text{Be}^{2+}\cdots\{\text{m-carborane}\}$ (**4b**), and (c) $\text{Be}^{2+}\cdots\{\text{p-carborane}\}$ (**4c**). All calculations are at the B3LYP/6-311+G(d, p) level of theory. All geometries correspond to energy minima. Also displayed are the closest $\text{Be}\cdots\text{H}$ distances.

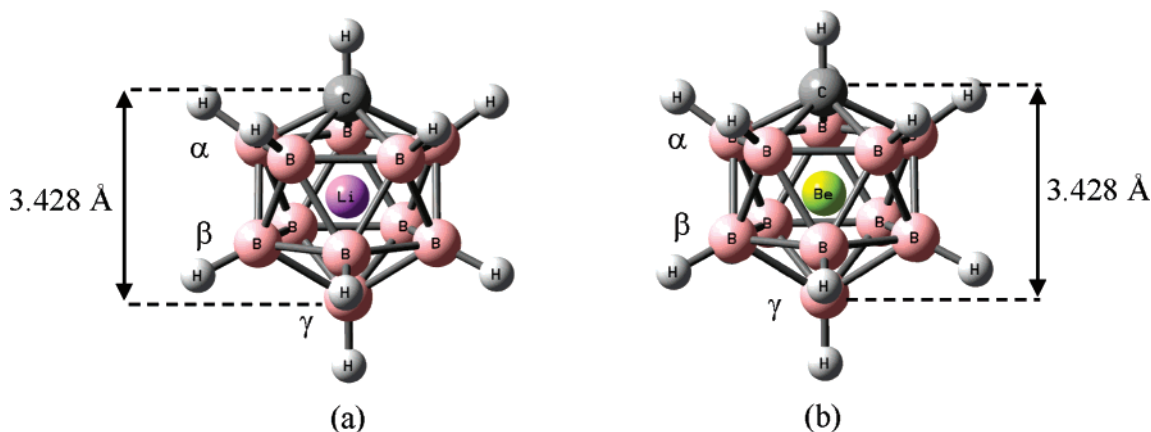


Figure 5. Optimized geometries of the endohedral complexes (a) $\text{Li}^+\text{@}\{\text{CB}_{11}\text{H}_{12}^-\}$ (**5**) and (b) $\text{Be}^{2+}\text{@}\{\text{CB}_{11}\text{H}_{12}^-\}$ (**7**). All calculations are at the B3LYP/6-311+G(d, p) level of theory. All geometries correspond to energy minima.

complex with $R(\text{H}-\text{H}) = 0.821 \text{ \AA}$ and $R(\text{Be}-\text{H}) = 1.629 \text{ \AA}$. The computed frequency for this complex corresponding to the $\text{H}\cdots\text{H}$ stretching is 3516 cm^{-1} ; for isolated H_2 , this computed frequency is 4416 cm^{-1} ($R(\text{H}-\text{H}) = 0.744 \text{ \AA}$), and for the radical cation $(\text{H}_2)^{\bullet+}$, 2051 cm^{-1} ($R(\text{H}-\text{H}) = 1.108 \text{ \AA}$). We can thus deduce that the bonding situation derived from the complexes $\text{Be}^{2+}\cdots\text{cage}$ **4a**–**c** corresponds—at least from the $\text{Be}\cdots\text{H}$ distances—more closely to a $(\text{Be}^{2+})\cdots(\text{H})$ attractive interaction rather than a $\text{Be}-\text{H}$ bond. Note that the $\text{Be}\cdots\text{H}$ distances in the exo complexes are $\sim 0.5 \text{ \AA}$ smaller than in the exo Li complexes (see Figure 3). The energies and ZPE corrections for exo Be complexes are gathered in Table 2: The energetic order is $E(\mathbf{4b}) < E(\mathbf{4a}) < E(\mathbf{4c})$, which is different from the exo Li complexes. Finally, we should point out that we found other local energy minima (higher in energy than those displayed in Figure 4) with Be^{2+} above triangular faces of the icosahedron where one or two carbon atoms are present in the face: As mentioned above, this is not the case for the exo Li-derived complexes.

3.3. Endohedral Complexes $\text{X}\text{@}\{\text{CB}_{11}\text{H}_{12}^-\}$, $\text{X} = \{\text{Li}^+, \text{Be}^{2+}\}$. We turn now to the endohedral systems derived from the monoanion $\{\text{CB}_{11}\text{H}_{12}^-\}$ and the cations Li^+ and Be^{2+} . Figure 5 shows the optimized structures for $\text{Li}^+\text{@}\{\text{CB}_{11}\text{H}_{12}^-\}$

(**5**) and $\text{Be}^{2+}\text{@}\{\text{CB}_{11}\text{H}_{12}^-\}$ (**7**). The labels for the corresponding exo complexes (see section 2.4) $\text{Li}^+\cdots\{\text{CB}_{11}\text{H}_{12}^-\}$ and $\text{Be}^{2+}\cdots\{\text{CB}_{11}\text{H}_{12}^-\}$ are **6** and **8**, respectively.

As shown in Figure 5, the encapsulation of Li^+ or Be^{2+} inside $\text{CB}_{11}\text{H}_{12}^-$ hardly changes the apical distances between C and B_γ . In **7**, the Be atom is pushed down far from C, $\sim 0.02 \text{ \AA}$ as compared to **5**, and closer to B_γ . From the values of Table 3, we can deduce that the $\text{X}\cdots\text{C}$ and $\text{X}\cdots\text{B}_\alpha$ distances in **5** are smaller than in **7**; however, for $\text{X}\cdots\text{B}_\beta$ and $\text{X}\cdots\text{B}_\gamma$ the situation is inverted, with longer distances for **5** as compared to **7**. In other words it is apparent that Be^{2+} is pushed toward the lower part of the carborane cage as compared to the Li^+ endohedral complex.

3.4. Exohedral Complexes $\text{X}\cdots\{\text{CB}_{11}\text{H}_{12}^-\}$, $\text{X} = \{\text{Li}^+, \text{Be}^{2+}\}$. Figure 6 depicts the optimized structures of the global energy minima derived from Li^+ and Be^{2+} and the monoanion $\text{CB}_{11}\text{H}_{12}^-$: $\text{Li}^+\cdots\{\text{CB}_{11}\text{H}_{12}^-\}$ (**6**) and $\text{Be}^{2+}\cdots\{\text{CB}_{11}\text{H}_{12}^-\}$ (**8**).

Comparison of Figure 3 and Figure 6a shows that the Li atom is closer to the cage by ~ 0.10 – 0.15 \AA as compared to the exo complex derived from neutral carboranes, which can be attributed to the negative charge of the cage monoanion, since the Mulliken charges on Li (in units of $|e|$) for **2a**–**c**

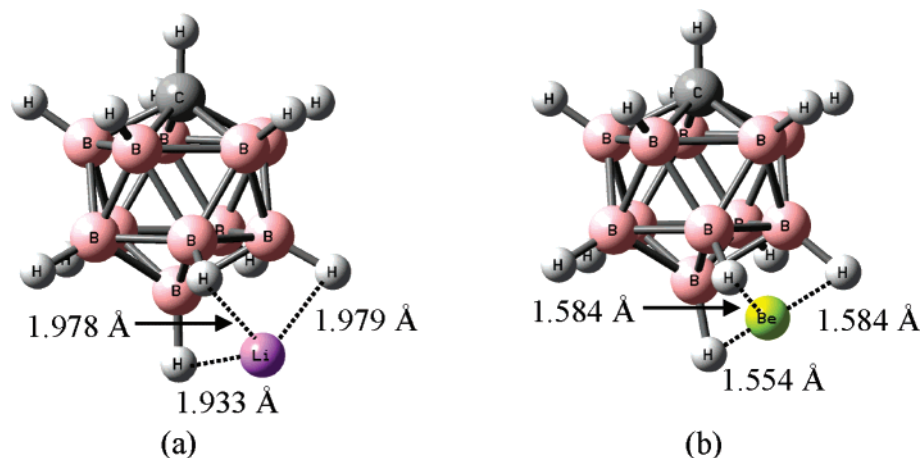


Figure 6. Optimized geometries of the exohedral complexes corresponding to global energy minima (a) $\text{Li}^+\cdots\{\text{CB}_{11}\text{H}_{12}^-\}$ (**6**) and (b) $\text{Be}^{2+}\cdots\{\text{CB}_{11}\text{H}_{12}^-\}$ (**8**). All calculations are at the B3LYP/6-311+G(d, p) level of theory.

Table 3. Selected Distances, Energies (au), Zero-Point Energy (ZPE) Corrections (kcal/mol), and Strain Energies E^c (kcal/mol) for Endohedral Complexes and Exohedral Complexes $\text{Li}^+\text{@}\{\text{CB}_{11}\text{H}_{12}^-\}$ = **5**, $\text{Li}^+\cdots\{\text{CB}_{11}\text{H}_{12}^-\}$ = **6**, $\text{Be}^{2+}\text{@}\{\text{CB}_{11}\text{H}_{12}^-\}$ = **7**, and $\text{Be}^{2+}\cdots\{\text{CB}_{11}\text{H}_{12}^-\}$ = **8**^a

compound	X...C	X...B _α	X...B _β	X...B _γ	<i>E</i>	ZPE	<i>E</i> ^c
5	1.692	1.802 ^b	1.762 ^b	1.736	-326.301887	108.7	149.9
7	1.722	1.830 ^b	1.750 ^b	1.706	-333.256216	106.0	96.7
6					-326.540828	109.8	
8					-333.410348	110.1	

^a All computations with the B3LYP/6-311+G(d, p) model. ^b Average of Li/Be...B/C distances.

and **6** are as follows: $q_{\text{Li}}(\mathbf{2a}) = 0.68$, $q_{\text{Li}}(\mathbf{2b}) = 0.68$, $q_{\text{Li}}(\mathbf{2c}) = 0.68$, and $q_{\text{Li}}(\mathbf{6}) = 0.56$.

As for the exo Be complex (**8**), comparison of Figures 4 and 6 shows a decrease of the $\text{Be}\cdots\text{H}$ distance by ~ 0.05 Å as compared to the exo neutral complexes **4a–c**. Again, this can be fairly attributed to the compensation of charges between the anion and the cation: $q_{\text{Be}}(\mathbf{4a}) = 0.36$, $q_{\text{Be}}(\mathbf{4b}) = 0.38$, $q_{\text{Be}}(\mathbf{4c}) = 0.36$, and $q_{\text{Be}}(\mathbf{8}) = 0.21$.

Again, the charge on Be has been reduced almost to half from the exo neutral carborane complexes **4** to the exo complex **8** with the carborane monoanion $\text{CB}_{11}\text{H}_{12}^-$.

3.5. Strain Energies. In this section we analyze the strain energies E^c , which are defined as the difference between the exohedral and endohedral structures. The strain energies indicate the amount by which the exohedral species are energetically more favorable than the endohedral complexes. As indicated in Table 1 (last column), the strain energies for the Li complexes with the neutral carboranes are 50–60 kcal/mol larger than the Be counterparts, hence indicating an “easier” path (at least thermodynamically) toward the encapsulated complex for the latter. Turning to the complexes derived from Li^+ and Be^{2+} and the monoanion $\text{CB}_{11}\text{H}_{12}^-$ —Table 3—the strain energies are smaller than in the previous complexes **1** and **3**. For Li^+ and Be^{2+} complexes, the differences in strain energies range from 30 to 35 kcal/mol for the Li^+ complexes and 20–40 kcal/mol for the Be^{2+} complexes when the latter interact with the neutral (ortho, meta, and para) and monoanionic carboranes.

4. Conclusions

The optimized structures for the endohedral and exohedral complexes derived from the interaction of Li^+ and Be^{2+} cations with icosahedral neutral and monoanionic carboranes are reported using the B3LYP/6-311+G(d, p) model. All endohedral structures reported correspond to local energy minima with a range of strain energies. With regards to exohedral structures, as opposed to Be^{2+} complexes, a comprehensive energy minimum search shows that for Li^+ no energy minima are found whereby a carbon atom is present in the triangular face of the icosahedron. We hope that the results presented in this work encourage new forthcoming routes toward the experimental syntheses of the endohedral compounds $\{\text{X}@C_n\text{B}_{12-n}\text{H}_{12}\}^{n-2}$ ($n = 1, 2$), ($\text{X} = \text{Li}^+, \text{Be}^{2+}$).

Acknowledgment. This work was financed under project MAT2006-13646-C03-02 from the Spanish Ministry of Science and Education. J.M.O. is grateful to Professor Neil L. Allan (University of Bristol, U.K.) for helpful discussions.

Supporting Information Available: Optimized geometries of all systems. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Jemmis, E. D.; Balakrishnarajan, M. M. *J. Am. Chem. Soc.* **2000**, *122*, 7392–7393.
- Kroto, H. W.; Heath, J. R.; O’Brien, S. C.; Curl, R. F.; Smalley, R. E. *Nature* **1985**, *318*, 162–163.
- Murry, R. L.; Scuseria, G. E. *Science* **1994**, *263*, 791–793.
- Yumura, T.; Sato, Y.; Suenaga, K.; Urita, K.; Iijima, S. *Nano Lett.* **2006**, *6*, 1389–1395.
- Tenne, R. *Nature* **2004**, *431*, 640–641.
- Regan, B. C.; Aloni, S.; Ritchie, R. O.; Dahmen, U.; Zettl, A. *Nature* **2004**, *428*, 924–927.
- Lee, J.; Kim, H.; Kahng, S.-J.; Kim, G.; Son, Y.-W.; Ihm, J.; Kato, H.; Wang, Z. W.; Okazaki, T.; Shinohara, H.; Kuk, Y. *Nature* **2002**, *415*, 1005–1008.
- Raston, C. L.; Cave, G. W. V. *Chem. Eur. J.* **2004**, *10*, 279–282.

- (9) Serrano-Andrés, L.; Oliva, J. M. *Chem. Phys. Lett.* **2006**, *432*, 235–239.
- (10) King, B. T.; Janoušek, Z.; Grüner, B.; Trammell, M.; Noll, B. C.; Michl, J. *J. Am. Chem. Soc.* **1996**, *118*, 3313–3314.
- (11) Charkin, O. P.; Klimenko, N. M.; Moran, D.; Mebel, A. M.; Charkin, D. O.; Schleyer, P. v. R. *J. Phys. Chem. A* **2002**, *106*, 11594–11602.
- (12) Charkin, O. P.; Klimenko, N. M.; Moran, D.; Mebel, A. M.; Charkin, D. O.; Schleyer, P. v. R. *Inorg. Chem.* **2001**, *40*, 6913–6922.
- (13) 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.
- (14) Oliva, J. M.; Allan, N. L.; Schleyer, P. v. R.; Viñas, C.; Teixidor, F. *J. Am. Chem. Soc.* **2005**, *127*, 13538–13547.
- (15) Davidson, M. G.; Hibbert, T. G.; Howard, J. A. K.; Mackinnon, A.; Wade, K. *Chem. Commun.* **1996**, *19*, 2285–2286.
- (16) Bethune, D. S.; Johnson, R. D.; Salem, J. R.; de Vries, M. S.; Yannoni, C. S. *Nature* **1993**, *366*, 123–128.
- (17) Wang, C.-R.; Dennis, J. S.; Inakuma, M.; Shinohara, H. In *Fullerenes: Recent Advances in the Chemistry and Physics of Fullerenes and Related Materials*; Kadish, K. M., Ruoff, R. S., Eds.; Electrochemical Society: Pennington, 1998; pp 1023–1030.
- (18) *Lithium Chemistry: A Theoretical and Experimental Overview*; Sapse, A.-M., Schleyer, P. v. R., Eds.; Wiley: 1995. CT700042Z

QM-MM Investigation of the Reaction of Peroxynitrite with Carbon Dioxide in Water

Mariano C. González Lebrero and Darío A. Estrin*

*Departamento de Química Inorgánica, Analítica y Química Física –
INQUIMAE-CONICET, Facultad de Ciencias Exactas y Naturales, Universidad de
Buenos Aires, Ciudad Universitaria, Pabellón 2, C1428EHA, Buenos Aires, Argentina*

Received February 13, 2007

Abstract: We have investigated the reaction of peroxynitrite with carbon dioxide in aqueous solution by means of combined quantum-classical (QM-MM) molecular dynamics simulations. In our QM-MM scheme, the reactant was modeled using density functional theory with a Gaussian basis set, and the solvent was described using the mean-field TIP4P force field. The free energy profile of this reaction has been computed using umbrella sampling and multiple steering molecular dynamics (MSMD) schemes. Umbrella sampling methods turned out to be much more efficient than MSMD schemes, due to the possibility of employing a combination of classical and QM-MM thermalization schemes. We found the presence of a significant barrier in the free energy profile associated with the reaction in solution, which is not present in vacuum, that may be ascribed to the significant charge redistribution upon reaction and the concomitant solvation pattern changes.

1. Introduction

Peroxynitrite anion (ONOO^-) is a stable species formed by the reaction of superoxide with nitric oxide in biological environments.¹ The formation of peroxynitrite has been linked to pathology. Research efforts directed to understand the mechanism of reaction of peroxynitrite were initially focused primarily on the reactions of peroxynitrite with substrates with zero-order kinetics (e.g., dimethyl sulfoxide and deoxyribose), with substrates with relatively small second-order rate constants (e.g., methionine and ascorbate), and with a few more reactive substrates, such as thiols.^{1–4} Although these experiments afforded important mechanistic information, the reactions of peroxynitrite with these substrates cannot compete with the reaction of peroxynitrite with CO_2 under physiological conditions, due to the relative high concentration of CO_2 in cellular environment and the fast reaction of these two compounds.

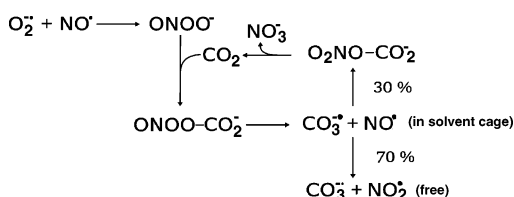
The reaction of peroxynitrite with carbonate buffers was observed initially by Keith and Powel,⁵ but its physiological

importance was not recognized until 1993 when Radi et al. noted the biological relevance of this reaction.⁶

A possible mechanism for this reaction was proposed in 1995.³ It is now firmly established that the reaction of peroxynitrite with CO_2 occurs between the peroxynitrite anion (ONOO^-) and dissolved CO_2 and forms the carbonate radical ($\text{CO}_3^{\bullet-}$) and nitrogen dioxide (NO_2^{\bullet}), as shown in Scheme 1.

Electronic structure calculations of the first step of this reaction, the formation of the adduct nitrosoperoxycarbonate, at different levels of theory show that the reaction is barrierless in vacuum.⁷ On the other hand, the experimental results in aqueous solution suggest the presence of a significant free energy barrier (about 12 kcal/mol).⁸ This fact indicates that the solvent plays a crucial role in the reaction, making it an ideal benchmark for explicit solvent QM-MM methodologies. In this work we have performed molecular dynamics simulations of the reaction in aqueous solution employing a QM-MM strategy to obtain free energy profiles and to understand the reaction mechanism from an atomistic point of view. We critically evaluate the performance of two different advanced sampling tools in the context of QM-

* Corresponding author phone: 54-11-4576-3368; fax: 54-11-4576-3341; e-mail: dario@qi.fcen.uba.ar.

Scheme 1. Reaction Pathways for Peroxynitrite in Vivo

MM calculations, namely, umbrella sampling and multiple steering molecular dynamics schemes.

2. Model and Simulation Methods

The Hybrid QM/MM Hamiltonian. Our computational scheme was constructed by partitioning the system into a quantum-mechanical (QM) and a classical-mechanical (MM) subsystems. Considering a configuration of N_c atoms in the MM subsystem with coordinates and partial charges $\{R_l, q_l, l = 1, \dots, N_c\}$ and N_q atoms in the QM subsystem with coordinates and nuclear charges $\{\tau_a, z_a, a = 1, \dots, N_q\}$, we propose the following expression for the ground state, Born–Oppenheimer potential energy surface that drives the dynamics of the nuclei

$$E[\{R_l\}, \{\tau_a\}] = E_{\text{KS}} + E_{\text{QM-MM}} + E_{\text{MM}} \quad (1)$$

where the first term is a purely QM piece given by the standard Kohn–Sham expression.⁹ The second term in eq 1 accounts for the coupling of the QM and MM subsystems and is given by

$$E_{\text{QM-MM}} = \sum_{l=1}^{N_c} q_l \int \frac{\rho(r)}{|r - R_l|} dr + \sum_{l=1}^{N_c} \sum_{\alpha=1}^{N_q} \left[v_{\text{LJ}}(|R_l - \tau_{\alpha}|) + \frac{q_l z_{\alpha}}{|R_l - \tau_{\alpha}|} \right] \quad (2)$$

where v_{LJ} is the Lennard-Jones potential between the classical and quantum part of the system and $\rho(r)$ is the electron density of the QM subsystem. The last term in eq 1 represents the potential energy contribution from the classical solvent potential, treated with the TIP4P mean-field potential.¹⁰

For the QM region, computations were performed at the generalized gradient approximation (GGA) level, using the BP86^{11–13} combination of exchange and correlation functionals. Gaussian basis sets of double- ζ plus polarization quality were employed for the expansion of the one-electron orbitals.¹⁴ The electronic density was also expanded in an auxiliary basis set;¹⁴ the coefficients for the fit were computed by minimizing the error in the Coulomb repulsion energy. The use of this procedure results in an important speedup of the computation.

In order to describe accurately dissociation processes, we have incorporated into our previously developed QM-MM code,¹⁵ a suitable cutoff scheme in the coupling QM-MM of the cutoff radii in the QM-MM component of the energy.¹⁶

The electron density of the quantum system is given by

$$\rho(r) = \sum_{i=1}^{N_{\text{occ}}} |\psi_i|^2 \quad (3)$$

where each KS molecular orbital, ψ_i , is defined as

$$\psi_i = \sum_k c_i^k g_k(r) \quad (4)$$

where $g_k(r)$ are the contracted basis functions, given by

$$g_k(r) = \sum_{j=1} c_k^j f_j(r) \quad (5)$$

where each $f_j(r)$ is a Gaussian function. Then, the density can be written as

$$\rho = \sum_{i=1}^{N_{\text{occ}}} \left| \sum_{k,j} c_i^{kj} f_j(r) \right|^2 \quad (6)$$

The product of two Gaussian functions of exponents α and β , centered on nuclei A and B, respectively, is proportional to another Gaussian function, centered on a point P

$$f_a(\alpha, r - R_A) f_b(\beta, r - R_B) = K_{\text{AB}} f_c(p, r - R_p) \quad (7)$$

where the constant K_{AB} is given by

$$K_{\text{AB}} = \left(\frac{2\alpha\beta}{(\alpha + \beta)\pi} \right)^{\frac{3}{4}} \exp \left[- \frac{\alpha\beta}{(\alpha + \beta)|R_A - R_B|^2} \right] \quad (8)$$

The exponent of the new Gaussian function centered in R_p is

$$p = \alpha + \beta$$

and the third center P lies on a line joining the centers A and B

$$R_p = \frac{\alpha R_A + \beta R_B}{\alpha + \beta}$$

Substituting eqs 5–7 into 2, we can express the first term of eq 9 as

$$\sum_{i=1}^{N_c} q_i \int \frac{\rho(r)}{|r - R_i|} dr = \sum_j \sum_{i=1}^{N_c} q_i \int \frac{K_j f_j(p_j, r - R_{p_j})}{|r - R_i|} dr \quad (9)$$

A possible way to compute eq 9 when using periodic boundary conditions is to include only the classical point charges located at a distance smaller than R_{cut} from the geometric center (or mass center) of the quantum subsystem, with R_{cut} equal to half the solvent box length.

However, this turns out to yield very poor results in processes in which the spatial extension of the quantum subsystem changes significantly upon reaction. This effect results in a very pronounced shift in the free energy profile when the size of the QM subsystem becomes similar to the box length. This fact has been noted by York et al.¹⁷ in a recent work.

An alternative scheme which alleviates this flaw consists of using a cutoff scheme in which we keep the integrals for

which the classical partial charge is located at a distance smaller than R_{cut} from the R_p corresponding to that integral

$$\sum_{i=1}^{N_c} q_i \int \frac{\rho(r)}{|r - R_i|} dr \approx \sum_j \sum_{i=1}^{N_c} q_i \int \frac{K_j f_j(p_j, r - R_{p_j})}{|r - R_i|} dr |R_i - R_{p_j}| < R_{\text{cut}} \quad (10)$$

Molecular Dynamics Simulations. In all our simulation experiments, the coordinate Verlet algorithm¹⁸ was employed to integrate Newton's equations of motion with a time step of 0.2 fs. Constraints associated with the intramolecular distances in water were treated using the SHAKE algorithm.¹⁹ The Lennard-Jones parameters for the quantum subsystem atoms are ϵ and σ of 0.200, 0.155, and 1.70 kcal/mol and 3.900, 3.154, and 3.65 Å, for N, O, and C, respectively. The solute was solvated in a cubic box of size $a = 24$ Å, containing 497 water molecules. Initial configurations were generated from preliminary 100 ps classical equilibration runs in which the quantum solute was replaced by a rigid peroxynitrite (or adduct) with partial charges obtained from a Mulliken population analysis in vacuo. At $t = 0$, the classical solute is replaced by a solute described at the DFT level, according to the hybrid methodology described above. An additional 2 ps of equilibration was performed using the QM-MM scheme. During the simulations, the temperature was held constant at 298 K by the Berendsen thermostat.²⁰ The solute and the rest of the system were coupled separately to the temperature bath. In order to compare solvation structures additional equilibrium simulations were performed for the reactants and products in water boxes with 497 solvent molecules and 24 Å of side.

If the free energy barriers are of the same order of magnitude as the thermal fluctuations, it is feasible to obtain the free energy profiles associated with a given process directly from the MD simulations. However, to have an appropriate sampling in accessible simulation times, the barriers should be smaller than thermal fluctuations. In cases where barriers are suspected to be high, biased sampling is required to obtain the free energy profile, also called potential of mean force (PMF). We will present here two different biased sampling methods: umbrella sampling and steered molecular dynamics.

Umbrella Sampling. This method²¹ attempts to overcome the sampling problem by modifying the potential function so that the unfavorable states are sampled sufficiently. The potential function is modified by adding a weighting function that usually takes a harmonic form. An umbrella sampling calculation involves a series of stages (called simulation windows), each characterized by a particular value of the reaction coordinate. The PMF is then obtained by superposing the results obtained for all the series of windows.

Multiple Steering Molecular Dynamics. The multiple steering molecular dynamics (MSMD) approach, originally proposed by Jarzynski,²² is based on the following relation

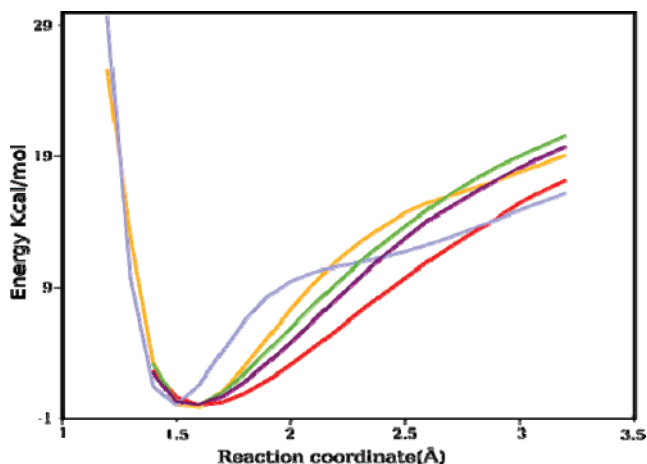


Figure 1. Potential energy profiles for the $\text{ONOO}^- + \text{CO}_2 \rightarrow \text{ONOOCO}_2^-$ reaction. Results obtained using MP2, HF, B3LYP, BP86, and BLYP are depicted in orange, blue, green, violet, and red lines, respectively.

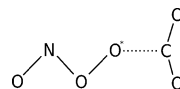


Figure 2. Schematic representation of the reactants. The reaction coordinate is depicted with a dotted line.

between the nonequilibrium dynamics and equilibrium properties

$$\exp[-\Delta A(\xi)]/k_B T = \langle \exp[-W(\xi)/k_B T] \rangle \quad (11)$$

in which $W(\xi)$ is the external work performed on the system as it evolves from the initial to the final state along the reaction coordinate ξ .

In MSMD the original potential is modified by adding to the potential energy a time-dependent external potential, usually harmonic, that moves the system along the reaction coordinate by varying the potential well according to

$$E'(r) = E(r) + k[\xi - (\xi_0 + v\Delta t)]^2 \quad (12)$$

where v is the pulling speed that moves the system along the reaction coordinate.

The PMF is obtained by performing several MSMD runs, collecting the work done at each time step, and then properly averaging it, according to eq 11. Usually, the pulling speed is chosen so that the system moves smoothly but faster than in a true reversible simulation.^{23–25}

Since the averages in this equation are exponential, the results are mostly determined by the trajectories of lower work. This can be addressed by replacing the exponential average by a Taylor expansion and keeping only the terms up to order 2.

For the umbrella sampling method we have taken 9 windows with 100 ps of total integration time. For the steering molecular dynamics method we performed 12 nonequilibrium trajectories going from a reaction coordinate of 1.4–5.4 Å, with a velocity of 0.5 Å/ps and a total time of 96 ps.

Validation of the Method. In order to validate the hybrid Hamiltonian we computed the potential energy profile of the

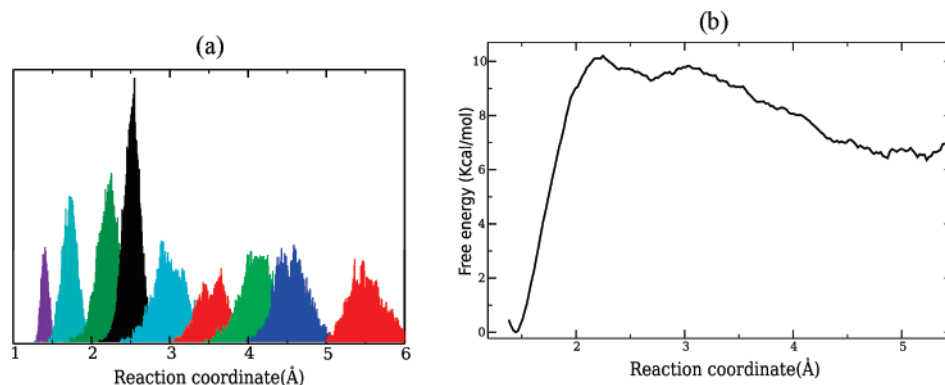


Figure 3. The reaction coordinate histogram for the umbrella sampling simulations (left panel) and the corresponding free energy profile (right panel).

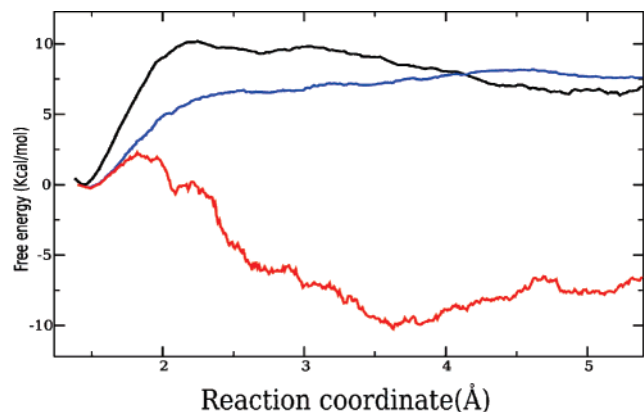


Figure 4. Free energy profile obtained by umbrella sampling (black line), Jarzinski's equation (blue line), and order two truncated expansion of the Jarzinski's equation (red line).

reaction in vacuum employing a variety of methods of electronic structure (Hartree Fock, MP2,²⁶ and DFT using the BP86,^{12,13} BLYP,²⁷ and B3LYP²⁸ functionals) and the basis set 6-31G**. These calculations have been performed using Gaussian 98²⁹ (Figure 1).

The reaction coordinate was chosen as the distance between the terminal O of peroxyinitrite and the carbon atom from the dioxide, as shown in Figure 2.

It can be seen in Figure 1 that the BP86 functional reproduces correctly the results obtained with more sophisticated methods like MP2 or B3LYP at a significantly lower cost. For this reason we use this functional to describe the QM subsystem.

The Lennard-Jones parameters for the peroxyinitrite ion were validated in a previous work,¹⁶ thus only the parameters for the carbon dioxide atoms have to be tested. We have performed an optimization of the adduct (OONO-CO₂) with one water molecule attached to one of the oxygen atoms of the CO₂ with the QM-MM and with a full quantum Hamiltonian. The computed binding energies of this aggregate were 12.7 kcal/mol and 11.9 kcal/mol for full quantum and QM-MM calculations, respectively.

3. Results

Umbrella Sampling. The free energy profile was obtained using 9 simulation windows of the umbrella potential, fixed in values that allow a correct sampling of the reaction

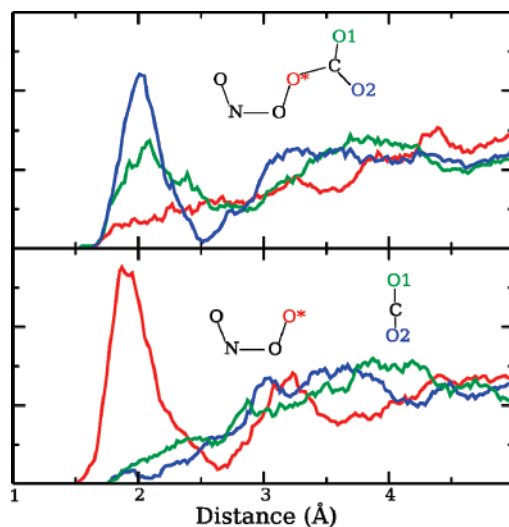


Figure 5. Radial correlation functions of selected atoms with water oxygen atoms from the product (upper panel) and reactant (lower panel). The radial correlations function of the terminal oxygen atom of the peroxyinitrite and of the oxygen atoms from the CO₂ are depicted in red, blue, and green, respectively.

coordinate range spanning from reactants to products. The total simulation time was 100 ps. The initial structures were thermalized for 100 ps each one with a full classical Hamiltonian, in which the solute was represented for a rigid structure represented with Lennard-Jones and Mulliken charges. Subsequently, 2 ps thermalization have been performed with the hybrid Hamiltonian.

In Figure 3 we show the reaction coordinated histogram for the different windows simulation.

The obtained free energy profile is also shown in Figure 3.

Multiple Steered Molecular Dynamics. The free energy profile was also obtained by using 12 independent steered molecular simulations of 8 ps each one using eq 11. The reaction coordinate was moved from 1.4 to 5.4 Å with a velocity of 0.5 Å /ps. The total simulated time was 96 ps, similar to the total time used in the umbrella sampling calculation. The results obtained using this scheme are shown in Figure 4. The results obtained using umbrella sampling are included for comparison. The significant difference between the results obtained using eq 11 and the results

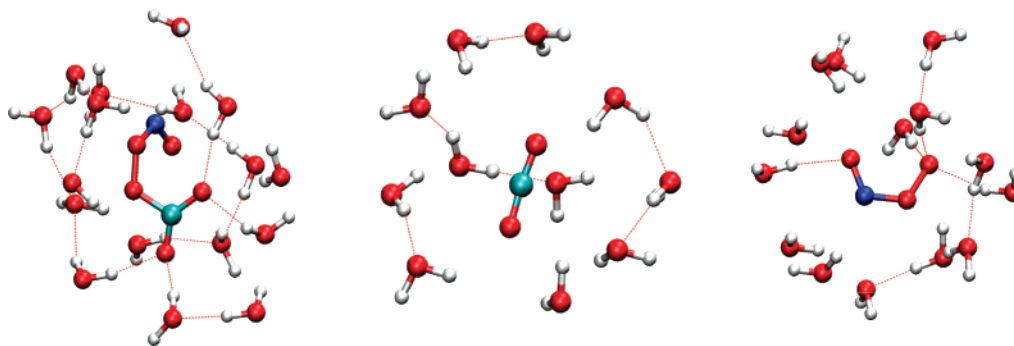


Figure 6. Typical snapshots of the adduct, CO₂, and peroxynitrite (left, middle, and right panels, respectively).

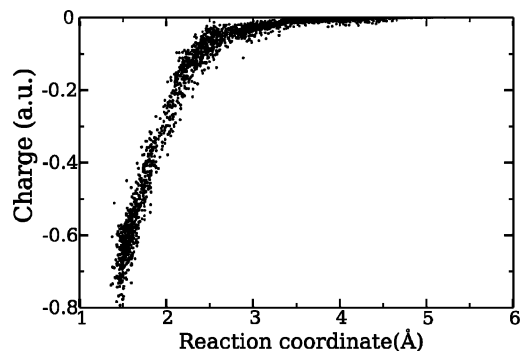


Figure 7. Net Mulliken population on the CO₂ moiety as a function of the reaction coordinate.

obtained using the second-order approximation indicate that there are large statistical errors in using this approach for describing this reaction.

4. Discussion

It can be concluded that the results obtained using the umbrella sampling methodology are more reliable than the results obtained using multiple steered molecular dynamics, for similar total simulation times. This fact can be associated mainly to the steering velocity. Unfortunately, the decrease of this velocity results in an important increase in the already high computational cost. This methodology results inefficiently in systems in which the characteristic relaxation times are not significantly smaller than the accessible total simulation time. In this system the final state of the constrained simulation is far from equilibrium, and this issue generates a systematic error which can only be reduced by decreasing the steering velocity and hence increasing the computational cost. This is due to the fact that there is significant charge redistribution upon reaction, with a concomitant change in solvation patterns which is not represented correctly if the simulation times of the steered molecular dynamics runs are not significantly higher than typical residence times of water molecules. This has also been reported by Cascella in a QM-MM investigation of formamide hydrolysis.³⁰

On the other hand, the results obtained using the umbrella sampling with the same total simulation time are more reliable and in qualitative agreement with experimental results. This fact can be related with the extensive thermalization of the different umbrella windows, which improves the convergence of the latter scheme. The hybrid methodol-

ogy allows us to perform a preliminary (and extensive) thermalization with the full classical Hamiltonian, before switching to the hybrid QM-MM scheme. This combination of classical and QM-MM thermalization schemes improves significantly the efficiency of the umbrella sampling technique compared to multiple steering molecular dynamics techniques at essentially the same computational cost.

The Free Energy Barrier. The free energy profile obtained by the umbrella sampling in aqueous solution shows the existence of a free energy barrier absent in vacuum. This barrier is about 3.8 kcal/mol, which is significantly lower than the experimental estimation. This difference can be tracked down to the DFT electronic structure at the GGA level flaws which typically underestimate the barriers or the absence of polarization effects in the TIP4P force field that can produce also an underestimation of the ion–solvent interaction energies. However, the result is qualitatively correct, and the microscopic view obtained with our simulations can offer important information about the origin of this barrier. In order to get an estimation of the possible DFT flaws in predicting the barriers, we have performed single point calculations of 10 selected snapshots extracted from the simulations, corresponding to reaction coordinates of 2.25 and 1.75 Å (approximate transition state and product, respectively). This provides us an estimation of the activation energy of the reverse reaction. For each snapshot, we have calculated the energy by employing the Gamess-US program³¹ at the DFT and MP2 levels, treating the reactant species quantum mechanically, and the 497 water molecules in the simulation box as TIP4P point charges. The average energy difference between the product and the approximate transition state at the MP2 level is 8.6 kcal higher than that calculated using DFT. This indicates that there is indeed an underestimation of the energy barriers by DFT in this case, compared to the MP2 calculations. This DFT flaw is not so evident in the vacuum calculations. The influence of the selected water model has been assessed by means of a scheme in which polarization is modeled by induced point dipoles on the O and H atoms of water molecules due to the electric field of the quantum subsystem as well as other water molecules. These induced dipoles are iterated to self-consistency.³² We have employed the 10 selected snapshots extracted from the simulations, corresponding to reaction coordinates of 2.25 and 1.75 Å (approximate transition state and product, respectively), and performed single point calculations using TIP4P charges and TIP4P charges plus

induced point dipoles centered at the O and H atoms (1.4146 and 0.0836 Å³, respectively). The average energy difference between the product and the approximate transition state for the MP2-TIP4P plus polarization is only 2.0 kcal/mol higher than the value computed using the MP2-TIP4P scheme. This indicates that the neglect of solvent polarization results also in the underestimation of the barrier. However, it seems that the errors are smaller than those due to DFT.

In order to obtain an atomistic picture of the solvent effects which produce this barrier, it may be a useful result to analyze the radial distribution function of solute atoms with water oxygen atoms corresponding to products and reactants, shown in Figure 5.

The solvation patterns around the oxygen atoms whose effective charge change during the reaction are, as expected, profoundly modified when going from the reactants to the adduct. In the adduct (upper panel) the oxygen atoms of the carbon dioxide are strongly solvated because they bear a significant high negative charge (Mulliken populations of these atoms are in average -0.55 e). On the other hand, the oxygen of peroxyinitrite (O*) is in a hydrophobic part of the molecule and exhibits a weak interaction with water (Mulliken population of this atom is on average -0.25 e). This is confirmed by inspecting typical snapshots (Figure 6).

In the lower panel (reactive) the oxygen atoms of CO₂ (mean values of the Mulliken population of the O atoms are -0.26) are poorly solvated (as expected) and the O* is strongly solvated (mean value of the Mulliken populations for this atom is -0.66). This means that during the reaction the strong hydrogen bonds of the O* atom with water molecules present in the reactant should weaken or break concomitantly with the formation of the adduct. This is probably the main microscopic determinant for the observed free energy barrier. Typical snapshots of CO₂, peroxyinitrite, and the adduct in aqueous solution are shown in Figure 6.

The dependence of the net Mulliken population over the CO₂ moiety upon reaction is shown in Figure 7. Since the system negative charge is localized mostly in the oxygen (O*) atom of the peroxyinitrite in the reactant and in the oxygen atoms of carbon dioxide in the adduct, the net CO₂ Mulliken charge turns out to be a good indicator of the degree of charge redistribution upon reaction.

In Figure 7 we can see the absence of charge transfer for reaction coordinate values larger than 2.6 Å. This means that the bond is practically broken for longer distances and is consistent with the hypothesis that the barrier is produced by the solvent, since the steep rise in the free energy profile is in the range of reaction coordinates 3.2–4.7 Å, in which the degree of charge transfer indicates that the bond has not yet formed.

5. Conclusion

The reaction of peroxyinitrite with carbon dioxide exhibits a barrier in the free energy profile produced by the solvent. The change in the solvation patterns upon reaction is the microscopic determinant of this barrier, since this change implicates the breaking of several hydrogen bonds and the formation of new ones. The results of the QM-MM simulation are in qualitative agreement with the available experi-

mental results. The differences may be due to both the treatment of the experimental data and to limitations of the computational scheme. Umbrella sampling methods turned out to be much more efficient than multiple steered molecular dynamics schemes, due to the possibility with the former methodology to employ a combination of classical and QM-MM thermalization schemes in each simulation window, which is not possible in the MSMD scheme.

Acknowledgment. This work was partially supported by the University of Buenos Aires, Agencia Nacional de Promoción Científica y Tecnológica (project PICT 25667), and CONICET (PIP 5218).

References

- (1) Beckman, J. S.; Beckman, T. W.; Chen, J.; Marshall, P. A.; Freeman, B. A. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 1620–1624.
- (2) Pryor, W. A.; Jin, X.; Squadrito, G. L. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 11173–11177.
- (3) Squadrito, G. L.; Jin, X.; Pryor, W. A. *Arch. Biochem. Biophys.* **1995**, *322*, 53–59.
- (4) Radi, R.; Beckman, J. S.; Bush, K. M.; Freeman, B. A. *J. Biol. Chem.* **1991**, *266*, 4244–4250.
- (5) Keith, W. G.; Powell, R. E. *J. Chem. Soc.* **1969**, A, 90–90.
- (6) Radi, R.; Cosgrove, T. P.; Beckman, J. S.; Freeman, B. A. *Biochem. J.* **1993**, *290*, 51–57.
- (7) Houk, K. N.; Condroski, K. R.; Pryor, W. A. *J. Am. Chem. Soc.* **1996**, *118*, 13002–13006.
- (8) Squadrito, G. L.; Pryor, W. A. *Chem. Res. Toxicol.* **2002**, *15*, 885–895.
- (9) Kohn, W.; Sham, L. J. *Phys. Rev. A* **1965**, *140*, 1133–1138.
- (10) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (11) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (12) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824; **1986**, *34*, 7406–7406.
- (13) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (14) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.* **1992**, *70*, 560–571.
- (15) González Lebrero, M. C.; Bikiel, D. E.; Elola, M. D.; Estrin, D. A.; Roitberg, A. E. *J. Chem. Phys.* **2002**, *117*, 2718–2725.
- (16) González Lebrero, M. C.; Perissinotti, L. L.; Estrin, D. A. *J. Phys. Chem. A* **2005**, *109*, 9598–9604.
- (17) Nam, K.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2005**, *1*, 2–13.
- (18) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon: Oxford, 1987.
- (19) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–331.
- (20) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (21) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–99.

- (22) Jarzynski, C. *Phys. Rev. Lett.* **1997**, 78, 2690–93.
- (23) Crespo, A.; Marti, M. A.; Estrin, D. A.; Roitberg, A. E. *J. Am. Chem. Soc.* **2005**, 127, 940–1.
- (24) Park, S.; Schulten, K. *J. Chem. Phys.* **2004**, 120, 5946–61.
- (25) Hummer, G.; Szabo, A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, 98, 3658–61.
- (26) Krishnan, R.; Pople, J. A. *J. Chem. Phys.* **1980**, 72, 4244–4245.
- (27) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, 37, 785–789.
- (28) Becke, A. D. *J. Chem. Phys.* **1993**, 98, 5648–5652.
- (29) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, J. A.; Montgomery, J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowki, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98, Revision A.1*; Gaussian, Inc.: Pittsburgh, PA, 1998.
- (30) Cascella, M.; Raugei, S.; Carloni, P. *J. Phys. Chem. B* **2004**, 108, 369–375.
- (31) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, 14, 1347–1363.
- (32) Li, H.; Netzloff, H. M.; Gordon, M. S. *J. Chem. Phys.* **2006** 125, 194103-
CT700038W

Understanding Rate Accelerations for Diels–Alder Reactions in Solution Using Enhanced QM/MM Methodology

Orlando Acevedo*[†] and William L. Jorgensen*[‡]

Department of Chemistry and Biochemistry, Auburn University, Auburn, Alabama 36849, and Department of Chemistry, Yale University, 225 Prospect Street, New Haven, Connecticut 06520-8107

Received March 30, 2007

Abstract: The Diels–Alder reactions of cyclopentadiene with 1,4-naphthoquinone, methyl vinyl ketone, and acrylonitrile have been investigated using QM/MM calculations in water, methanol, acetonitrile, and hexane. This extends an earlier AM1-based QM/MM study (*J. Phys. Chem. B* **2002**, *106*, 8078) that only investigated the reactions in water and utilized gas-phase optimized structures as starting points for computations of one-dimensional potentials of mean force (PMFs). Presently, the stationary points were located automatically in multiple solvents by computing two-dimensional PMFs, and the QM method is now PDDG/PM3. The resultant geometries are improved, and relative free energies of activation are well reproduced, e.g., ΔG^\ddagger for the reaction with naphthoquinone is computed to increase upon transfer from water to methanol, acetonitrile, and hexane by 3.2, 4.1, and 5.1 kcal/mol, while the experimental values are 3.4, 4.0, and 5.0 kcal/mol. Ab initio MP2/6-311+G(2d,p) calculations using the CPCM continuum solvent model on gas-phase CBS-QB3 geometries were also found to yield accurate ΔG^\ddagger values in water. However, only the QM/MM methodology reproduced the large rate increases in proceeding from aprotic solvents to water. The dominant factors for the rate variations are enhanced hydrogen bonding for the polarized transition states and reduction in hydrophobic surface area.

Introduction

The Diels–Alder reaction is one of the most powerful carbon–carbon bond forming processes and continues to be an important subject for both computational^{1,2} and experimental studies.³ Solvent effects on the reaction have received much attention due to striking rate accelerations that have been observed in aqueous solution.^{4–11} However, not all Diels–Alder reactions benefit equally in water; e.g., the rate of the reaction between cyclopentadiene and 1,4-naphthoquinone is enhanced by up to 10 000-fold in aqueous over aprotic solvents,⁹ while with acrylonitrile as the dienophile the acceleration is only 31-fold.⁴ Reviews are available on

the mechanistic aspects and solvent effects for Diels–Alder reactions,^{3,7,12} and computational studies have clarified the microscopic variations in solvation along the reaction paths.^{1,13–20} Early proposals that the primary factors responsible for the aqueous acceleration are reduction in hydrophobic surface area as the cycloaddition proceeds⁴ and enhanced hydrogen bonding between water molecules and the transition state¹³ are now largely accepted.^{3–21}

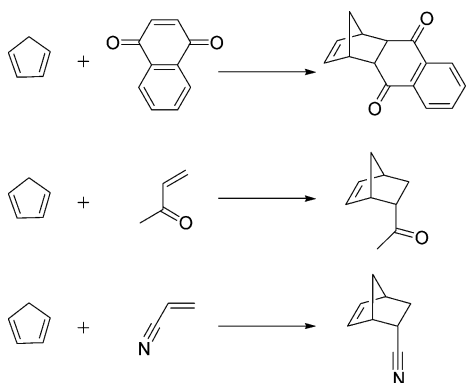
To further explore the solvent-dependence of the reaction rates and geometries of the transition states at the atomic level, the Diels–Alder reactions between cyclopentadiene and three different dienophiles, 1,4-naphthoquinone, methyl vinyl ketone (MVK), and acrylonitrile (Scheme 1), have been investigated using the recently developed PDDG/PM3 semiempirical molecular orbital method^{22,23} in mixed quantum and molecular mechanics (QM/MM) simulations. In our prior QM/MM study of these reactions,¹ only water was

* Corresponding author e-mail: orlando.acevedo@auburn.edu (O.A.) and william.jorgensen@yale.edu (W.L.J.).

[†] Auburn University.

[‡] Yale University.

Scheme 1. Diels–Alder Reactions between Cyclopentadiene and 1,4-Naphthoquinone, Methyl Vinyl Ketone, and Acrylonitrile



considered, and the QM/MM approach used AM1 as the QM method and relied on gas-phase optimized structures as starting points for computations of one-dimensional potentials of mean force (PMFs). In the current study, the reactants, transition structures, and cycloadducts have been located in a fully automated manner in up to four different solvents (water, methanol, acetonitrile, and hexane) using two-dimensional PMF calculations. Transition structures and activation barriers were computed with complete sampling of the geometry for the reacting systems and explicit representation of the solvent molecules. Problems with the use of one-dimensional PMF calculations, particularly for computed transition structures, are illustrated. In addition, changes in solvation along the reaction paths are fully characterized, and comparison is made with results of *ab initio* calculations with the CPCM solvation model.

Computational Methods

QM/MM calculations,²⁴ as implemented in BOSS 4.6,²⁵ were carried out with the reacting system treated using the semi-empirical PDDG/PM3 method. PDDG/PM3 has been extensively tested for gas-phase structures and energetics^{22,23} and has given excellent results in solution-phase QM/MM studies for a variety of organic reactions.^{26,27} The solvent molecules are represented with the TIP4P water model²⁸ and the united-atom OPLS force field for the nonaqueous solvents,²⁹ with the exception of hexane which used the all-atom version.³⁰ The systems consisted of the reactants, plus 390–395 solvent molecules for the nonaqueous solvents, or 730 molecules for water. The systems are periodic and tetragonal with $c/a = 1.5$; a is ca. 25, 27, 29, and 40 Å for water, methanol, acetonitrile, and hexane. To locate the minima and maxima on the free-energy surfaces, two-dimensional free-energy maps were constructed for each reaction using the lengths of the two forming CC bonds as the reaction coordinates (Figure 1). Free-energy perturbation (FEP) calculations were performed in conjunction with NPT Metropolis Monte Carlo (MC) simulations at 25 °C and 1 atm. The reactant state was defined by $R_{C1} = R_{C2} = 4.0$ Å, and the free-energy surfaces were flat in this vicinity.

In the present QM/MM implementation, the solute's intramolecular energy is treated quantum mechanically using PDDG/PM3; computation of the QM energy and atomic

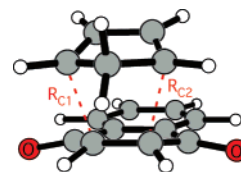


Figure 1. Reaction coordinates, R_{C1} and R_{C2} , for the Diels–Alder reaction between cyclopentadiene and 1,4-naphthoquinone. Illustrated structure is the transition structure from gas-phase PDDG/PM3 calculations ($R_{C1} = R_{C2} = 2.13$ Å).

charges is performed for each attempted move of the solute, which occurred every 100 configurations. For electrostatic contributions to the solute–solvent energy, CM3 charges³¹ were obtained for the solute using PDDG/PM3 calculations with a scaling factor of 1.14. This is augmented with standard Lennard-Jones interactions between solute and solvent atoms using OPLS parameters.³⁰ This combination is appropriate for a PM3-based method as it minimizes errors in computed free energies of hydration.³²

Solute–solvent and solvent–solvent intermolecular cutoff distances of 12 Å were employed based on all heavy atoms of the solute, the oxygens of water and methanol, the central carbon of acetonitrile, and carbon atoms of hexane. If any distance is within the cutoff, the entire solute–solvent or solvent–solvent interaction was included. Quadratic feathering of the intermolecular interactions within 0.5 Å of the cutoff was applied. Total translations and rotations were sampled in ranges that led to overall acceptance rates of about 41–47% for new configurations. FEP windows were run simultaneously on a Linux cluster at Yale and on computers located at the Alabama Supercomputer Center.

The complete basis set method CBS-QB3³³ was also used to characterize the transition structures and ground states in vacuum using Gaussian 03.³⁴ In a recent study, the CBS-QB3 method gave energetic results in the closest agreement to experiment for a set of 11 different pericyclic reactions compared to other *ab initio* and density functional theory methods.³⁵ The CBS-QB3 calculations were used for geometry optimizations and computations of vibrational frequencies, which confirmed all stationary points as either minima or transition structures and provided thermodynamic corrections. The effect of solvent was approximated by subsequent single-point calculations using the conductor-like polarizable continuum model (CPCM)³⁶ and the MP2/6-311+G(2d,p) theory level; default Gaussian 03 dielectric constants of 78.39, 32.63, 36.64, and 1.92 were used for water, methanol, acetonitrile, and hexane.

Results and Discussion

Structures. Geometries for the Diels–Alder reactions in solution were located with the QM/MM/MC calculations by starting from the gas-phase PDDG/PM3 cycloadduct structures and perturbing the two reacting carbon bonds between the diene and the dienophiles to find the transition structures (Figure 1). The endo addition mode was chosen in all cases, and for MVK, the *s-cis* conformation was used. These choices correspond to the preferred transition state from *ab initio* calculations³⁷ as well as experimental stereoselective preferences.^{6,11} All internal degrees of freedom other than

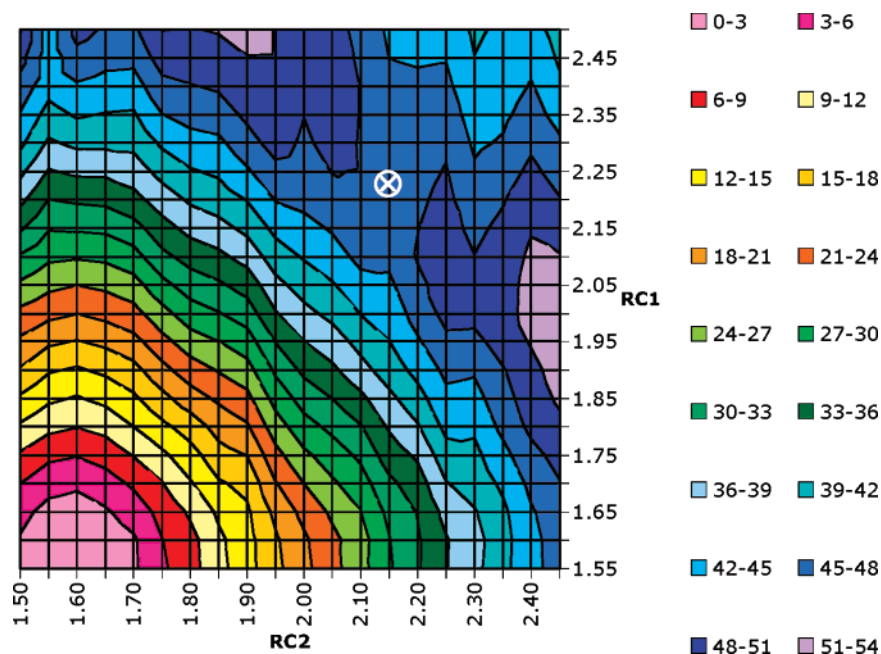


Figure 2. Two-dimensional potential of mean force (free-energy map) for the Diels–Alder reaction between cyclopentadiene and 1,4-naphthoquinone in hexane; \otimes marks the saddle point that represents the solution-phase transition structure. All distances are in Å, and relative free energy is in kcal/mol.

the two reaction coordinates R_{C1} and R_{C2} were fully sampled during the simulations. The initial ranges for R_{C1} and R_{C2} were 1.5–2.5 Å. Each FEP calculation entailed 5 million (M) configurations of equilibration and 10 M configurations of averaging and was computed using increments of 0.05 Å. As an example, the resultant map for the Diels–Alder reaction between cyclopentadiene and 1,4-naphthoquinone in hexane is shown in Figure 2. To locate the critical points more precisely, the regions surrounding the cycloadduct and transition state from the initial maps were explored in increments of 0.01 Å.

The previous one-dimensional PMF approach required a number of intermediate geometries connecting the transition structure to the reactants and cycloadduct for use as initial geometries in the QM/MM simulations.¹ The geometries were derived by following the intrinsic reaction coordinate with *ab initio* calculations and by interpolation. A single reaction coordinate was defined between two dummy atoms located at the midpoint of the reacting carbons in the diene and dienophile; however, this approach provides uncertainty in locating the transition structures. To illustrate the problem, one-dimensional PMF calculations were performed here with three different choices for the reaction coordinate (Figure 3); these calculations used PDDG/PM3, 0.01 Å increments, 5–10 M configurations of equilibration, and 10–30 M configurations of averaging for each FEP window. Without the *ab initio* reference points, the predicted transition structures are found to be highly dependent on the chosen reaction coordinate. Similar results were found for all three Diels–Alder reactions in Scheme 1. The current 2-D approach, in contrast, does not require the *ab initio* calculations and effectively samples all geometries including the transition structure. The geometrical results for the transition structures from the 2-D QM/MM/FEP maps are listed in Table 1 along with the gas-phase CBS-QB3 findings.

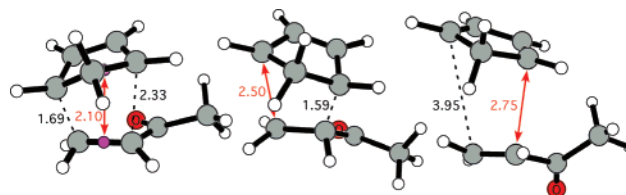


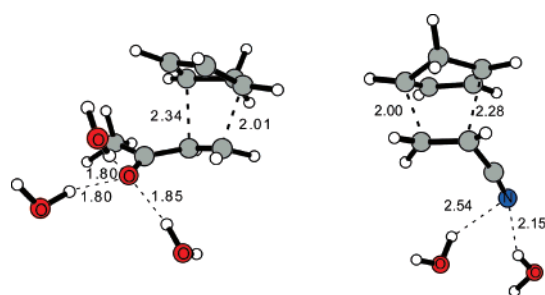
Figure 3. Transition-state geometries for the Diels–Alder reaction between cyclopentadiene and methyl vinyl ketone in water from one-dimensional potential of mean force calculations using three choices for the reaction coordinates (colored in red). All distances are in Å.

The PDDG/PM3 and CBS-QB3 results for the cyclopentadiene plus 1,4-naphthoquinone transition structure in the gas phase are notably similar and reflect a symmetrical, synchronous process. The two bond lengths remain essentially the same in solution, though they are lengthened by about 0.1 Å from the gas-phase values. The asynchronicity Δr is close to the level of uncertainty in the results, ca. ± 0.02 Å. Since the bond-lengthening is similar in all solvents, it can be attributed to the thermal averaging and location of the variational transition state in solution (free-energy saddle-point) as opposed to the conventional transition state from the gas-phase potential energy calculations. The degree of asynchronicity in vacuum and solution becomes significant for the reactions with the unsymmetrical dienophiles, MVK and acrylonitrile. For the gas-phase reactions, the PDDG/PM3 method yields Δr values of ca. 0.1 Å, which underestimates the CBS-QB3 results of 0.6 and 0.4 Å. The latter values are similar to prior results, and the ordering is consistent with the greater capacity for resonance stabilization of negative charge by a keto group than a cyano group. In both cases the computed asynchronicity from the PDDG/PM3-based QM/MM simulations increases in water and methanol to ca. 0.3 Å, while Δr is predicted to be 0.18 Å

Table 1. Computed Bond Lengths (Å) and Asynchronicity ($\Delta r = R_{C2} - R_{C1}$) for the Transition Structures of the Diels–Alder Reactions with Cyclopentadiene at 25 °C and 1 Atm

	gas ^a	CBS-QB3 ^b	water ^c	CH ₃ OH ^c	CH ₃ CN ^c	hexane ^c
1,4-naphthoquinone						
R_{C1}	2.13	2.18	2.22	2.18	2.25	2.22
R_{C2}	2.13	2.17	2.25	2.21	2.19	2.19
Δr	0.0	-0.01	0.03	0.03	-0.06	-0.03
methyl vinyl ketone						
R_{C1}^d	2.08	1.99	2.01	2.02	2.06	
R_{C2}	2.18	2.60	2.34	2.28	2.24	
Δr	0.10	0.61	0.33	0.26	0.18	
acrylonitrile						
R_{C1}^d	2.09	2.05	2.00	2.01		
R_{C2}	2.18	2.45	2.28	2.28		
Δr	0.09	0.40	0.28	0.27		

^a PDDG/PM3 optimizations. ^b Gas-phase optimizations. ^c From the QM/MM/FEP free-energy maps using PDDG/PM3. ^d Shorter distance is with the terminal carbon of the dienophile.

**Figure 4.** Typical snapshots of transition structures for the reactions between cyclopentadiene and methyl vinyl ketone and acrylonitrile in water from the QM/MM/MC simulations. Only water molecules nearest the carbonyl or cyano group are illustrated; distances are in Å.

for the reaction with MVK in the dipolar aprotic solvent acetonitrile. As discussed previously, the greater asynchronicity in the protic solvents can be attributed to enhanced hydrogen bonding at the oxygen of MVK and nitrogen of acrylonitrile in the transition states.^{1,13,16} Typical structures at the transition states from the simulations of these two reactions in water are provided in Figure 4. The number of hydrogen-bonded water molecules increases from two for acrylonitrile to three for methyl vinyl ketone. The hydrogen bonds are also shorter and stronger to the keto oxygen than cyano nitrogen. The hydrogen-bond lengths agree well with results of previous ab initio calculations on the same systems complexed to a single water molecule.¹⁶ The changes in solvation are discussed further below.

Energetics. The computed activation barriers for the Diels–Alder reactions are summarized in Tables 2 and 3. Uncertainties for the free energies are calculated by propagating the standard deviation (σ_i) on each individual ΔG_i . Free energy changes were obtained with statistical uncertainties of only 0.008–0.03 kcal/mol in each window; this implies overall uncertainties in the computed values for ΔG^\ddagger and ΔG_{rxn} of 0.5 and 0.6 kcal/mol, respectively.

Similar to previous QM/MM results for Diels–Alder reactions,^{1,27} the relative free energies of activation are in good agreement with experiment, while the absolute values are overestimated by 10–15 kcal/mol (Tables 2 and 3). As discussed previously,¹ the accuracy of the absolute free

Table 2. Free Energy Changes, ΔG (kcal/mol), at 25 °C for the Diels–Alder Reactions between Cyclopentadiene and the Three Dienophiles Using PDDG/PM3/MM/MC

	water	CH ₃ OH	CH ₃ CN	hexane
1,4-naphthoquinone				
ΔG^\ddagger (calc)	26.0	29.2	30.1	31.1
ΔG^\ddagger (exptl) ^a	16.6	20.0 ^b	20.6	21.6
ΔG_{rxn} (calc)	-20.1	-17.7	-17.2	-15.4
$\Delta G_{\text{retro}}^\ddagger$ (calc)	46.0	46.8	47.3	46.4
methyl vinyl ketone				
ΔG^\ddagger (calc)	32.2	36.4	35.7	
ΔG^\ddagger (exptl) ^a	19.2	21.6 ^b	22.6	
ΔG_{rxn} (calc)	-25.8	-19.3	-19.4	
$\Delta G_{\text{retro}}^\ddagger$ (calc)	58.0	55.7	55.1	
acrylonitrile				
ΔG^\ddagger (calc)	34.0	35.2		
ΔG^\ddagger (exptl) ^c	22.2	23.8		
ΔG_{rxn} (calc)	-16.7	-15.6		
$\Delta G_{\text{retro}}^\ddagger$ (calc)	50.7	50.9		

^a Reference 9. ^b In ethanol. ^c Reference 4; 30 °C.

Table 3. Free Energy of Activation, $\Delta\Delta G^\ddagger$ (kcal/mol), at 25 °C Relative to Water for the Diels–Alder Reactions Using PDDG/PM3/MM/MC

	water	CH ₃ OH	CH ₃ CN	hexane
1,4-naphthoquinone				
$\Delta\Delta G^\ddagger$ (calc)	0.0	3.2	4.1	5.1
$\Delta\Delta G^\ddagger$ (exptl) ^a	0.0	3.4 ^b	4.0	5.0
methyl vinyl ketone				
$\Delta\Delta G^\ddagger$ (calc)	0.0	4.2	3.5	
$\Delta\Delta G^\ddagger$ (exptl) ^a	0.0	2.4 ^b	3.4	
acrylonitrile				
$\Delta\Delta G^\ddagger$ (calc)	0.0	1.2		
$\Delta\Delta G^\ddagger$ (exptl) ^c	0.0	1.6		

^a Reference 9. ^b In ethanol. ^c Reference 4; 30 °C.

energies is adversely affected by several issues. First, the entropy of the reactants is underestimated owing to incomplete sampling and the need for a cratic entropy correction. The latter is small, ca. 3 cal/mol·K, and stems from constraining the reactants to a sphere of 4.0 Å radius with 1 M standard states. A small rate reduction can also be expected from dynamical effects or solvent friction.¹⁹

However, the largest contribution to the overestimation of the activation barriers lies in the quantum mechanics methodology. PDDG/PM3 calculations yield gas-phase activation enthalpies of 34.8, 33.4, and 33.0 kcal/mol for the reactions of cyclopentadiene with naphthoquinone, methyl vinyl ketone, and acrylonitrile, respectively. The prior AM1 calculations yielded similar results, 31.0, 30.1, and 29.8 kcal/mol, while the experimental values are in the 10–20 kcal/mol range.¹ The use of more intensive QM calculations does not guarantee better accuracy. For example, the gas-phase activation barrier for the cyclopentadiene plus MVK reaction has been computed to be in the range of 2–35 kcal/mol for a variety of ab initio and density functional methods; the experimental ΔH^\ddagger is 12.8 kcal/mol in isoctane.^{1,9,37} For reasonable quantitative accuracy for the absolute activation barriers, MP3/6-31G(d), B3LYP/6-31G(d), or higher levels are required. Unfortunately, the use of such ab initio or DFT methods in the current QM/MM approach is impractical in view of the system sizes and need for thorough configurational sampling in the fluid simulations. While the overestimated activation energy barrier for MVK in methanol (Table 3) came primarily from semiempirical QM error, part of the deviation was due to a higher noise level in the Monte Carlo simulations; a reduction of sampling in exchange for an ab initio method would likely increase the noise level further offsetting any accuracy gained from using a high-level QM method. For the previous QM/MM/MC calculations of one-dimensional free-energy profiles, ca. 3.5 million single-point QM calculations were required per profile.¹ The present computations of two-dimensional free-energy surfaces increase the demands to ca. 50 million QM calculations per map. Though there is clearly room for improvement in the semiempirical QM methods, it is notable that the present QM/MM/MC methodology reproduces well the observed rate acceleration in water over methanol and aprotic solvents (Table 3) and provides gas-phase geometries which compare favorably to CBS-QB3 results (Table 1).

Continuum Solvent Models. Previous theoretical work on the Diels–Alder reaction explored implicit solvent models to describe the effects of hydration.^{15,17,20} For example, Cativiela et al. used the self-consistent reaction-field (SCRf) continuum approach coupled with the PM3 semiempirical method to model the reaction between cyclopentadiene and methyl vinyl ketone and found the exo *s*-trans conformation to be the most stable transition structure.¹⁷ This appears to be an artifact of PM3 as the exo *s*-trans transition structure is the least stable of the four endo/exo and *s*-cis/*s*-trans options at the MP3/6-31G(d) level.³⁷ In fact, AM1, PM3, and PDDG/PM3 all yield very similar results for the energetics of the four transition structures, as summarized in the Supporting Information (Table S1). It may also be noted that an abnormally short H–H distance predicted by PM3 between a methylene hydrogen on cyclopentadiene and a hydrogen from methyl vinyl ketone in the exo *s*-trans TS is corrected with PDDG/PM3 (Supporting Information Figure S2). The problem can be traced to the PM3 core repulsion formula (CRF), which incorrectly describes H–H nonbonded interactions in the 1.7–1.8 Å range.²²

Table 4. MP2/6-311+G(2d,p)/CPCM Results for ΔG^\ddagger (kcal/mol) at 25 °C for the Diels–Alder Reactions with Cyclopentadiene^a

	water	CH ₃ OH	CH ₃ CN	hexane
1,4-naphthoquinone				
ΔG^\ddagger (calc)	16.7	17.9	16.5	18.4
ΔG^\ddagger (exptl) ^b	16.6	20.0 ^d	20.6	21.6
methyl vinyl ketone				
ΔG^\ddagger (calc)	19.5	20.9	19.6	
ΔG^\ddagger (exptl) ^b	19.2	21.6 ^d	22.6	
acrylonitrile				
ΔG^\ddagger (calc)	22.1	23.5		
ΔG^\ddagger (exptl) ^c	22.2	23.8		

^a CPCM single point on CBS-QB3 optimized geometries. ^b Reference 9. ^c Reference 4; 30 °C. ^d In ethanol.

The use of the conductor-like polarizable continuum model (CPCM) in conjunction with ab initio single-point energy calculations has been shown to provide good accuracy for computing free energies of hydration for a variety of organic molecules and ions.³⁸ To explore the accuracy of CPCM for the present systems, single-point energy calculations at the MP2/6-311+G(2d,p) level were carried out on the gas-phase CBS-QB3 geometries. The absolute ΔG^\ddagger values computed with this MP2/CPCM approach are in excellent agreement with the experimental values in water (Table 4). However, the continuum model significantly underestimates the effects of switching from water as the solvent to methanol or acetonitrile. Specific changes in hydrogen bonding are expected to be important along the reaction path, and they are not reflected in the continuum treatment. In addition, differences in the structures of the transition states due to solvation (Table 1) are not taken into account when using the gas-phase geometries, though this is expected to be a secondary issue and apparently did not adversely effect the CPCM results for water. However, for acetonitrile and water, the CPCM approach yields nearly identical activation barriers for the cyclopentadiene plus 1,4-naphthoquinone, 16.5 and 16.7 kcal/mol, while the experimental difference is 4.0 kcal/mol.⁹ A similar pattern is seen for the cyclopentadiene plus MVK reaction (Table 4). The QM/MM/MC calculations with their explicit representation of the solvent molecules overcome this limitation (Table 3).

Solvent Effects. The QM/MM/MC simulations capture the key contributions of the medium effects as evidenced by the good agreement between the computed and observed changes in the free energies of activation (Table 3). In water, hydrophobicity-promoted aggregation of diene and dienophile certainly contributes to the lowering of activation barriers for Diels–Alder reactions.^{4–10} The burial of surface area is relatively constant, and the associated rate enhancement is steady at about a factor 10.^{1,13} The key contributor to larger rate increases in protic solvents is preferential stabilization of the transition structures through enhanced hydrogen bonding.^{1,13,14} The present results are fully consistent with these ideas. As an example, the solute–solvent energy pair distributions for the reaction between cyclopentadiene and 1,4-naphthoquinone in water and hexane are presented in Figure 5. The interaction energies are quantified by analyzing the QM/MM/MC results in three representative

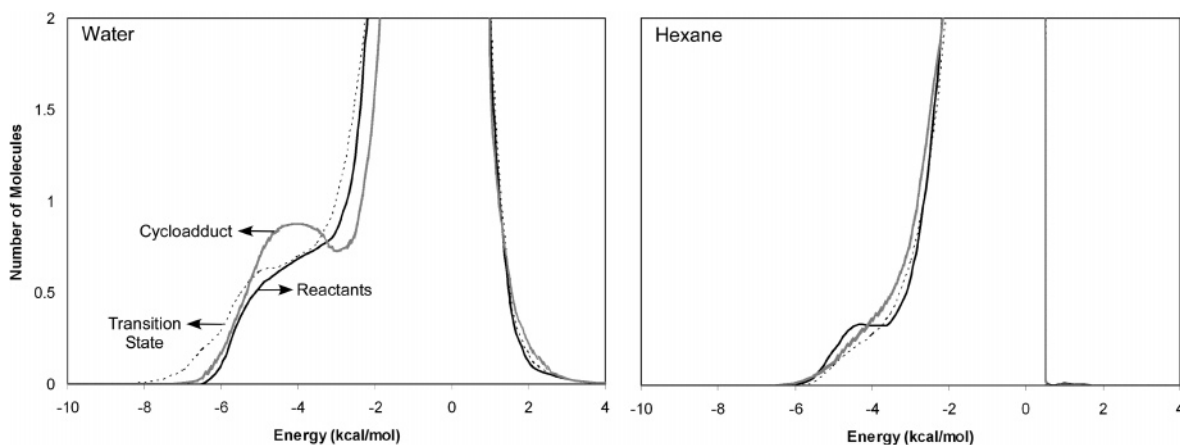


Figure 5. Solute–solvent energy pair distributions for the Diels–Alder reaction between cyclopentadiene and 1,4-naphthoquinone for the reactants, transition state, and cycloadduct in water and hexane at 25 °C. The ordinate records the number of solvent molecules that interact with the solutes and their interaction energy on the abscissa. Units for ordinate are number of molecules per kcal/mol.

FEP windows, near the reactants, transition structure, and cycloadducts. The distributions record the average number of water or hexane molecules that interact with the reacting system and their corresponding energies. Hydrogen bonding between the solute and solvent is found in the left-most region with the most attractive interaction energies. The large bands near 0 kcal/mol is the result of the many solvent molecules located in the outer shells. The critical difference in water is the extension of the distribution for the transition state to low energy. The hydrogen bonding is similar for the reactant and product as they both have keto groups; however, the hydrogen bonds become stronger for the transition state owing to diene to dienophile charge transfer and concomitant enhanced C^+O^- polarization of the carbonyl groups.^{13,16} Consistent with the previous QM/MM simulations,¹ it is also found that the number of solute–water hydrogen bonds increases by about one in going from the reactant to TS in water. If the curves in Figure 5 are integrated to a cutoff energy of -4.0 kcal/mol, the number of water molecules is 2.3, 3.0, and 3.0 for the reactant, TS, and product. If the integration is extended to -3.5 kcal/mol, the corresponding values are 3.1, 3.8, and 3.9.

The solute–solvent structure for the Diels–Alder reactions in water can be further characterized by radial distribution functions, $g(R)$. Hydrogen bonding between the oxygens of naphthoquinone and methyl vinyl ketone and the hydrogens of water, O(dienophile)–H(water), should yield contacts shorter than ca. 2.5 \AA . The corresponding $g_{OH}(R)$ gives the probability of finding a hydrogen of water at a distance R from oxygens of the dienophile. Accordingly, both Diels–Alder reactions show a well-defined first peak centered around 1.9 \AA with minima near 2.5 \AA that reflects the hydrogen bonds (Figure 6). In both cases, the hydrogen bonding is clearly greatest for the transition state. Integration of the first peaks to the minima near 2.5 \AA reveals averages of 3.0 hydrogen bonds between the dienophile oxygens and water molecules for the transition states in both cases. This is well illustrated in Figure 4 for the MVK transition state, while in the Supporting Information, Figure S1, a snapshot for the naphthoquinone transition state illustrates a configuration with two hydrogen bonds for each carbonyl oxygen.

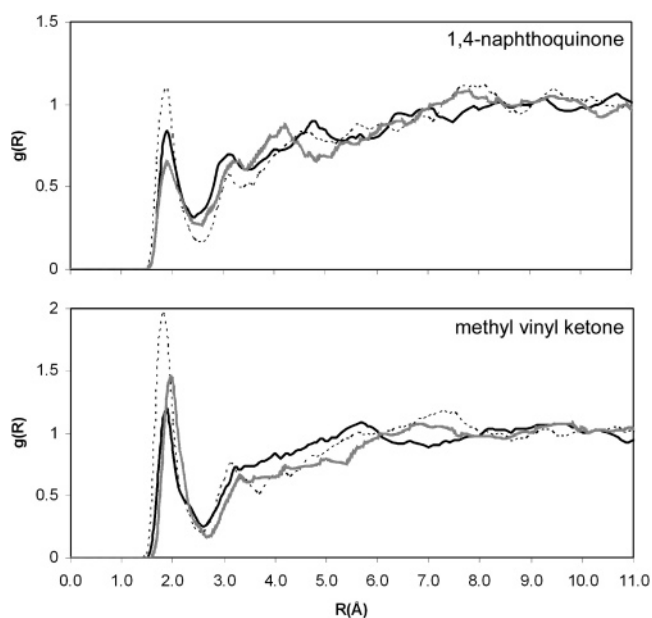


Figure 6. Computed O(dienophile)–H(water) radial distribution functions for the reactions of cyclopentadiene with 1,4-naphthoquinone and methyl vinyl ketone: reactants (solid black curve), transition state (dashed curve), and cycloadduct (solid gray curve) at 25 °C.

The hydrogen bonds are a little longer on average for the naphthoquinone case as reflected in Figures 4, 6, and S1.

Conclusion

QM/MM/MC simulations have been carried out for three Diels–Alder reactions in water, methanol, acetonitrile, and hexane yielding good accord between the computed and observed variations in the free energies of activation. In an advance over the prior related study,¹ two-dimensional free energy surfaces were computed as a function of the lengths of the two forming bonds in a fully automated manner. This avoids potential artifacts and uncertainty in location of transition states associated with the use of single reaction coordinates. The present results also confirm the general view that rate accelerations for such Diels–Alder reactions in

protic solvents arise primarily from enhanced hydrogen-bonding to hydrogen-bond accepting groups in the dienophile. Ab initio MP2/CPCM calculations were also carried out to examine the solvent effects on reaction rates. Excellent results were obtained for the reactions in water; however, the substantial rate enhancements over the aprotic solvents were not reproduced by the continuum methodology. Clearly, a QM/MM/MC approach with the QM at the ca. MP2/6-311+G(d,p) level is most desirable but difficult to achieve with the present level of MC sampling that is required to give acceptable precision for the free-energy surfaces.

Acknowledgment. Gratitude is expressed to the National Science Foundation (CHE-0446920) and the Alabama Supercomputer Center for support of this research.

Supporting Information Available: Illustration of the naphthoquinone transition state in water, a comparison of PM3 and PDDG/PM3 transition structures for the MVK reaction, and a table with gas-phase activation energies for the MVK reaction. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Chandrasekhar, J.; Shariffskul, S.; Jorgensen, W. L. *J. Phys. Chem. B* **2002**, *106*, 8078–8085.
- (2) (a) Acevedo, O.; Evanseck, J. D. *Org. Lett.* **2003**, *5*, 649–652. (b) DeChancie, J.; Acevedo, O.; Evanseck, J. D. *J. Am. Chem. Soc.* **2004**, *126*, 6043–6047. (c) Guimarães, C. R. W.; Udier-Blagovic, M.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2005**, *127*, 3577–3588. (d) Pieniazek, S. N.; Houk, K. N. *Angew. Chem., Int. Ed.* **2006**, *45*, 1442–1445.
- (3) (a) Corey, E. J. *Angew. Chem., Int. Ed.* **2002**, *41*, 1650–1667. (b) Nicolaou, K. C.; Snyder, S. A.; Montagnon, T.; Vassilikogiannakis, G. *Angew. Chem., Int. Ed.* **2002**, *41*, 1668–1698.
- (4) Rideout, D. C.; Breslow, R. *J. Am. Chem. Soc.* **1980**, *102*, 7816–7817.
- (5) (a) Breslow, R.; Maitra, U.; Rideout, D. *Tetrahedron Lett.* **1983**, *24*, 1901–1904. (b) Breslow, R.; Guo, T. *J. Am. Chem. Soc.* **1988**, *110*, 5613–5617. (c) Breslow, R. *Acc. Chem. Res.* **1991**, *24*, 159–164. (d) Sarma, D.; Pawar, S. S.; Deshpande, S. S.; Kumar, A. *Tetrahedron Lett.* **2006**, *47*, 3957–3958.
- (6) Breslow, R.; Maitra, U. *Tetrahedron Lett.* **1984**, *25*, 1239–1240.
- (7) (a) Breslow, R.; Rizzo, C. J. *J. Am. Chem. Soc.* **1991**, *113*, 4340–4341. (b) Blokzijl, W.; Blandamer, M. J.; Engberts, J. B. F. N. *J. Am. Chem. Soc.* **1991**, *113*, 4241–4246. (c) Blokzijl, W.; Engberts, J. B. F. N. *J. Am. Chem. Soc.* **1992**, *114*, 5440–5442. (d) Blokzijl, W.; Engberts, J. B. F. N. In *Structure and Reactivity in Aqueous Solution*; Cramer, C. J., Truhlar, D. G., Eds.; American Chemical Society: Washington, DC, 1994; Vol. 568, pp 303–317. (e) Breslow, R.; Zhu, Z. *J. Am. Chem. Soc.* **1995**, *117*, 9923–9924. (f) Breslow, R.; Connors, R.; Zhu, Z. *Pure Appl. Chem.* **1996**, *68*, 1527–1533. (g) Otto, S.; Bertoncin, F.; Engberts, J. B. F. N. *J. Am. Chem. Soc.* **1996**, *118*, 7702–7707. (h) Wijnen, J. W.; Zavarise, S.; Engberts, J. B. F. N.; Charton, M. *J. Org. Chem.* **1996**, *61*, 2001–2005. (i) Otto, S.; Engberts, J.; Kwak, J. C. T. *J. Am. Chem. Soc.* **1998**, *120*, 9517–9525. (j) Otto, S.; Boccaletti, G.; Engberts, J. B. F. N. *J. Am. Chem. Soc.* **1998**, *120*, 4238–4239. (k) van Mersbergen, D.; Wijnen, J. W.; Engberts, J. B. F. N. *J. Org. Chem.* **1998**, *63*, 8801–8805. (l) Otto, S.; Engberts, J. B. F. N. *J. Am. Chem. Soc.* **1999**, *121*, 6798–6806. (m) Otto, S.; Engberts, J. B. F. N. *Pure Appl. Chem.* **2000**, *72*, 1365–1372. (n) Rispens, T.; Engberts, J. B. F. N. *Org. Lett.* **2001**, *3*, 941–943. (o) Schreiner, P. R. *Chem. Soc. Rev.* **2003**, *32*, 289–296. (p) Wittkopp, A.; Schreiner, P. R. *Chem. Eur. J.* **2003**, *9*, 407–414. (q) Breslow, R. *Acc. Chem. Res.* **2004**, *37*, 471–478. (r) Rispens, T.; Engberts, J. B. F. N. *J. Phys. Org. Chem.* **2005**, *18*, 725–736. (s) Tiwari, S.; Kumar, A. *Angew. Chem., Int. Ed.* **2006**, *45*, 4824–4825. (t) Kleiner, C. M.; Schreiner, P. R. *Chem. Commun.* **2006**, 4315–4317.
- (8) (a) Otto, S.; Blokzijl, W.; Engberts, J. B. F. N. *J. Org. Chem.* **1994**, *59*, 5372–5376. (b) van der Wel, G. K.; Wijnen, J. W.; Engberts, J. B. F. N. *J. Org. Chem.* **1996**, *61*, 9001–9005.
- (9) Engberts, J. B. F. N. *Pure Appl. Chem.* **1995**, *67*, 823–828.
- (10) Wijnen, J. W.; Engberts, J. B. F. N. *J. Org. Chem.* **1997**, *62*, 2039–2044.
- (11) Cativiela, C.; García, J. I.; Gil, J.; Martínez, R. M.; Mayoral, J. A.; Salvatella, L.; Urieta, J. S.; Mainar, A. M.; Abraham, M. H. *J. Chem. Soc., Perkin Trans.* **1997**, *2*, 653–660.
- (12) (a) Sauer, J.; Sustmann, R. *Angew. Chem., Int. Ed.* **1980**, *19*, 779–807. (b) Cativiela, C.; García, J. I.; Mayoral, J. A.; Salvatella, L. *Chem. Rev.* **1996**, *25*, 209–218. (c) Wittkopp, A.; Schreiner, P. R. In *The chemistry of dienes and polyenes*; Rappoport, Z., Ed.; Wiley: New York, 2000; Vol. 2, pp 1029–1088.
- (13) Blake, J. F.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1991**, *113*, 7430–7432.
- (14) (a) Cativiela, C.; García, J. I.; Mayoral, J. A.; Avenoza, A.; Peregrina, J. M.; Roy, M. A. *J. Phys. Org. Chem.* **1991**, *4*, 48–52. (b) Cativiela, C.; García, J. I.; Mayoral, J. A.; Royo, A. J.; Salvatella, L.; Assfeld, X.; Ruiz-lopez, M. F. *J. Phys. Org. Chem.* **1992**, *5*, 230–238. (c) Jorgensen, W. L.; Blake, J. F.; Lim, D.; Severance, D. L. *J. Chem. Soc., Faraday Trans.* **1994**, *90*, 1727–1732. (d) Harano, Y.; Sato, H.; Hirata, F. *J. Am. Chem. Soc.* **2000**, *122*, 2289–2293.
- (15) Ruiz-López, M. F.; Assfeld, X.; García, J. I.; Mayoral, J. A.; Salvatella, L. *J. Am. Chem. Soc.* **1993**, *115*, 8780–8787.
- (16) Blake, J. F.; Lim, D.; Jorgensen, W. L. *J. Org. Chem.* **1994**, *59*, 803–805.
- (17) Cativiela, C.; Dillet, V.; García, J. I.; Mayoral, J. A.; Ruiz-López, M. F.; Salvatella, L. *J. Mol. Struct. (THEOCHEM)* **1995**, *331*, 37–50.
- (18) Furlani, T. R.; Gao, J. *J. Org. Chem.* **1996**, *61*, 5492–5497.
- (19) (a) Pak, Y.; Voth, G. A. *J. Phys. Chem. A* **1999**, *103*, 925–931. (b) Hu, H.; Kobrak, M. N.; Xu, C.; Hammes-Schiffer, S. *J. Phys. Chem. A* **2000**, *104*, 8058–8066.
- (20) Kong, S.; Evanseck, J. D. *J. Am. Chem. Soc.* **2000**, *122*, 10418–10427.
- (21) Blokzijl, W.; Engberts, J. B. F. N. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 1545–1579.
- (22) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601–1622.
- (23) (a) Tubert-Brohman, I.; Guimarães, C. R. W.; Repasky, M. P.; Jorgensen, W. L. *J. Comput. Chem.* **2003**, *25*, 138–150. (b) Tubert-Brohman, I.; Guimarães, C. R. W.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2005**, *1*, 817–823.

- (24) (a) Aqvist, J.; Warshel, A. *Chem. Rev.* **1993**, *93*, 2523–2544. (b) Gao, J. *Acc. Chem. Res.* **1996**, *29*, 298–305. (c) Kaminski, G. A.; Jorgensen, W. L. *J. Phys. Chem. B* **1998**, *102*, 1787–1796.
- (25) Jorgensen, W. L.; Tirado-Rives, J. *J. Comput. Chem.* **2005**, *26*, 1689–1700.
- (26) (a) Acevedo, O.; Jorgensen, W. L. *Org. Lett.* **2004**, *6*, 2881–2884. (b) Acevedo, O.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2005**, *127*, 8829–8834. (c) Acevedo, O.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2006**, *128*, 6141–6146. (d) Acevedo, O.; Jorgensen, W. L. *J. Org. Chem.* **2006**, *71*, 4896–4902.
- (27) Acevedo, O.; Jorgensen, W. L.; Evanseck, J. D. *J. Chem. Theory Comput.* **2007**, *3*, 132–138.
- (28) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (29) (a) Jorgensen, W. L. *J. Phys. Chem.* **1986**, *90*, 1276–1284. (b) Jorgensen, W. L.; Briggs, J. M. *Mol. Phys.* **1988**, *63*, 547–558. (c) Jorgensen, W. L.; Briggs, J. M.; Contreras, M. L. *J. Phys. Chem.* **1990**, *94*, 1683–1686. (d) Briggs, J. M.; Matsui, T.; Jorgensen, W. L. *J. Comput. Chem.* **1990**, *11*, 958–971.
- (30) (a) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236. (b) Price, M. L.; Ostrovsky, D.; Jorgensen, W. L. *J. Comput. Chem.* **2001**, *22*, 1340–1352. (c) Thomas, L. L.; Christakis, T. J.; Jorgensen, W. L. *J. Phys. Chem. B* **2006**, *100*, 21198–21204.
- (31) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **2003**, *24*, 1291–1304.
- (32) Udier-Blagovic, M.; De Tirado, P. M.; Pearlman, S. A.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 1322–1332.
- (33) Ochterski, J. W.; Petersson, G. A.; Montgomery, J. A. *J. Chem. Phys.* **1996**, *104*, 2598–2619.
- (34) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, revision D.01*; Gaussian, Inc.: Wallingford, CT, 2004.
- (35) Guner, V.; Khuong, K. S.; Leach, A. G.; Lee, P. S.; Bartberger, M. D.; Houk, K. N. *J. Phys. Chem. A* **2003**, *107*, 11445–11459.
- (36) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Comput. Chem.* **2003**, *24*, 669–681.
- (37) Jorgensen, W. L.; Lim, D.; Blake, J. F. *J. Am. Chem. Soc.* **1993**, *115*, 2936–2942.
- (38) Takano, Y.; Houk, K. N. *J. Chem. Theory Comput.* **2005**, *1*, 70–77.

CT700078B

The Gradient Curves Method: An Improved Strategy for the Derivation of Molecular Mechanics Valence Force Fields from *ab Initio* Data

T. Verstraelen,^{*,†} D. Van Neck,[†] P. W. Ayers,[‡] V. Van Speybroeck,[†] and M. Waroquier[†]

Center for Molecular Modeling, Ghent University, 9000 Gent, Belgium, and Department of Chemistry, McMaster University, Hamilton, ON L8S 4M1, Canada

Received June 23, 2006

Abstract: A novel force-field development strategy is proposed that tackles the well-known difficulty of parameter correlations arising in a conventional least-squares optimization. In the first step of the new gradient curves method (GCM), continuity criteria are imposed to transform the raw multidimensional *ab initio* training data to distinct sets of one-dimensional data, each associated with an individual energy term. In the second step, the transformed data suggest suitable analytical expressions, and the parameters in these expressions are fitted to the transformed data; that is, one does not have to postulate a priori analytical expressions for the force-field energy terms. This approach facilitates the derivation of valence terms. Benchmarks have been performed on a set of small molecules. The results show that the new method yields physically acceptable energy terms exactly when a conventional parametrization would suffer from parameter correlations, that is, when an increasing number of redundant internal coordinates is used in the force-field model. The generic treatment of parameter correlations in the proposed method facilitates an intuitive physical interpretation of the individual terms in the force-field expression, which is a prerequisite for the transferability of force-field models.

1. Introduction

The development of a molecular mechanics force field based on an *ab initio* parametrization is a tedious task plagued by model selection and parameter correlations, especially when one wants to extend its applicability to a broad range of molecular systems. The final goal of this study lies in the construction of an accurate all-atom zeolite-guest force field that is applicable both to unconstrained bulk zeolite structures and to unconstrained interfaces between zeolite nanoparticles and their environment. It is highly ambitious to assert that such a broad domain of applications can be covered by a single force-field model. Most of the force fields proposed in the literature can be used only for a subset of the applications mentioned above.^{1–7} There are two reasons for

the limited applicability of existing force fields. On the one hand, molecular mechanics models are limited, in general, to a specific domain of application due to the reduction of the full *ab initio* description of a molecule into a set of parametrized analytical energy terms. This failure is inherent to the nature of force-field models. On the other hand, the determination of reliable and transferable parameters for the analytical expressions in a force field is a nontrivial task. The main focus of this paper is the development of a reliable parametrization technique.

Parameter correlations, which are inherent to least-squares parametrization in general, represent the major difficulty in the development of force fields based on *ab initio* data. In the naive approach of an accurate force-field model, a large number of parameters should be introduced to describe all possible types of interactions. The optimization then usually leads to many degenerate solutions; that is, many disparate parameter sets have nearly the same goodness of fit. Only a

* Corresponding author e-mail: Toon.Verstraelen@UGent.be.

[†] Ghent University.

[‡] McMaster University.

very small number of these “good” fits are physically acceptable and transferable to molecules not belonging to the training set. In an attempt to fulfill these requirements, several techniques have been proposed in literature that select a physically meaningful and potentially transferable set of parameters yielding an acceptable goodness of fit.

(i) An intuitive procedure first parametrizes a coarse force field that only contains the most important energy terms, using a traditional least-squares method. Second, the residual error is further reduced by including corrective energy terms whose parameters are optimized without modifying the original coarse force field.¹⁰ This approach keeps the contribution of the corrective energy terms small compared to the coarse force field, but the optimal goodness of fit is not reached. Moreover, only the correlations between parameters in the coarse force field and the corrective energy terms are treated. More generally, a global optimization is divided in smaller piecewise optimizations, to make the parametrization more tractable. A piecewise optimization only considers a subset of variable parameters, for example, the parameters associated with all bond-stretch terms. The global optimum is then approximated with a limited number of iterations in which each subset of parameters is optimized or reoptimized as to give an optimal fit with respect to the training data and the other subsets of parameters that are kept fixed.^{8,9}

(ii) Another procedure avoids degeneracies in the first-order energy terms (i.e., correlations between first-order force constants and reference coordinates) by imposing constraints on their coefficients,² but degeneracies in higher-order terms are neglected.

(iii) The most systematic approach adds quadratic penalty functions to the χ^2 cost function.¹¹ With each parameter, a penalty function is associated that restrains this parameter to a physically acceptable value. This regularization technique is similar to restrained electrostatic potential fitting.¹² Unfortunately, one must choose the weight for each penalty function to be small enough so that the penalty functions only make small contributions to the total cost function but large enough so that the parameters are forced to retain a physically reasonable value. This “weight determination problem” is also ill-conditioned. Essentially, one has just replaced one ill-conditioned problem (parameter fitting) with another one (weight determination).

(iv) Parameter correlations can also be avoided by reducing the number of parameters in a force-field model.¹ One has to select carefully the energy terms that can be omitted and the analytical form of the retained terms. The disadvantage of this approach is that the absence of some molecular interaction terms in the force field will be compensated by biased parameters in the retained terms. Consequently, it is a common practice to exclude the atomic charges from the optimization procedure and to assign formal charges to these atoms instead; this prevents unphysical atomic charges.

(v) The most extreme approach in this comparative overview is represented by the rule-based force fields that do not contain fitted parameters.^{13–15} All parameters are directly derived from semiempirical rules or are estimated

on the basis of common sense. Such force fields sacrifice accuracy to achieve transferability.

Except for the second method, all the techniques mentioned above require additional subjective choices to tackle the problem of parameter correlations: the separation of coarse- and fine-grained components, a vast amount of weight factors, and so forth. Only the third method is truly systematic since it treats all parameter correlations, but it depends on a series of manually tuned weight factors.

This work aims to present a new force-field parametrization procedure—the gradient curves method (GCM)—which is innovative in its concept and which addresses the main concerns raised above. First, the method does not rely on subjective choices, for example, predefined analytical expressions for the energy terms, manually tuned parameters, repetitive parametrizations where at each iteration some parameters are included or excluded, and so forth. Second, the new method treats the problem of parameter correlations in a rigorous way. The only input is a set of *ab initio* training data and a list of the internal coordinates that will be used in the force-field model.

The gradient curves method is designed to extract the maximum amount of information from the *ab initio* training data set. A two-step procedure is used to achieve this objective. The first step encompasses a transformation of the raw multidimensional *ab initio* data into distinct one-dimensional data sets, each associated with a single energy term. During this transformation, a consistent treatment of parameter correlations guarantees a unique and physically acceptable series of transformed data sets. In this context, “physically acceptable” indicates that it is possible to give an intuitive physical interpretation to the individual transformed data sets. The analytical expressions enter the procedure only in the second step, where they can be easily estimated from the transformed data sets and may be modeled with nonlinear parameters without major difficulties.

For several reasons, the present version of the gradient curves method is less appropriate to parametrize long-range interactions. These interactions (i.e., the classical electrostatic and the dispersion interactions) obey well-known physical laws. Therefore, it would be highly inefficient to derive these long-range interactions without relying on their asymptotic behavior during the first step of the new method. Specific parametrization techniques for chemically accurate electrostatic models have already been actively studied during the past decades.^{12,16–18} Due to the enormous computational cost of post-Hartree–Fock *ab initio* calculations that describe dispersion interactions properly,¹⁹ it is more efficient to use such calculations specifically for the parametrization of dispersion interactions.^{20,21}

Most of the ingredients of the gradient curves method are new, but the idea to express a multivariate function in terms of functions depending on a smaller number of variables is frequently applied. We refer to the high dimensional model representation²⁵ (HDMR) which has been applied in several fields, ranging from molecular modeling²⁶ to global atmospheric models.²⁷ This technique guarantees a unique multivariate expansion; that is, it treats parameter correlations, by imposing orthogonality constraints between all the

components in the expansion. HDMR is very efficient when the primary concern is only to reproduce a given set of training data. The end result is an efficient and reliable input–output model. At this point, our focus is different; that is, we would like to ensure that all the distinct energy terms are physically intuitive instead of orthogonal. Less popular black-box approaches where the expansion consists solely of one-dimensional functions^{28,29} are based on Kolmogorov’s solution³⁰ to Hilbert’s 13th problem³¹ and rely on nonsmooth component functions.

The applications in this paper are limited to a set of small molecules such as H₂O, NH₃, and CH₄. For the short-range aspects of interest, this is sufficient to illustrate and benchmark the new method. The aim of these examples is not to obtain transferable force-field parameters for these three molecules but rather to show how the prerequisites for transferability can be met. Additionally, it is not the intention to derive definitive force-field parameters for these three molecules that can be directly tested against experimental data, but we focus on the aspect of how well a reasonable force-field model can simulate a given set of ab initio calculations. We have intentionally generated ab initio training data for these molecules that include a significant portion of the anharmonic part of the potential energy surface, in order to test to what extent the gradient curves method is capable of parametrizing force fields that also reproduce the nonharmonic part of the potential energy surface of the three benchmark molecules. Work is in progress to extend the applicability of the gradient curves method to larger systems, taking into account long-range interactions.

The remainder of this article is organized as follows. In section 2, the new procedure is derived. The benchmark protocol that evaluates the merits of this new procedure is presented in section 3. Section 4 discusses the results obtained by the benchmarks. Finally, conclusions are given in section 5.

2. Gradient Curves Method

2.1. Outline. For the sake of simplicity, we limit ourselves to force fields of the class-I form:

$$E_{\text{FF}} = \sum_{k=1}^K E_k(q_k) \quad (1)$$

where $K > 3N - 6$ and $N =$ the number of atoms. The force-field energy E_{FF} of a molecular geometry is expressed as a sum over functions E_k of only one internal coordinate q_k , where the q_k ’s are not restricted to the $(3N - 6)$ molecular degrees of freedom and may stand for a redundant set of internal coordinates. The redundancy originates from the observation that even a coarse valence force field^{13,14} includes terms for all bond lengths, all bending angles, and some dihedral angles. Force fields that are accurate in the prediction of both structural and vibrational properties have to include cross terms $E_{k_1,k_2}(q_{k_1},q_{k_2})$ in the force-field expression.²² In class-II force fields,²³ this is resolved by adding functions that depend on products of internal coordinates, that is, $q_{k_1}q_{k_2}$. We prefer to label products and other constructions of internal coordinates as new internal

coordinates, which allows us to work with the class-I form in eq 1. This implies that for accurate force fields $K \gg 3N - 6$.

As a consequence of the redundancy, a direct fit of parametrized expressions for the E_k to a set of ab initio training data contains severe parameter correlations even when an abundant amount of training data is available. By selecting one arbitrary set of parameters that minimizes the residual errors, the resulting force field contains energy terms with an unphysical behavior and consequently lacks transferability.^{2,11} Similar considerations about redundant internal coordinates in the theory of molecular vibrations have led to the canonical force-field concept, which is useful for the analysis of vibrational spectra.²⁴

The detailed mathematical derivation of the gradient curves method will be presented in the next subsection. We now continue with a general outline of the method. The training data used in the gradient curves method are the ab initio calculated gradients for M different geometries of a given molecule

$$Y_i^{(m)} = \left(\frac{\partial E_{\text{AI}}}{\partial x_i} \right)_{x=x^{(m)}} \quad (2)$$

where $m = 1 \dots M$ and $x^{(m)}$ is the vector that contains all the Cartesian coordinates of the atoms in geometry m . For an energy surface of the class-I form in eq 1, one factorizes the Cartesian gradient for geometry m according to

$$G^{(m)} = J^{(m)} g^{(m)} \quad (3)$$

where the matrices in expression 3 are defined as

$$\begin{aligned} G_i^{(m)} &= \left(\frac{\partial E_{\text{FF}}}{\partial x_i} \right)_{x=x^{(m)}} \\ g_k^{(m)} &= \left(\frac{\partial E_{\text{FF}}}{\partial q_k} \right)_{q=q^{(m)}} = \left(\frac{dE_k}{dq_k} \right)_{q_k=q_k^{(m)}} \\ J_{i,k}^{(m)} &= \left(\frac{\partial q_k}{\partial x_i} \right)_{x=x^{(m)}} \end{aligned} \quad (4)$$

The convention for matrix notation in this article uses upper indexes to indicate different matrices and lower indexes to identify the matrix elements; for example, $G^{(m)}$ and $g^{(m)}$ are column matrices of dimension $3N$ and K , respectively, whereas $J^{(m)}$ is a rectangular matrix of dimensions $3N \times K$.

Since we want to find a suitable class-I representation of the true (ab initio) energy surface E_{AI} sampled in M geometries, we first identify the Cartesian gradient of the force-field energy in expression 1 with the ab initio training data

$$G_i^{(m)} \equiv Y_i^{(m)} \quad (5)$$

and try to solve the linear system

$$Y^{(m)} = J^{(m)} y^{(m)} \quad (6)$$

for the “ab initio gradient in internal coordinates”, $y^{(m)}$. Due to the redundancy of the coordinates q_k , this equation has many solutions, that is, a particular solution plus an arbitrary

vector from the null space of $J^{(m)}$. Step I of the gradient curves method determines which vector from the null space must be taken for each geometry by an optimization procedure. In other words, the first step defines how the ab initio training data are transformed into one-dimensional data sets of the form $D_k = \{(q_k^{(m)}, y_{k(\text{opt})}^{(m)}) | m = 1 \dots M\}$. Through the identification $y_k^{(m)} \equiv g_k^{(m)}$, or $y_k^{(m)} \equiv (dE_k/dq_k)_{q_k=q_k^{(m)}}$, step II of the gradient curves method consists of proposing a functional form for the derivative of each energy term, (dE_k/dq_k) , based on its corresponding transformed data set, D_k , and the expected asymptotic behavior. Finally, each functional form can be fitted to its corresponding data set with conventional fitting procedures.

The purpose of the transformation in step I of the gradient curves method is to make step II as successful as possible. This means that—for each geometry—the vector from the null space will be taken so as to optimize the continuity conditions of the data sets D_k . In practice, this is achieved by selecting the solutions of eq 6 for all geometries that minimize a cost function, Z , which is a measure for the continuity of the data sets D_k . In this work, continuity is measured by the goodness of fit of a generic high-order polynomial to a set of data points.

Unfortunately, this continuity requirement alone will in general not result in a uniquely defined transformation. In other words, the cost function, Z , as a function of the solutions of eq 6, can have a degenerate minimum. It will be shown in the next subsection that the transformation will always be ill-defined when the number of energy terms, K , is much larger than the number of independent internal degrees of freedom, $3N - 6$. To guarantee a unique minimum, we must introduce additional but subordinate criteria that will select from all the possible transformations to continuous data sets the one solution that corresponds optimally to what we expect from physical intuition. In this work “physical intuition” is interpreted as “having minimal forces along the internal coordinates”. This prescription can be implemented as a least-norm criterion on the y values of the data sets D_k , in addition to the continuity criterion. Formally, such a least-norm criterion is implemented as an extra term in the cost function $Z^* = Z + \epsilon L$, where ϵ is a very small positive number and L is the contribution from the least-norm criterion. For small values of ϵ , the minimum of the new cost function approximately also minimizes the original cost function. This least-norm criterion is also known as zeroth-order regularization, and—as shown in the next subsection—it ensures that the transformation is always uniquely defined.

In order to understand the remainder of this paper, it is not strictly required to read the next subsection which describes the detailed mathematical derivation of the gradient curves method. Nevertheless, it is highly recommended for a deeper understanding, and mandatory when one is interested in implementing or extending the method.

2.2. Detailed Procedure. Since step II is a standard fitting procedure, we now concentrate on the details of step I. The general solution of the linear system (6) is given by

$$y^{(m)} = p^{(m)} + \mathcal{N}^{(m)} s^{(m)} \quad (7)$$

where $p^{(m)}$ is a particular solution, $\mathcal{N}^{(m)}$ is a matrix with orthogonal columns spanning the null space of the Jacobian $J^{(m)}$, and the vector $s^{(m)}$ contains arbitrary coefficients that determine which vector from the null space is added to the particular solution. One can derive the particular solution and the null space of a given linear system through the singular value decomposition algorithm.³²

The coefficients $s^{(m)}$ are fixed by imposing continuity criteria: we select the $s^{(m)}$'s that minimize the sum of squared residual errors, obtained in a linear fit of a set of generic auxiliary functions, $f_n(q_k)$ (e.g., polynomials), to the “ab initio gradient in internal coordinates”, $y_k^{(m)}$,

$$y_k^{(m)\text{fit}} \equiv \sum_n a_n^{(k)} f_n(q_k^{(m)}) \quad (8)$$

The sum of the squared residual errors in the fit to the data set D_k is given by the expression

$$R_k^2 = \sum_m \left(\sum_n a_n^{(k)} f_n(q_k^{(m)}) - y_k^{(m)} \right)^2 \quad (9)$$

In this equation, and in the following analysis, we find it convenient to switch to a notation where the different matrix quantities are labeled by the index of the internal coordinates under scrutiny, k , for example,

$$F_{m,n}^{(k)} = f_n(q_k^{(m)}) \quad \tilde{y}_m^{(k)} = y_k^{(m)} \quad (10)$$

In the revised notation, the sum of squared residuals (using standard manipulations) is

$$R_k^2 = (F^{(k)} a^{(k)} - \tilde{y}^{(k)})^T (F^{(k)} a^{(k)} - \tilde{y}^{(k)}) \quad (11)$$

Minimizing this expression with respect to the expansion coefficients, $a_n^{(k)}$, allows one to discern how well the gradient information can be represented by a continuous function. The least-squares expansion coefficients from eq 8 are given by the expression

$$a_{(\text{opt})}^{(k)} = [F^{(k)T} F^{(k)}]^{-1} F^{(k)T} \tilde{y}^{(k)} \quad (12)$$

and the residual error is

$$\min_{a^{(k)}} R_k^2 = \tilde{y}^{(k)T} [1 - F^{(k)} (F^{(k)T} F^{(k)})^{-1} F^{(k)T}] \tilde{y}^{(k)} = \tilde{y}^{(k)T} C^{(k)} \tilde{y}^{(k)} \quad (13)$$

which is indicative of the continuity of the data set D_k . Note that $C^{(k)}$ projects on the complement of the range of $F^{(k)}$. In analogy to eq 10, we can introduce relabeled matrix quantities

$$\tilde{p}_m^{(k)} = p_k^{(m)} \quad \tilde{\mathcal{N}}_{m',\beta m}^{(k)} = \mathcal{N}_{k,\beta}^{(m)} \delta_{m',m} \quad \tilde{s}_{\beta m} = s_{\beta}^{(m)} \quad (14)$$

in terms of which eq 7 can be rewritten as

$$\tilde{y}^{(k)} = \tilde{p}^{(k)} + \tilde{\mathcal{N}}^{(k)} \tilde{s} \quad (15)$$

This allows a compact expression for the desired cost function, which is a weighted sum of the continuity measures

of all the data sets D_k

$$Z(\tilde{s}) = \sum_k w_k^2 (\min_{q_k} R_k^2) = \sum_k (\tilde{p}^{(k)} + \mathcal{N}^{(k)} \tilde{s})^T w_k^2 C^{(k)} (\tilde{p}^{(k)} + \mathcal{N}^{(k)} \tilde{s}) \quad (16)$$

For the practical applicability of the gradient curves method, the weight factors w_k^2 which convert the R_k^2 's to the dimension of an energy squared should be easy to obtain. A simple physical interpretation of $w_k R_k$ is illustrated in Figure 1. R_k is the RMS error obtained by fitting the auxiliary functions $f_n(q_k)$ to the optimized data set D_k . One obtains a tentative energy term by integrating the fitted function $\sum_n a_n^{(k)} f_n(q_k)$ over the physically relevant interval $[q_k^{(\min)}, q_k^{(\max)}]$. The error accumulated during this integration is equal to $(q_k^{(\max)} - q_k^{(\min)}) R_k$. It always has the dimension of an energy, and it is a quality measure for the energy terms obtained by fitting functional forms for (dE_k/dq_k) to the data sets D_k . Therefore, it is both practical and acceptable to identify the conversion factor w_k with the width of the physically relevant interval of q_k . One can intuitively estimate w_k , or alternatively one can obtain these widths from the geometries in the training set if this training set is generated by a well-behaving and extensive sampling procedure. We have observed that the gradient curves method is insensitive to any reasonable changes in the values w_k , and that it is sufficient to estimate the correct order of magnitude.

The $\tilde{s}_{(\text{opt})}$ that minimizes expression Z can be substituted back into expression 7, after reordering this solution into vectors $s_{(\text{opt})}^{(m)}$. This yields the sets of data points D_k that are optimally continuous and thus slightly scattered around a continuous curve. The minimization of Z makes sure that this scattering is minimal. The selection of a suitable functional form for each E_k is easily accomplished by inspecting the scatter plots of the transformed data sets D_k .

Unfortunately, the solution $\tilde{s}_{(\text{opt})}$ is in general not unique. Since Z is a quadratic expression, only one global minimum exists, although that minimum can still be degenerate. In the case of a degenerate minimum, there is a subspace S that contains all the arguments of Z that yield the minimum value. The dimension of S is equal to the dimension of the null space of the matrix

$$\begin{aligned} \mathcal{H} &= \sum_k \mathcal{N}^{(k)T} C^k \mathcal{N}^{(k)} \\ &= \begin{pmatrix} \mathcal{N}^{(1)} \\ \vdots \\ \mathcal{N}^{(K)} \end{pmatrix}^T \begin{pmatrix} w_1^2 C^{(1)} & & 0 \\ & \ddots & \\ 0 & & w_K^2 C^{(K)} \end{pmatrix} \begin{pmatrix} \mathcal{N}^{(1)} \\ \vdots \\ \mathcal{N}^{(K)} \end{pmatrix} \\ &= \mathcal{N}^T C \mathcal{N} \end{aligned} \quad (17)$$

This matrix is a projection of the singular matrix C on a lower-dimensional space. Note that the matrix \mathcal{N}^T is a nonsquare full-rank matrix by construction. Therefore, a unique solution $\tilde{s}_{(\text{opt})}$ will only be available if the intersection of the range of \mathcal{N}^T and the null space of C is empty.

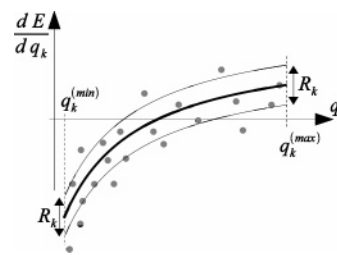


Figure 1. Schematic overview of how the weight factor w_k can be identified with the physically relevant interval of q_k . The fit of the auxiliary functions to the data set D_k is plotted, together with the RMS error on the fitted curve. The error on the integrated curve is approximated by $w_k R_k$.

Since

$$\mathcal{N} \in \mathbb{R}^{KM \times [K - (3N - 6)]M} \quad (18)$$

with N = the number of atoms and $K > 3N - 6$, one should expect \mathcal{H} to be singular when $K \gg 3N - 6$, because then \mathcal{N} is almost a square matrix. As stated in the introduction, an accurate force field always uses many more internal coordinates than independent coordinates. Consequently, for practical applications, a unique solution $\tilde{s}_{(\text{opt})}$ will not be available, no matter how much training data are used. This is a reformulation of the parameter correlations that occur when conventional least-squares fitting is used to parametrize force-field models.

The degeneracy of the cost function gives us the opportunity to select a solution $\tilde{s}_{(\text{opt})}$ that both minimizes Z and that will also result in a physically intuitive model. In this work, the physically intuitive character of a data set will be measured by a least-norm criterion: $\sum w_k^2 \|\tilde{y}^{(k)}\|^2$. The lower this value, the smaller the forces along the internal coordinates in the resulting force-field model, and the more plausible the model. In general, \mathcal{H} is much too large to store in any reasonable computer memory. It is therefore not feasible to perform a singular value decomposition of \mathcal{H} in order to find the least-norm solution in S . Instead, a standard modification to the matrices C^k assures that Z has a unique solution that approximates the least-norm solution:

$$\begin{aligned} Z^*(\tilde{s}) &= Z(\tilde{s}) + \epsilon \sum_k w_k^2 (\tilde{y}^{(k)})^T \tilde{y}^{(k)} \\ &= \sum_k (\tilde{p}^{(k)} + \mathcal{N}^{(k)} \tilde{s})^T w_k^2 (C^{(k)} + \epsilon I) (\tilde{p}^{(k)} + \mathcal{N}^{(k)} \tilde{s}) \end{aligned} \quad (19)$$

where ϵ is a positive constant that is small compared to one. This approximation (of the least-norm solution) becomes exact in the limit of ϵ toward zero, but for numerical applications, the optimal value of ϵ depends on the floating point accuracy. The minimization of Z^* can now be accomplished by a conjugate gradient method and a sparse notation for all the matrices in expression 19.

For reasons of transparency, no restrictions on the functional dependencies of the different internal coordinates have been imposed in the above derivation, and we only considered geometries of a single molecule. When creating realistic force fields, the method is complicated by two practical aspects. First, a useful force field should describe the energy

Table 1. Overview of the a Priori Information Used by the Force-Field Models^a

benchmark model	sets of equivalent internal coordinates	number of elements
Water_default	OH bond lengths	2
	HOH bending angles	1
	HOH span	1
Water_ext1	in addition to the internal coordinates of Water_default (HOH bending cosine) × (OH bond lengths)	2
	(HOH span) × (OH bond lengths)	2
Water_ext2	in addition to the internal coordinates of Water_ext1 (OH1 bond length) × (OH2 bond length)	1
Ammonia_default	NH bond lengths	3
	HNH bending angles	3
	HNH spans	3
Ammonia_ext1	in addition to the internal coordinates of Ammonia_default N(HHH) distance	1
	(N(HHH) distance) × (NH bond lengths)	3
Ammonia_ext2	in addition to the internal coordinates of Ammonia_ext1 (HNH bending cosines) × (NH bond lengths)	6
	(HNH spans) × (NH bond lengths)	6
Methane_default	CH bond lengths	4
	HCH bending angles	6
	HCH spans	6
Methane_ext1	in addition to the internal coordinates of Methane_default (HCH bending cosines) × (CH bond lengths)	12
	(HCH spans) × (CH bond lengths)	12
Methane_ext2	in addition to the internal coordinates of Methane_ext1 (CH bond lengths) × (CH bond lengths)	6

^aAll internal coordinates that belong to the same set are modeled with the same function $E_k(q_k)$ (see eq 1).

dependence of equivalent internal coordinates with the same expression E_k . Second, for a good parametrization, one would sample geometries of different molecules. Because both extensions merely introduce more indexes in the derivation, the same method applies.

3. Benchmark Protocol

The comparison of our novel procedure with conventional force-field parametrizations follows a strict protocol that will be applied on three small benchmark molecules: H₂O, NH₃, and CH₄. The protocol consists of six steps: (i) the generation of training data by a sampling procedure that performs ab initio calculations on a set of different geometries of the given molecule, in addition to the generation of test data by a similar sampling procedure that covers a larger part of the potential energy surface, (ii) the selection of the internal coordinates that are used in the force-field model and the sets of equivalent internal coordinates q_k that are modeled with the same functional dependence, (iii) the gradient curves method presented in this paper, (iv) conventional force-field constructions, using the analytical expressions generated in the former step as input, and (v) the individual validation of each force-field model based on training and test data, and the comparison of all the force-field models.

3.1. Sampling Procedure. The sampling procedure starts with a geometry optimization of the given molecule. The optimized geometry is chosen as the origin of an equidistant $(3N - 6)$ -dimensional grid. The training set is then extended iteratively, by selecting the neighboring grid point of the already calculated geometries that has the lowest estimated

ab initio energy. For each benchmark molecule, 200 training samples and 200 test samples have been generated. The samples in the training data set span an energy range from 0 to 60 kJ mol⁻¹ with respect to the optimized geometry (the origin), while the test samples have a higher upper limit of 100 kJ mol⁻¹. This sampling procedure is only appropriate for small molecules. For larger systems, Monte Carlo sampling should be used. Since our main aim is to test the gradient curves method (while the resulting parameters are of minor importance), a rather low level of theory (DFT/B3LYP) and a small basis set (3-21G*) were used. All ab initio calculations were performed with the MPQC program.³³

3.2. Selection of Internal Coordinates. When developing a force field, one has to select sets of equivalent internal coordinates on which the force-field energy depends. For the gradient curves method, this is the only information that must be given in advance. In this work, nine benchmark force fields are extensively studied, using the different choices of coordinates described in Table 1. The default models for the three molecules use all the interatomic distances and all the cosines of the bending angles, as illustrated in Figure 2. These internal coordinates correspond to those in the well-known Urey–Bradley-type force field, but in this work, no quadratic functional dependencies are imposed. Additionally, two extended force fields are studied for each molecule. The products of internal coordinates in the extended models only contain products of different internal coordinates, and it is always assured that only products of related internal coordinates are considered; for example, a product of two bond

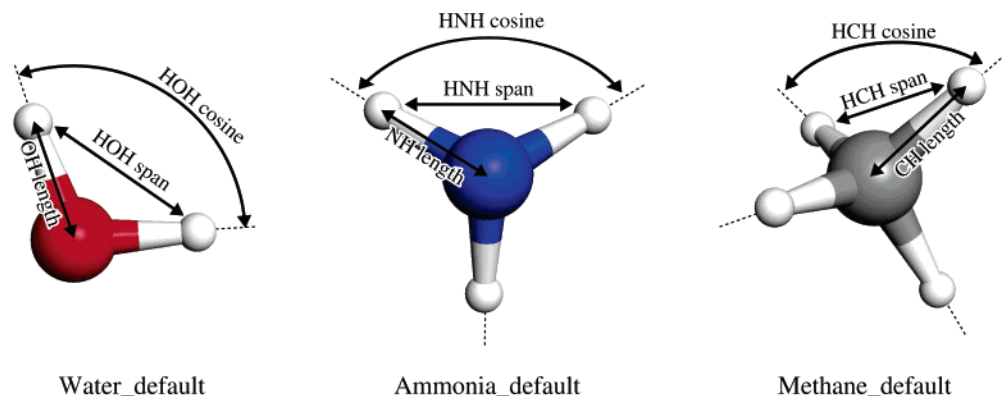


Figure 2. Schematic representation of the internal coordinates in the default models.

lengths will only be considered if the two bonds share exactly one atom. A detailed listing of which products have been used is given in the first section of the Supporting Information. Notice that the term “XYZ span” is defined as “the distance between the atoms X and Z that are both connected to the same atom Y”, and the “A(BCD) distance” is defined as “the distance between an atom A and the plane that is defined by the atoms B, C, and D”. The “XYZ span” is an internal coordinate initially introduced by Urey and Bradley³⁴ in their attempt to derive force fields for small molecules that show an improved reproduction of experimental vibrational frequencies. In their work, it is assumed that the corresponding energy term should be repulsive. We do not make this assumption a priori.

3.3. The Gradient Curves Method. For the auxiliary set of functions $f_n(q_k)$ in eq 8, polynomials up to the 11th order have been used. Two variants of the new gradient curves method are applied: **GCI** is the ill-conditioned variant of the new method, that is, without the least-norm criterion; **GCL** is the variant in which the least-norm correction is applied with $\epsilon = 10^{-6}$.

3.4. The Conventional Methods. In addition to the gradient curves method presented in this work, a series of conventional force-field parametrizations has been performed. They are conventional in the sense that the parameters have been obtained by directly minimizing a well-defined least-squares cost function, although in the literature, additional techniques are used to deal with parameter correlations. The different types of cost functions are listed below. Optionally, a constraint has been applied that compels the force field to reproduce the ab initio Hessian and the zero gradient for the ab initio optimized geometry.

CEU is an unconstrained minimization of the residual error on the energies³⁵

$$Z_{\text{CEU}} = \sum_{m=1}^M [(E_{\text{AI}}^{(m)} - E_{\text{AI}}^{(\text{opt})}) - (E_{\text{FF}}^{(m)} - E_{\text{FF}}^{(\text{opt})})]^2 \quad (20)$$

where the sum over m contains all the molecules in the training set and corrections due to the difference in reference energies of the ab initio and the force-field model have been taken into account.

CEC is a minimization of the residual error on the energies constrained so that the ab initio Hessian and zero gradient are reproduced at the ab initio equilibrium geometry: $Z_{\text{CEC}} = Z_{\text{CEU}}$.

CGU is an unconstrained minimization of the residual error on the gradients³⁶

$$Z_{\text{CGU}} = \sum_{m=1}^M \sum_{i=1}^{3N} \left(\frac{\partial E_{\text{AI}}^{(m)}}{\partial x_i} - \frac{\partial E_{\text{FF}}^{(m)}}{\partial x_i} \right)^2 \quad (21)$$

where i iterates over the Cartesian coordinates.

CGC is a minimization of the residual error on the gradients constrained such that the ab initio Hessian and zero gradient are reproduced at the ab initio equilibrium geometry: $Z_{\text{CGC}} = Z_{\text{CGU}}$.

CCU is an unconstrained minimization of the residual error on the energies and gradients of all the training geometries, as well as the Hessian of the optimized molecule where (i,j) iterates over all the pairs of the Cartesian

$$Z_{\text{CCU}} = W_{\text{CEU}} Z_{\text{CEU}} + W_{\text{CGU}} Z_{\text{CGU}} + W_{\text{CHU}} \sum_{i=1}^{3N} \sum_{j=i}^{3N} \left(\frac{\partial^2 E_{\text{FF}}^{(\text{opt})}}{\partial x_i \partial x_j} - \frac{\partial^2 E_{\text{AI}}^{(\text{opt})}}{\partial x_i \partial x_j} \right)^2 \quad (22)$$

coordinates. The three contributions to the cost function have been weighted to ensure that they have a proportional influence on the obtained parameters. Alternative cost functions that combine ab initio energies, gradients, and/or Hessians have also been reported in the literature for the optimization of force-field parameters.^{1,10,11}

The conventional parametrizations will serve as a reference for the results of the gradient curves method. To guarantee a fair comparison, the analytical expressions used in the conventional methods were obtained with GCL and these expressions only contain linear parameters.

3.5. Validation and Comparison. The generated force-field models are validated with three different criteria. (i) The standard deviation on $E_{\text{FF}} - E_{\text{AI}}$ for all geometries, defined as $\langle [(E_{\text{FF}} - E_{\text{AI}}) - \langle E_{\text{FF}} - E_{\text{AI}} \rangle]^2 \rangle^{(1/2)}$, should be small. The standard deviation is not sensitive to the reference energies of both ab initio and force-field models, in contrast to the root mean square of $E_{\text{FF}} - E_{\text{AI}}$, given by $\langle (E_{\text{FF}} - E_{\text{AI}})^2 \rangle^{(1/2)}$. (ii) The root mean square of $|\nabla E_{\text{FF}} - \nabla E_{\text{AI}}|$ should

be small where ∇ indicates the Cartesian gradient. (iii) At the ab initio optimized geometry, the ratios of the eigenvalues for matching eigenvectors of the force field and of the ab initio Hessian should be near unity.

The third quality criterion is calculated as follows. First, the ab initio Hessian and the force-field Hessian are calculated at the ab initio optimized geometry. The eigenmodes corresponding to the external degrees of freedom are removed by projecting both Hessians on the same basis of $3N - 6$ independent internal coordinates. Then, both projected matrices are diagonalized. The overlap matrix of the corresponding sets of eigenvectors shows clearly which two eigenvectors of the ab initio Hessian and the force-field Hessian correspond with each other. Significant mismatches have not been observed. Finally, the ratios of the eigenvalues associated with the corresponding eigenvectors are calculated.

The quality of the force fields will be compared by the three criteria defined above. In order to assess the robustness of the parametrization, validations i and ii are in addition applied to the set of test data. Finally, we have examined the possibility of giving a physical interpretation to the force-field expressions E_k obtained in the different models.

4. Results and Discussions

To illustrate the usage of the new procedure, we first discuss the three gradient curves generated by GCL applied to Water_default. For each geometry m , the Jacobian, $J^{(m)}$ (see eq 3) is a $N \times K$ matrix or 9×4 matrix of rank $3N - 6 = 3$. The matrix $\mathcal{N}^{(m)}$ describing the null space of such a Jacobian has the dimension $N \times K - (3N - 6)$ or 9×1 . Consequently, given the 200 geometries in the training set, 200 unknown coefficients must be obtained by minimizing the cost function, Z^* . Although there are four distinct internal coordinates in this specific force-field model, the two transformed data sets corresponding to the OH-bond length have been merged into one; that is, their continuity is measured as a whole. Consequently, the data set associated with the bond length consists of 400 data points, while the two others contain 200 data points each. The continuity of each data set is measured by the goodness of fit of an auxiliary 11th-order polynomial. We used generic high-order polynomials to prevent any assumptions about the resulting energy terms being imposed by the continuity criterion; that is, these polynomials will not enforce specific features in the final energy terms. The results are depicted in Figure 3. The data sets D_k obtained by substituting $s_{(\text{opt})}$ into eq 7 are plotted as black crosses. The minimization of Z^* guarantees that these data points lie on continuous curves. The (optimized) auxiliary polynomials that are used to measure the continuity are plotted as dashed lines. Their unphysical asymptotic behavior and the oscillations at the boundaries clarify that the auxiliary polynomials can only be regarded as a measure for the continuity and that they cannot be used as functional forms for the force-field model. In a next step, the analytical form of the derivative of E_k is estimated, on the basis of the data sets. For the energy curve of the OH stretch, a sixth-order polynomial in $1/r_{\text{OH}}$ gives an accurate fit, and the resulting expression has the expected asymptotic

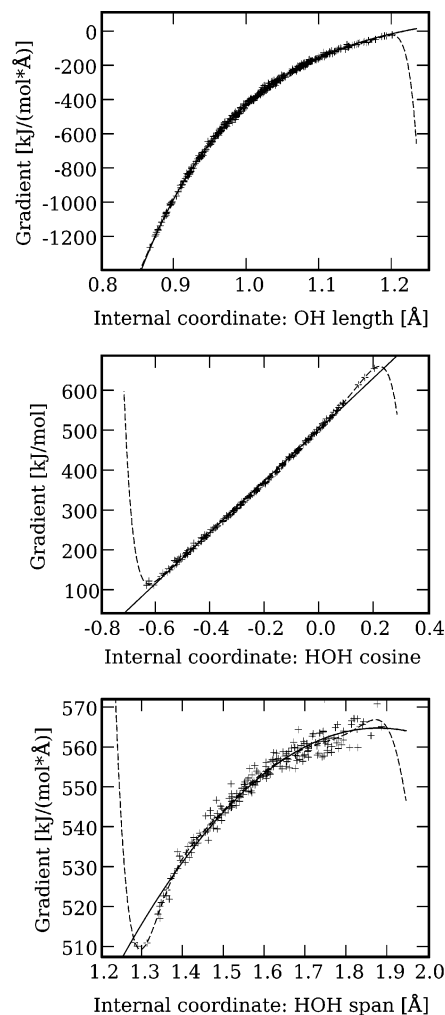


Figure 3. Gradient curves dE_k/dq_k (solid line) obtained for the Water_default model with the GCL method. The black crosses represent the transformed one-dimensional data (see text). The dashed curves are the fitted auxiliary functions for evaluating the continuity criterion.

behavior. The energy curves of the cosine of the bending angle and the interatomic HH distance are estimated to be quadratic and cubic, respectively. Finally, the parameters in the functional forms are optimized using one-dimensional least-squares optimization to the data sets D_k . The resulting curves, dE_k/dq_k , are plotted as solid lines in Figure 3, and the optimized parameters are given in Table 2. The GCL parameters for all nine benchmark models are included in the second section of the Supporting Information.

An overview of the quality criteria for each parametrization is given in Figure 4. The x axis shows the force-field models, and for each force-field model, the different parametrization methods (GCI, GCL, CEU, and so forth) are indicated with different colors. On the y axes, the quality criteria are plotted on a logarithmic scale. Figure 4a and b display respectively the standard deviation on $(E_{\text{FF}} - E_{\text{AI}})$ and the root mean square of $|\nabla E_{\text{FF}} - \nabla E_{\text{AI}}|$ for both training geometries (filled circles) and test geometries (open circles). Figure 4c gives an overview of the validation with the third criterion, represented by the ratios of corresponding Hessian eigenvalues (force-field over ab initio estimates) at the ab initio optimized geometry. It is clear that the overall quality of

Table 2. The Parameters for the Water Default Model Obtained with GCL^a

OH bond length r		HOH bending cosine c		HOH span d	
terms	coefficients	terms	coefficients	terms	coefficients
r^{-1}	-4.608e-01	c	1.931e-01	d	9.898e-03
r^{-2}	5.210e-02	c^2	1.228e-01	d^2	2.933e-02
r^{-3}	3.578e-01			d^3	-2.758e-03
r^{-4}	3.988e-01				
r^{-5}	3.103e-01				
r^{-6}	1.943e-01				

^a The functional form of each energy term, $E_k(q_k) = \sum_{i=1}^I c_i \mathcal{A}_i(q_k)$, is a linear combination of terms listed in the first column of each table. The corresponding coefficients in this linear combination are given in the second column. All parameters are given in atomic units.

the force fields constructed with GCL is comparable to that obtained by the conventional methods. Nevertheless, some interesting discrepancies appear, which will be discussed below.

The ammonia molecule serves as a good example of how to obtain relevant sets of internal coordinates. Initially, the new method was applied on the Ammonia_default model, which only contains the basic internal coordinates: bond lengths, interatomic distances, and bending angles. As shown in Figure 4c, the constructed force field predicts one eigenvalue of the Hessian that deviates significantly from the ab initio value. This eigenvalue corresponds to the inversion of the ammonia molecule. At the transition state of this umbrella inversion, the NH bond length increases due to the alteration from sp^3 to sp^2 hybridization. To describe the inversion more accurately, the extended ammonia model contains two extra sets of internal coordinates: the out-of-plane distance and the products of the out-of-plane distance with the bond lengths. It is striking to observe that the parametrization of the extended ammonia model results in a seriously improved reproduction of the eigenvalues. An attempt was made to avoid the inclusion of more internal coordinates, by constraining the parameters in order to reproduce the ab initio Hessian. This failed drastically for ammonia and methane, since these constraints led to unacceptable errors on the energies and gradients for both training and test data. The corresponding quality criteria falls out of the scope of Figure 4a and b. The performance of CCU in the parametrization of the Ammonia_default model manifestly suffers from the attempt to use information of the ab initio Hessian in the optimization.

The parametrization of ammonia demonstrates that, in some cases, the inclusion of additional redundant internal coordinates in a force-field model is indispensable. This is in agreement with previous studies where it was shown that a pure Urey–Bradley force field, that is, the default model in this work, is not sufficient for an adequate description of the ammonia molecule.^{37,38} Unfortunately, the parametrization of a force field with a high number of internal coordinates ($K \gg 3N - 6$) is sensitive to parameter correlations, and a good treatment of these correlations is required to obtain a useful force field.

In the remainder of this section, we discuss the effect of increasing model complexity. The main effect of the exten-

sions to the force-field models is visible in Figure 4. An improved reproduction of the energies, gradients, and the Hessian is obtained for all methods except the GCI method. This general trend is understandable: the more parameters a model contains, the further a cost function can be optimized. The poor performance of the GCI method for the extended models needs some explanation. Both GCI and GCL yield the same transformed data sets D_k for the default models. For these models, the cost function Z (see eq 16) has a unique solution, even without applying the least-norm correction. This is no longer true for the extended models. In these cases, the minimum of the cost function, Z , becomes highly degenerate, and GCI selects from this minimum an essentially random solution, in the sense that a small change in the training data would imply a very large change in the transformed data sets D_k . On average, such a random solution consists of transformed data sets D_k with very high ranges. This is unacceptable because the absolute errors from fitting energy terms to the transformed data sets (step II of the gradient curves method) scale with the range of D_k . Consequently, the absolute errors shown in Figure 4 are much higher for GCI when applied to the most extended models. We conclude that, of the new methods, GCL is to be preferred over GCI. Both are equivalent for a small number of internal coordinates, but GCL produces superior fits for the more extended models.

The most important trend noticed by increasing the complexity of the model is the behavior of the functions E_k , which is different for GCL as compared to all other methods (i.e., the conventional methods and GCI). Figures 5–7 display all the energy terms E_k obtained with CCU and GCL, for the water, ammonia, and methane molecules, respectively. In these figures, CCU could have been replaced by any other method except GCL without generating significant differences in the global trends. Each row in these figures contains the plots of the energy terms that belong to a specific force-field parametrization, while every column corresponds to a specific set of equivalent internal coordinates. In what follows, we will first discuss the global trends in these figures, and consequently some more specific aspects will be discussed that are not applicable to all the results.

Figures 5a, 6a, and 7a show that CCU yields energy terms E_k with increasing amplitudes, when the force-field model is extended with extra internal coordinates. The conventional methods use the extra degrees of freedom to improve the accuracy, but this improvement is the result of a nonrobust cancellation of high-energy contributions. We have tested an implementation of the conventional methods that applies a singular value decomposition to the design matrix,^{32,39} but a singular value cutoff that gives a good balance between accuracy and reasonable behavior of the functions E_k is not available. The reason is that a least-norm solution in the parameter space is not meaningful since the parameters have different units. A weighted least-norm solution, where the norm of dimensionless weighted parameters is minimal, would be more correct, but then one has to determine a weight value for each parameter as in the work of Ewig et al.¹¹ It is highly remarkable that, as depicted in Figures 5b,

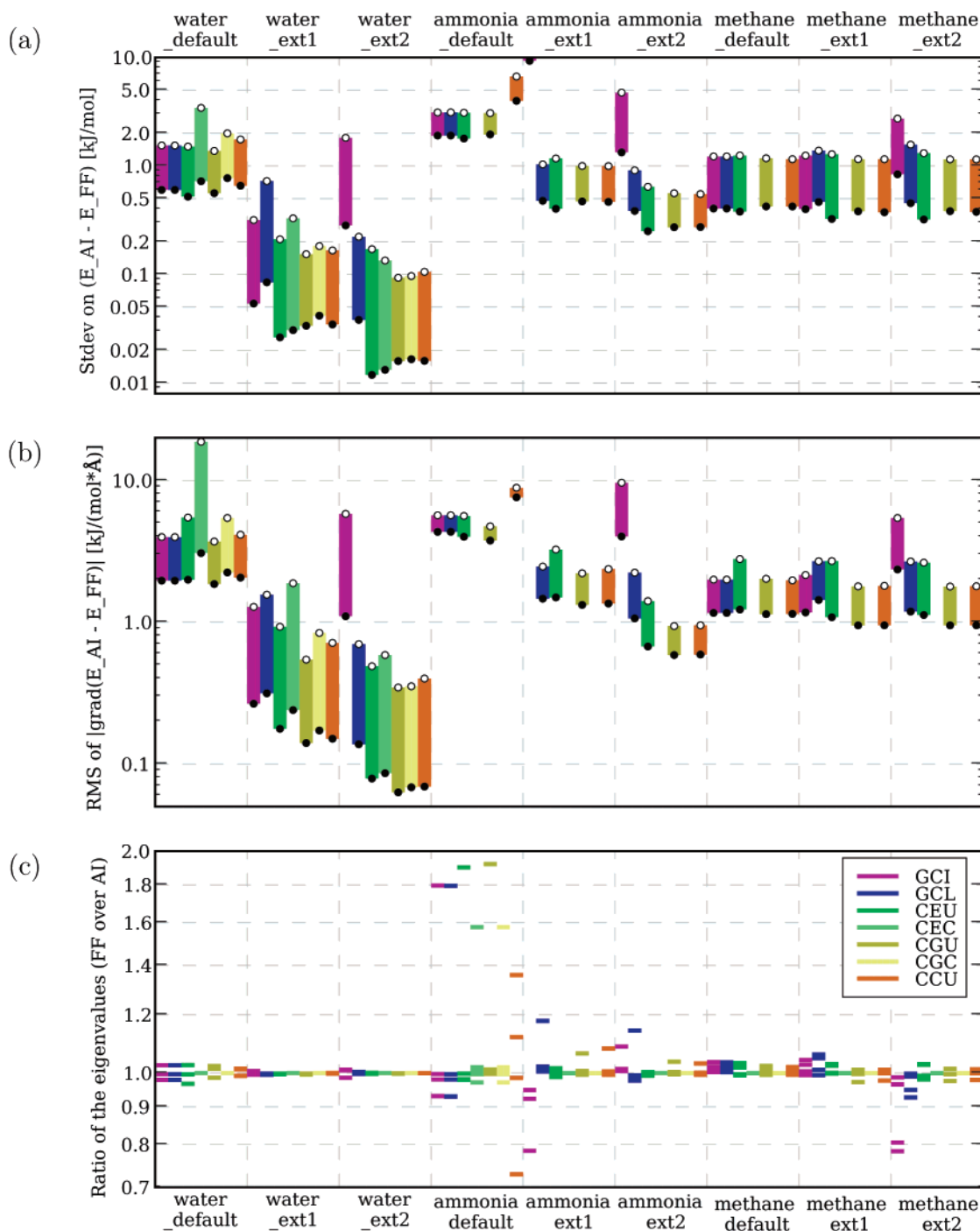


Figure 4. Overview of the force-field validations. Upper figure (a): Standard deviation of the energy differences. Middle figure (b): Root mean square of the gradient differences. Lower figure (c): Ratios of corresponding Hessian eigenvalues (force field over ab initio values), at the ab initio optimized geometry (see text). In parts a and b, the errors for the constrained methods applied on ammonia and methane are too large to fit in the scale of both plots.

6b, and 7b, GCL shows exactly the opposite trend from CCU: the ranges of the functions E_k are reduced in the extended force-field models, and for the ext2 models, it is even possible to give a physical interpretation to the important energy dependencies. For example, the minima of E_k correspond approximately to the internal coordinates of the ab initio optimized geometry. For the terms E_{OH} , E_{NH} , and E_{CH} , even a Morse-like behavior (i.e., the left side of the curve is steeper than the right side) is reproduced. It should be remarked that GCL does not depend on constraints,

model selection, or ad hoc interventions to obtain physical force-field terms. When the gradient curves method will be applied on larger systems, we expect that the absence of cancellation effects will yield transferable and accurate force fields.

In addition to the global trends discussed above, some interesting specific features show up in the results. The most remarkable outcome is that the energy terms for the Ammonia_default model obtained with CCU are very reasonable, and at first instance, this appears to contradict the previous

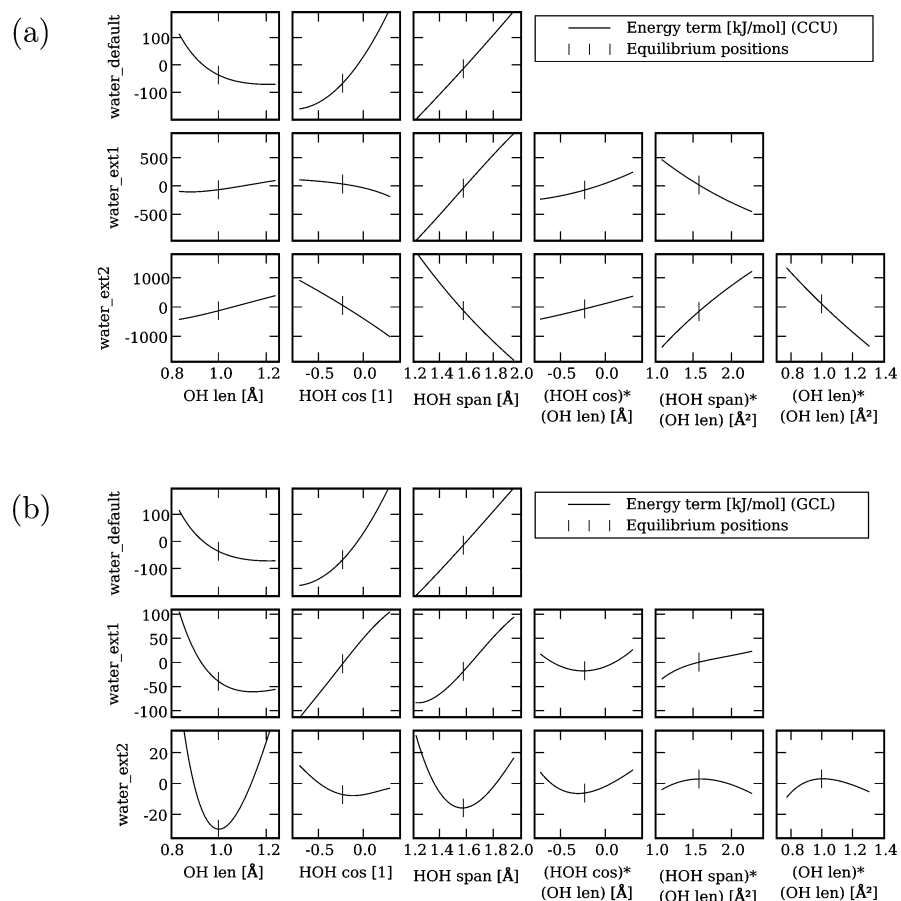


Figure 5. Energy terms E_k for the three different water models, generated (a) by CCU, a conventional parametrization method, and (b) by GCL, the gradient curves method with the least-norm correction. The values of the internal coordinates at the ab initio equilibrium geometry are marked by vertical lines.

paragraph where we stated that reasonable models could only be obtained with GCL. The explanation is that the Ammonia_default model with CCU parameters is indeed reasonable but less accurate compared to other parametrizations of Ammonia_default (see Figure 4a and b). The energy terms for the Ammonia_default model obtained with CGU (see Figure 6c) reveal that the incorporation of the ab initio Hessian of the optimized ammonia geometry in the CCU cost function forces the energy terms of the Ammonia_default model to behave reasonably.

A more subtle result is that the first row of Figure 5a contains virtually the same energy terms as the first row in Figure 5b. Similarly, the first row of both parts a and b of Figure 7 are virtually equal. This situation can be summarized as follows: CCU, a method that does not handle parameter correlations, yields the same energy terms as GCL, a method that does treat parameter correlations. The reason is that none of the parametrization methods in this paper suffer from parameter correlation problems in case of the default models. In the case of the Water_default or the Methane_default model, all the uniquely defined minima of the cost functions of CCU, CEU, CGU, GCL, and GCI even result in the same energy terms. As already discussed above, the different cost functions in the case of Ammonia_default have different—but each of them uniquely defined—optimal parameters. The absence of parameter correlations does however not imply reasonable energy terms. Actually, the sets of equivalent

internal coordinates in the default models are too limited for an accurate reproduction of all the training data with reasonable energy terms. The OH-stretch term represents a repulsive interaction, whereas the energy terms for the HH distance and HOH cosine are both attractive interactions. Correct behavior is obtained only when the three energy terms are combined. For reasons of clarity, we note that the GCL curves in the default models are not supposed to coincide perfectly with the quadratic energy terms in a standard Urey–Bradley parametrization, which are fitted so as to reproduce experimental frequencies.^{40–42} In the present case, the curves are fitted not only to molecular configurations near equilibrium but to higher-energy configurations as well. In fact, when the curves in the first row of Figures 5b and 7b are quadratically expanded around the equilibrium values, a fair correlation with the quadratic force constants and the minima in the work of Kuchitsu and Bartell^{40,41} is observed.

At this point, we have shown how the gradient curves method is able to reconcile the accuracy and the physical interpretation of a force-field model. However, one could wonder how the energy terms, as shown in Figures 5b, 6b, and 7b, evolve when the force-field model is extended with even more additional sets of equivalent internal coordinates (higher-order products, cubic terms, etc.). In the HDMR approach,²⁵ orthogonality criteria are introduced to assert that the addition of higher-order terms does not have any

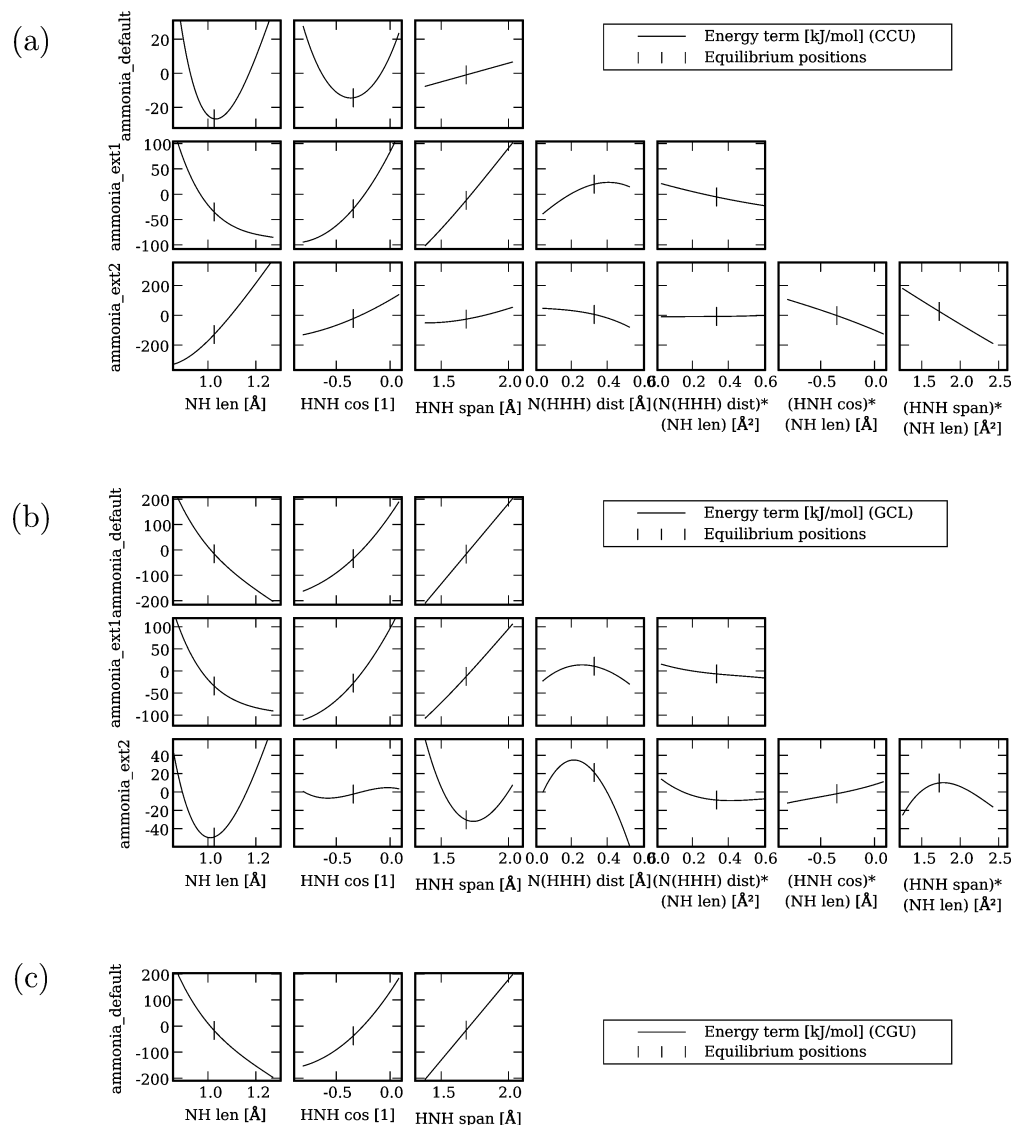


Figure 6. Energy terms E_k for the three different ammonia models, generated (a) by CCU, a conventional parametrization method, (b) by GCL, the gradient curves method with the least-norm correction, and (c) by CGU, a conventional parametrization method that only uses ab initio gradient training data. For part c, only the default model is shown. The values of the internal coordinates at the ab initio equilibrium geometry are marked by vertical lines.

influence on the lower-order terms in the model. The gradient curves method never relies on such orthogonality criteria; for example, this is the reason why the energy terms for the bond length of the three models in Figure 5b are different. There is no “theoretical guarantee” that modifications will not occur when the water model is extended with even more sets of equivalent internal coordinates. Additional energy terms make the continuity criterion extremely degenerate, and in such cases, the least-norm criterion might become an overly naive representation of our physical intuition. Figure 8 demonstrates the behavior of the energy terms for a series of additional extended water models. Similar plots for ammonia and methane are included in the third section of the Supporting Information. Except for the highest-order terms in the two most extended models for the water molecule, the modifications in the energy terms seem to converge once the model is extended enough to show a physically intuitive behavior. The inclusion of second-order derivatives of the ab initio energy in the training data and

more sophisticated criteria for our physical intuition are viable candidates to cure the situation for the two most extended water models and are the subject of our current active research. Nevertheless, one should realize that also these additional measures would suffer from the same defects for the very hypothetical case of even more extended models.

5. Conclusions

This work shows how the gradient curves method can surmount several difficulties that are associated with the development of force fields using least-squares parametrization. Technically, the new method is a two-step procedure: in the first step, continuity criteria and subordinate least-norm criteria are imposed to transform the multidimensional training data into a series of separate one-dimensional data sets, each associated with an energy term of the proposed force field. In this work, the training data are the gradients of the ab initio energy for different molecular geometries.

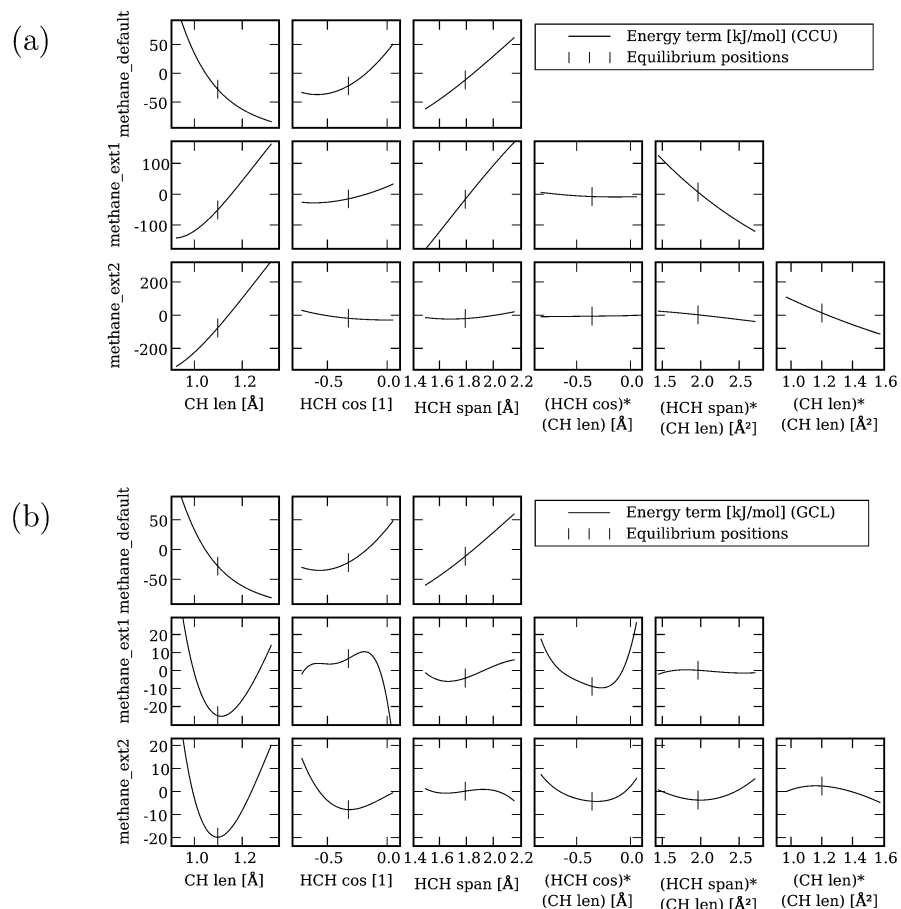


Figure 7. Energy terms E_k for the three different methane models, generated (a) by CCU, a conventional parametrization method, and (b) by GCL, the gradient curves method with the least-norm correction. The values of the internal coordinates at the ab initio equilibrium geometry are marked by vertical lines.

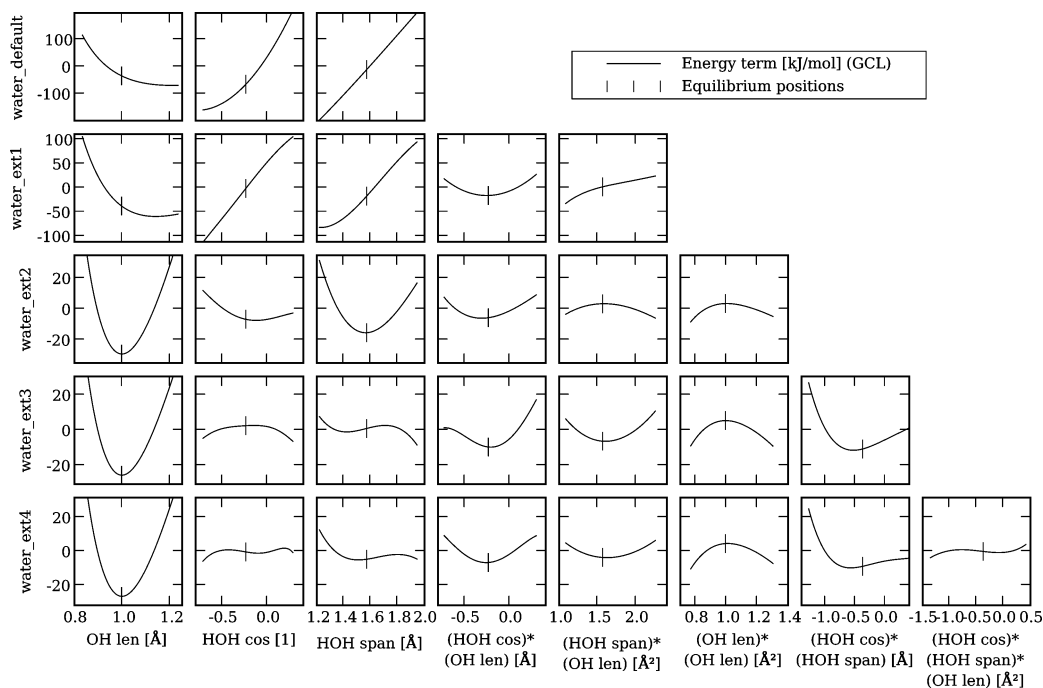


Figure 8. Overview of the energy terms for additional extended water models parametrized with GCL.

During the second step, the derivative of each energy term in the force field is fitted to the corresponding transformed data set.

The gradient curves method has several advantages. Only the internal coordinates have to be defined in advance, instead of a complete analytical ansatz of the force-field model. The

problem of parameter correlations that troubles the conventional force-field development is tackled during the transformation from the multidimensional training data to separate one-dimensional data sets. The continuity and least-norm criteria that are imposed do not only guarantee that the transformed data sets are unique but they also facilitate the physical interpretation of the energy terms fitted to these data sets. In fact, the least-norm criteria express the argument that a plausible force-field model should not contain large derivatives in the energy terms to acquire a marginal increase of accuracy. This prescription fixes all the parameter correlations that originate from the redundancy of the internal coordinates in the force-field model. Once the first step is completed, suitable analytical expressions for the energy terms can be easily proposed after analysis of the transformed data sets and taking into account the expected asymptotic behavior of these energy terms. Because the ability of interpreting the individual force-field terms is known to be a prerequisite for transferable force fields,^{2,11} we expect this method to be very helpful when developing accurate and robust force-field models for larger systems.

The current research mainly focuses on an extended variation of the gradient curves method which is also capable of efficiently deriving the nonbonding interactions from ab initio training data. The primary application on a large system will be the construction of an accurate all-atom zeolite-guest force field. Other active areas include the extension of the gradient curves method to include the ab initio energy and Hessian in the training data, and a more sophisticated formalism for the intuitive character of the energy terms that will eventually supersede the least-norm criterion. We also expect a generalization of the gradient curves method (beyond the scope of force fields) to be useful whenever data parametrization is complicated by parameter correlations and the absence of theoretically supported analytical models.

Acknowledgment. T.V. would like to thank the Flemish organization IWT for its financial support. P.W.A. would like to thank NSERC for funding. V.V.S and M.W. thank the Fund for Scientific Research—Flanders and the Research Board of Ghent University.

Supporting Information Available: A listing of the internal coordinates, the GCL parameters, and an overview of additional extended models. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Schröder, K. P.; Sauer, J. *J. Phys. Chem.* **1996**, *100*, 11043–11049.
- (2) Hill, J.; Sauer, J. *J. Phys. Chem.* **1995**, *99*, 9536–9550.
- (3) Sierka, M.; Sauer, J. *Faraday Discuss.* **1997**, *106*, 41–62.
- (4) Smirnov, K. S.; Bougeard, D. *Chem. Phys.* **2003**, *292*, 53–70.
- (5) Ermoshin, V. A.; Engel, V. *J. Phys. Chem. A* **1999**, *103*, 5116–5122.
- (6) Chandross, M.; Webb, E. B.; Grest, G. S.; Martin, M. G.; Thompson, A. P.; Roth, M. W. *J. Phys. Chem. B* **2001**, *105*, 5700–5712.
- (7) Pascual, P.; Ungerer, P.; Tavitian, B.; Pernot, P.; Boutin, A. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3684–3693.
- (8) Allinger, N. L.; Chen, K.; Lii, J.-H. *J. Comput. Chem.* **1996**, *17*, 642–668.
- (9) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490–519.
- (10) Sun, H.; Rigby, D. *Spectrochim. Acta, Part A* **1997**, *53*, 1301–1323.
- (11) Ewig, C.; Berry, R.; Dinur, U.; Hill, J.; Hwang, M.; Li, H.; Liang, C.; Maple, J.; Peng, Z.; Stockfisch, T.; Thacher, T.; Yan, L.; Ni, X.; Hagler, A. *J. Comput. Chem.* **2001**, *22*, 1782–1800.
- (12) Bayly, C.; Cieplak, P.; Cornell, W.; Kollman, P. *J. Phys. Chem.* **1993**, *97*, 10269–10280.
- (13) Mayo, S.; Olafson, B.; Goddard, W. *J. Phys. Chem.* **1990**, *94*, 8897–8909.
- (14) Rappe, A.; Casewit, C.; Colwell, K.; Goddard, W.; Skiff, W. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.
- (15) Shi, S.; Yan, L.; Yang, Y.; Fisher-Shaulsky, J.; Thacher, T. *J. Comput. Chem.* **2003**, *24*, 1059–1076.
- (16) Mortier, W.; Ghosh, S.; Shankar, S. *J. Am. Chem. Soc.* **1986**, *108*, 4315–4320.
- (17) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141–6156.
- (18) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621–627.
- (19) Chalasiński, G.; Szczesniak, M. *Chem. Rev.* **2000**, *100*, 4227–4252.
- (20) Bordner, A. J.; Cavasotto, C. N.; Abagyan, R. A. *J. Phys. Chem. B* **2003**, *107*, 9601–9609.
- (21) Giese, T.; York, D. *Int. J. Quantum Chem.* **2004**, *98*, 388–408.
- (22) Maple, J.; Dinur, U.; Hagler, A. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 5350–5354.
- (23) Maple, J. R.; Hwang, M. J.; Stockfisch, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T. *J. Comput. Chem.* **1994**, *15*, 162–182.
- (24) Martinez, E.; Lopez, J. J.; Vazquez, J. *J. Mol. Struct.* **2004**, *705*, 141–145.
- (25) Rabitz, H.; Aliş, O.; Shorter, J.; Shim, K. *Comput. Phys. Commun.* **1999**, *117*, 11–20.
- (26) Manzhos, S.; Carrington, T. *J. Chem. Phys.* **2006**, *125*, 084109.
- (27) Shorter, J. A.; Ip, P. C.; Rabitz, H. A. *J. Phys. Chem. A* **1999**, *103*, 7192–7198.
- (28) Gorban, A. N. *Appl. Math. Lett.* **1998**, *11*, 45–49.
- (29) Frisch, H. L.; Borzi, C.; Ord, G.; Percus, J. K.; Williams, G. *Phys. Rev. Lett.* **1989**, *63*, 927–929.
- (30) Kolmogorov, A. *Dokl. Akad. Nauk SSSR* **1957**, *114*, 679.
- (31) Hilbert, D. *Bull. Am. Math. Soc.* **1902**, *8*, 461.

- (32) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. Singular Value Decomposition. In *Numerical Recipes in C: The Art of Scientific Computing*; Cowles, L., Harvey, A., Hahn, R., Eds.; Press Syndicate of the University of Cambridge: Cambridge, United Kingdom, 2002; Chapter 2.6, pp 59–70.
- (33) Janssen, C. L.; Nielsen, I. B.; Leininger, M. L.; Valeev, E. F.; Seidl, E. Y. *The Massively Parallel Quantum Chemistry Program (MPQC)*, version 2.3.0; Sandia National Laboratories: Livermore, CA, 2004.
- (34) Urey, H. C.; Bradley, C. A. *Phys. Rev.* **1931**, *38*, 1969–1978.
- (35) Kramer, G.; Farragher, N.; van Beest, B.; van Santen, R. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1991**, *43*, 5068–5080.
- (36) Ercolessi, F.; Adams, J. *Europhys. Lett.* **1994**, *26*, 583–588.
- (37) Pariseau, M.; Wu, E.; Overend, J. *J. Chem. Phys.* **1962**, *37*, 217–223.
- (38) King, W. T. *J. Chem. Phys.* **1961**, *36*, 165–170.
- (39) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. General Linear Least Squares. In *Numerical Recipes in C: The Art of Scientific Computing*; Cowles, L., Harvey, A., Hahn, R., Eds.; Press Syndicate of the University of Cambridge: Cambridge, United Kingdom, 2002; Chapter 15.4, pp 671–681.
- (40) Kuchitsu, K.; Bartell, L. S. *J. Chem. Phys.* **1962**, *36*, 2460–2469.
- (41) Kuchitsu, K.; Bartell, L. S. *J. Chem. Phys.* **1962**, *36*, 2470–2481.
- (42) Simanouthi, T. *J. Chem. Phys.* **1949**, *17*, 245–248.

CT6002093

Molecular Dynamics and Free Energy Study of the Conformational Equilibria in the UUUU RNA Hairpin

Nan-Jie Deng*

Accelrys Inc., 10188 Telesis Court, San Diego, California 92121

Piotr Cieplak

Burnham Institute for Medical Research, La Jolla, California 92037

Received November 17, 2006

Abstract: A series of molecular dynamics (MD) simulations was performed to elucidate the thermodynamic basis for the relative stabilities of hairpin, duplex, and single stranded forms of the 5'-CGC(UUUU)GCG-3' oligonucleotide. According to a recent NMR study this sequence exhibits dynamic conformational equilibrium in aqueous solution in the vicinity of room temperature. Free energy calculations using the molecular mechanics-Poisson Boltzmann-surface area (MM-PB/SA) approach support a shift in the conformational equilibrium from duplex to hairpin as the temperature is increased from 276 to 300 K, in agreement with the NMR results. The effect of added salt on the relative stabilities of RNA conformers is also reproduced by our calculations. The calculated ΔH^\ddagger for the equilibrium between hairpin and single stranded forms is estimated to be -23.4 kcal/mol, in reasonable agreement with experimental values. Our results reveal that the conformational equilibrium strongly depends on the solute entropy and the electrostatic interactions modulated by added salt. Simulations of hairpin loop conducted at two different temperatures converged to the same lowest energy loop conformation. This conformer is stabilized by favorable van der Waals interactions as a result of U5-U6-U7 base stacking, a hydrogen bond between the U4 base and the phosphate linking U6 and U7, and hydrogen bonds involving the 2'OH groups at U4 and U6. However, the sugar pucker of the four uridines in the lowest energy conformer is different from that reported by a NMR study. While the NMR study found that U5 and U7 adopt the C2'-endo conformation, the simulation results suggests that overall the structure with the U5 and U7 in the C3'-endo conformation is thermodynamically more stable than the structure containing the C2'-endo pucker by approximately 8 kcal/mol. Calculations based on the MM-PB/SA scheme show that although the electrostatic solvation free energy favors the C2'-endo conformation for the U5 and U7 riboses, it is offset by the less favorable intramolecular electrostatic and van der Waals energies. To enhance the conformational sampling, a replica exchange molecular dynamics (REMD) simulation was conducted in a generalized Born (GB) continuum solvent for the hairpin loop. This simulation indicates that the stable loop structure observed in the explicit solvent simulations corresponds to the free energy minimum. It also reveals that while the U4, U5, and U6 sugar rings are predominantly in the C3'-endo conformation, there is considerable variation in the sugar pucker of the U7 ribose ring.

1. Introduction

Hairpin loops are common secondary structural elements in RNA structure. RNA tetraloops contain four nucleotides

forming a loop closed by a Watson–Crick base-paired doubly helical stem. They are widely believed to play an important role in RNA folding and structure.¹ Tetraloops can also participate in RNA tertiary interactions and protein-RNA binding, serving as important molecular recognition sites.

* Corresponding author e-mail: ndeng@accelrys.com.

For example, binding of bacteria signal recognition particle (SRP) to its receptors is influenced by the structure and conformational flexibility of the tetraloop region of the 4.5S RNA.²

The UUUU tetraloop motif is found in the domain IV hairpin loops of yeast *Saccharomyces cerevisiae* SRP RNA.³ Compared with other RNA tetraloops such as the UNCG or GNRA tetraloops which are more frequently occurring and exhibit remarkable stability,⁴ the solution structure of UUUU tetraloops is more dynamic. The structural flexibility of UUUU tetraloop is believed to be important in enhancing the catalytic activity of the hammerhead ribozyme: for example, changing the loop II sequence from UUUU to the thermodynamically more stable GCAA tetraloop reduces the catalytic cleavage rate considerably for the hammerhead ribozyme variants at low Mg²⁺ concentrations.⁵

The melting temperature (T_m) and the thermodynamic parameters of a 12 nucleotide cUUUUg tetraloop have been determined in thermal denaturation experiments.^{6,7} In one study, Tinoco et al.⁶ measured the UV absorbance melting profiles at 260 nm and estimated the T_m to be 60.4 °C. More recently, Proctor et al.⁷ found $T_m = 59.2$ °C, based on measurements of both UV absorbance (at 260 and 280 nm) and NMR spectroscopy. Estimated errors in T_m are ± 1 °C in these studies. A recent NMR study on a 10 nucleotide cgcUUUUg tetraloop revealed interesting conformational equilibria involving duplex form, hairpin loop, and single stranded RNA.⁸ Information on the relative population of secondary structural species is obtained by measuring the temperature dependence of the 1D ¹H NMR spectra. The results indicate that the duplex form is only stable at 3 °C, while the hairpin loop is a predominant species at 27 °C. At higher temperatures the hairpin melts and forms single stranded RNA. The duplex conformation is found to be favored by the addition of salt. The NMR spectroscopy also provides detailed information on the conformations of the ribose moiety in the loop composed of uridines.⁸ The pseudorotation phases and amplitudes of the ribose in the four uridines were obtained by measuring the homonuclear ³J(H,H) coupling constants. The results of sugar puckering modes were corroborated by independent measurements of the heteronuclear ⁿJ(C,H) coupling from cross-correlated relaxation experiment. According to these solution NMR studies, nucleotide U5 and U7 in the loop region are in the C2'-endo conformation, instead of the C3'-endo conformation that is consistent with the canonical A-form of RNA.

These experimental works provide motivation for the present computational investigation in which we conduct force-field based simulations on a cgcUUUUg tetraloop sequence. The goal of the present study is to interpret and understand the conformational properties and the energetics of the short RNA oligonucleotides and to elucidate the origins of the relative stabilities of various secondary structural forms. This is essential for a physical interpretation of the temperature dependent conformational equilibria and may be helpful in understanding the activity of the hairpin loop under different physical conditions.

Molecular dynamics (MD) simulations have been a powerful tool in studying the physical properties of nucleic

acids^{9–11} because of their ability to provide detailed structural and energetic information¹² and to reveal the time dependent conformational transitions. In recent years, a number of MD simulations studies of RNA hairpins have been carried out in explicit solvent and/or by using implicit solvent models.^{13–21} Using MD simulations in explicit solvent and MM-PB/SA approach for postanalysis, Srinivasan et al.¹³ were able to correctly discriminate between different hairpin conformations of a UUCG tetraloop. Hall and Williams^{14,21} conducted MD simulations on a UUCG tetraloop using a GB/SA implicit solvent model. In one of their studies,¹⁴ the simulated RNA hairpin showed some tendency of converting to experimental loop conformation from incorrect structures in nanosecond time scale. In another study,²¹ they examined the effect of substitution of a G-C for a C-G closing base-pair in the UUCG tetraloop by experimental and computational methods. The GB/SA simulation results are consistent with the increased chain flexibility in the UUCG tetraloop closed by the G-C base-pair. Li et al.²⁰ studied the thermal denaturation and refolding of a GAAG tetraloop by running MD simulations in both explicit solvent and a GB/SA solvent. The melting temperature obtained from the simulation is in fair agreement with experiments. Based on the structures sampled during relatively short heating and cooling cycles, they concluded that the folding of the tetraloop proceeds in a stepwise manner. Pande and coworkers^{15,17,18} have conducted a series of large scale MD simulations to investigate the folding dynamics in a GCAA tetraloop. Experimentally, the folding of this hairpin follows two-state kinetics. In their first paper on this hairpin sequence,¹⁵ folding pathways were inferred from the conformations populated along unfolding trajectories generated by multiple high-temperature MD simulations in a GB/SA solvent. The folding/unfolding was shown to be a three-state event, with a globular intermediate state separating the folded and unfolded states in the free energy surface. In the second study,¹⁸ massively parallel simulations were achieved using distributed computing network, which resulted in much more extensive sampling (hundreds of microseconds) of the configuration space. Two types of folding/unfolding pathways, compaction and zipping, are identified. Both pathways are described by two-state free energy surfaces. The importance of explicitly including the solvent and ions is underscored by their most recent simulation study using explicit solvent,¹⁷ in which the picture of two distinct folding pathways was shown to be an oversimplification. Using explicit solvent simulations, significantly greater diversity in the intermediate structures was sampled during the folding process. The folding in explicit ionic solvent was shown to be driven by the collapse of the extended structures, and no simple pathway can be easily distinguished in a highly stochastic conformational search process. Case studies of RNA hairpin simulations include a recent work by Spackova and Sponer¹⁹ on the sarcin-ricin domain motif from 23S (*Escherichia coli*) and 28S(rat) rRNA, which features a GAGA tetraloop region. This tetraloop was found to be the most dynamic part of the RNA motif, and long-residency water molecules were shown to be important in mediating non-Watson–Crick base pairing in the tetraloop.

While these studies demonstrated the usefulness of MD simulation in providing valuable insights into the conformational dynamics of the RNA hairpins at atomic resolutions, limitations in the current generation force field and the use of implicit solvent model can lead to errors in the description of conformational energy surface for noncanonical nucleic acid structures such as single stranded loops. Recently, Fadna and co-workers²² conducted an extensive study on the four-thymidine DNA loops in guanine quadruplexes (G-DNA) using explicit solvent simulation, locally enhanced sampling (LES), and MM-PB/SA free energy calculations. They found that while the force field yields correct characterization of the G-DNA stem structure, it has problems in the description of the flexible loop region interacting with monovalent cations.

The molecular mechanics-Poisson Boltzmann-surface area (MM-PB/SA) method is an approximate approach to the calculation of the free energy difference between two conformational states. This method was first developed to estimate the relative stability of A- and B-form DNA and RNA duplex,²³ the RNA hairpin loops and helices,¹³ and the conformational preferences of A- and B-form DNA in aqueous and mixed solutions.²⁴ It considers the two end points in a configuration space, where the free energy of a structure is decomposed into contributions from the gas-phase molecular mechanics energy, the electrostatic and nonpolar components of the solvation free energy, and the solute entropy. The electrostatic component of the solvation free energy is calculated by solving Poisson–Boltzmann equation (PB) or by using the generalized Born approximation (GB) for a solute in a dielectric medium mimicking water. According to this approach the nonpolar term which accounts for the hydrophobic effect is approximated by a solvent accessible surface area (SA) term. The solute conformational entropy associated with vibrational motions in a single energy well may be obtained by normal-mode calculations. The loss of translational and rotational entropies from molecular association or binding may be estimated using expressions for ideal gas systems or by considering the change in the volume of the conformational space upon association. An ensemble of conformation is collected from snapshots along trajectories generated by running molecular dynamics simulations in explicit or implicit solvents. The MM-PB/SA approach²⁵ and its variants have been widely used to study the problems involving protein–ligand,²⁶ protein–protein,^{27,28} protein–DNA,²⁹ DNA–ligand,³⁰ and RNA–ligand binding.³¹ It has also been applied to analyze the free energies of various conformational species sampled in the 1- μ s folding simulation of villin headpiece³² and the relative stabilities of the Hoogsteen duplex, the reverse Watson–Crick parallel duplex, and the antiparallel Watson–Crick duplex of d(A:T)-based DNA molecules.³³

While the MM-PB/SA analysis can be an effective method in many situations, the simplified treatment of the solvation effects could lead to errors in estimating free energy: (1) The van der Waals interactions between the solute and solvent may be inadequately represented by the surface area SA term, which uses a single surface tension constant for all types of atoms. (2) In the continuum dielectric model

such as PB, the solvent and solute phases are considered as uniform dielectric media, which responds linearly to external electric field. It also assumes that the solvent equilibration around a given solute conformation is complete at every instant. These assumptions may be inadequate for characterizing the atomic nature of the solvent–solute interactions. For example, hydration patterns emerged from explicit solvent simulations indicate that water densities in the vicinity of the nucleic acids structures are far from uniform.^{11,34} The nonuniform distribution of water density should have an impact on the dielectric properties of the solvent phase, and such an effect is not taken into account by the current continuum electrostatics models. (3) The calculation of PB is sensitive to the choice of parameters such as the input atomic radii, the effective dielectric constant of the solute phase, and the definition of dielectric boundary separating the solute and solvent. Since the extent of errors caused by these factors is case dependent, caution is needed when applying the method and interpreting the free energy results.

In the present work, we perform a series of molecular dynamics (MD) simulations on the RNA sequence 5'-CGC-(UUUU)GCG-3' in aqueous solutions initiated from duplex, hairpin loop, and single stranded forms at 276 and 300 K. The MM-PB/SA method is applied to estimate the free energy differences among different conformational species, and the results are compared with experimental data. We investigate the energy profiles of the sugar puckering of the loop uridines and discuss the results in relation to the findings from a NMR study. The structural transition revealed from a 100 ns folding simulation in explicit solvent initiated from single stranded RNA is also analyzed. To improve the efficiency of conformational sampling, we apply replica exchange molecular dynamics (REMD) simulations in a generalized Born continuum solvent described by the GBSW model.^{35,36} We analyze the free energy surface of the hairpin and compare the results with those obtained from the explicit solvent simulations.

2. Methods

Explicit Solvent MD Simulation. The initial structures for the duplex simulations and single stranded RNA simulations were built according to the structural parameters for the standard A-form RNA. One of the initial structures used for the hairpin simulations was built using the coordinates of the cGUAAG tetraloop region in the RNA hammerhead ribozyme³⁷ (PDB ID 1mme) as a template. The four uridines in this initial structure have the C3'-endo sugar pucker, which are different from the sugar puckering modes reported from a NMR study.⁸ One aspect of the present study is to observe reversible conversions in the sugar pucker of the loop residues. To investigate the energy basis for the sugar pucker preferences, a second initial coordinate set was used in which the sugar puckering configuration was modified to resemble those determined by the NMR study, i.e., the U5/U7 are in the C2'-endo and the U4/U6 are in the C3'-endo conformation, respectively. The conformations of the three secondary structural forms are shown in Figure 1(a). The base-pairing and stacking interactions in the hairpin conformation are illustrated in Figure 1(b).

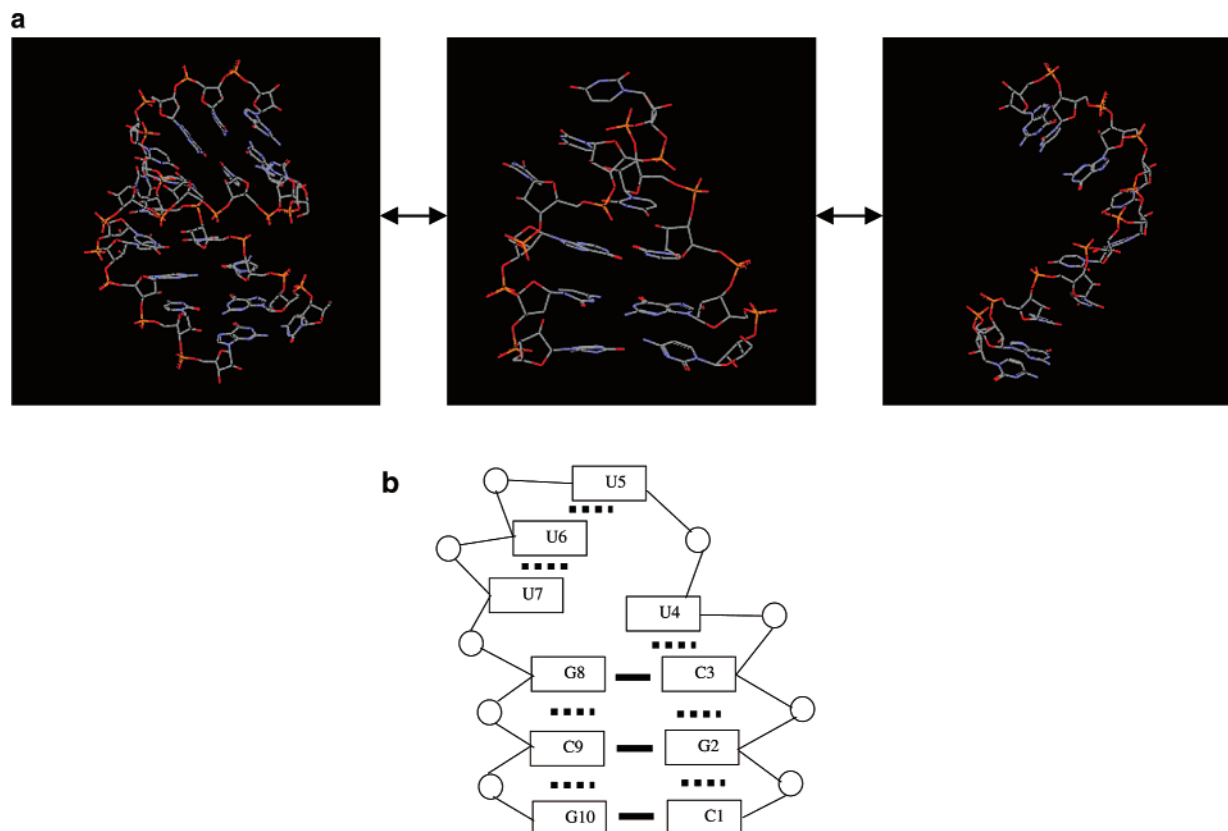


Figure 1. (a) 3D views of the three secondary structural forms. (b) Schematic diagram of the hairpin conformation: circle, phosphate group; solid line, Watson–Crick base pairing; and dotted line, base stacking interaction.

In explicit solvent simulations, sodium counterions were added to neutralize the net charges at each phosphate group. The solute molecule plus counterions were then solvated in a truncated octahedral box containing TIP3P water molecules³⁸ previously equilibrated at 300 K and 1 atm pressure, with the solute atoms separated from nearest walls of the box by 9 Å. Waters within 2.8 Å of solute atoms or counterions were removed from the solvated systems. The MD simulations were performed using the CHARMM³⁹ program versions 29b1 and 30b1. The all-atom CHARMM27 parameter set^{40,41} was used to model nucleic acid molecules. Electrostatic interactions were computed using the particle-mesh Ewald (PME) method.⁴² A switching function between 8.5 Å and 10 Å was used to calculate van der Waals interactions. The Verlet leapfrog integrator was used to solve the equation of motion with an integration step of 2 fs. MD simulations were performed in the NpT ensemble, under atmospheric pressure, using constant pressure/temperature (CPT) dynamics. In each case the following protocol has been applied to minimize and equilibrate the simulated system. Prior to the MD production run, the molecular system was equilibrated according to the following protocol: the solvent alone was first minimized for 1000 steps using the steepest descent method followed by 1000 steps of the adopted basis Newton–Raphson (ABNR) method, with the solute molecules fixed in space. The whole system was then minimized for 1000 steepest descent steps and 1000 ABNR steps without constraints. Following the minimization steps, the system was heated to the desired temperature from 50 K within 40 ps. The system was then equilibrated at the targeted

temperature for 40 ps, before the production run was started. The MD trajectories were saved every 1 ps for analysis.

Free Energy Estimations from MM-PB/SA. As described in the Introduction, a fundamental assumption in the MM-PB/SA approach is that the free energy of a structure may be decomposed into additive contributions: gas-phase molecular mechanics energy $E(\text{gas})$, solvation free energy $G(\text{solv})$, and entropic terms originating from translational, rotational, and vibrational motions of the solute molecules. The free energy of an ensemble of solute structures is obtained by taking the averages over each snapshot in the ensemble generated by molecular dynamics method, i.e.

$$\langle G(\text{tot}) \rangle = \langle E(\text{gas}) \rangle + \langle G(\text{solv}) \rangle - T \langle S(\text{tot}) \rangle \quad (1)$$

The gas-phase energy of the solute is the sum of the internal energy $E(\text{intra})$, electrostatic energy $E(\text{elec})$, and van der Waals energy $E(\text{vdw})$

$$E(\text{gas}) = E(\text{intra}) + E(\text{elec}) + E(\text{vdw}) \quad (2)$$

where the internal $E(\text{intra})$ includes energies from bond, angle, dihedral, and improper dihedral terms. The solvation free energy is further decomposed into electrostatic and nonpolar contributions and is written as

$$G(\text{solv}) = G(\text{PB}) + G(\text{np}) \quad (3)$$

The electrostatic contribution to the solvation free energy $G(\text{PB})$ accounts for the electrostatic interaction between the solute and the polarizable solvent. The latter is treated as a continuum dielectric medium, described by the Poisson–

Boltzmann equation. The nonpolar term $G(\text{np})$ includes the unfavorable hydrophobic contribution and the favorable van der Waals interaction between the solute and the solvent. This term is often assumed to be proportional to the solvent accessible surface area (SASA), i.e.

$$G(\text{np}) = \gamma \times \text{SASA} + \beta \quad (4)$$

The surface tension coefficient γ of $0.5 \text{ cal/mol}\cdot\text{\AA}^2$ and zero for constant β are used in the present study. The atomic radii set for nucleic acids derived by Banavali and Roux⁴³ for the CHARMM27 force-field parameters was used to define dielectric boundary separating solute and solvent. The Poisson–Boltzmann equation was solved by the PBEQ module in the CHARMM program, using a grid spacing of 0.4 \AA and the re-entrant molecular surface for dielectric boundary.

The entropic term $T\langle S(\text{tot}) \rangle$ includes contributions from solute translation, rotation, and vibrational movements:

$$T\langle S(\text{tot}) \rangle = T\langle S(\text{trans}) \rangle + T\langle S(\text{rot}) \rangle + T\langle S(\text{vib}) \rangle \quad (5)$$

The $T\langle S(\text{trans}) \rangle$ and $T\langle S(\text{rot}) \rangle$ are calculated using the standard statistical mechanics expressions for entropies associated with rigid-body translation and rotation in ideal gas.⁴⁴

Two methods, the normal-mode analysis and the quasiharmonic analysis (QH), are used to estimate the solute vibrational entropy. In the normal-mode analysis, we calculate the normal-mode frequencies and use the standard statistical mechanics expression for independent harmonic oscillators⁴⁴ (eq 6) to estimate the entropy due to molecular vibration:

$$TS(\text{vib}) = \sum_{i=1}^{3N-6} \left[\frac{hv_i}{\exp(hv_i/kT) - 1} - kT \ln(1 - \exp(-hv_i/kT)) \right] \quad (6)$$

Here ν_i is the frequency of the i th normal mode, k is the Boltzmann constant, and h is Planck's constant. Prior to normal-mode calculations, the molecular system was fully minimized using a distance dependent dielectric medium with $\epsilon = 4r$, until the root-mean-square energy gradient was less than $10\text{--}5 \text{ kcal/mol}\cdot\text{\AA}$. The VIBRAN utility in the CHARMM program was used to diagonalize the second derivative matrix and generate normal-mode frequencies. The harmonic approximation neglects the effects of anharmonicity in the potential energy surface. The entropy arising from hopping between different energy minima is also not accounted for by the normal-mode analysis.

To semiquantitatively estimate the anharmonic nature of the energy surface, we calculate the vibrational entropy using the quasiharmonic approximation (QH). The QH method and its variants have been widely used for studying the dynamics of proteins and nucleic acids. For example, recently it has been used to explain the thermodynamic basis for the cooperativity of molecular association in a drug–DNA system⁴⁵ and in glycopeptide antibiotics.⁴⁶ The QH approach was first proposed for estimating the configurational entropy

difference for two conformers of a flexible molecule⁴⁷ and has been extended over the years.^{48–52} The original method used internal coordinates because of the singularity of the covariance matrix in Cartesian space.^{47,48,52} Schlitter proposed a heuristic formula which provides an upper bound for the quasiharmonic entropy⁵⁰ while elegantly circumventing the need to use internal coordinates. Recently, Andricioaei and Karplus⁵¹ have shown that the quasiharmonic analysis can be directly performed to calculate entropy from the Cartesian-coordinates covariance matrix, without conversions to internal coordinates or the use of a heuristic formula. The calculation of QH entropy in the present work is based on their analysis.⁵¹ In the QH approach, the configuration probability distribution is a multivariate Gaussian, i.e. $P(x) \propto \exp[-1/2(x-\langle x \rangle)^T \sigma^{-1}(x-\langle x \rangle)]$, where the covariance matrix $\sigma_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle$ is accumulated along a molecular dynamics trajectory. The form of $P(x)$ may be compared with the probability distribution in the canonical ensemble, i.e. $P(x) \propto \exp(-E(x)/kT)$. This comparison suggests that the effective potential is a quadratic function of coordinates $E(x) = 1/2 x^T F x$, with the effective force matrix $F = kT\sigma^{-1}$. Following the standard normal-mode procedure, the mass-weighted covariance matrix $\sigma' = M^{1/2}\sigma M^{1/2}$ is diagonalized to obtain the eigenvalues λ_i and hence the quasiharmonic frequencies $\nu_i = 1/2\pi \sqrt{kT/\lambda_i}$. The quasiharmonic entropy was calculated from the $3N-6$ internal quasiharmonic modes by substituting ν_i into the vibrational entropy formula for harmonic oscillators eq 6.

The VIBRAN module in CHARMM was used to perform the quasiharmonic analysis in this study. The covariance matrix σ_{ij} is accumulated from the molecular dynamics trajectory recorded at time intervals of 1 ps. To focus on configurational entropy due to internal motions, the global translation and rotation were removed from the dynamics trajectory prior to the quasiharmonic analysis, by fitting each coordinate set of the trajectory to a reference structure. The snapshot at the middle point of the trajectory was chosen as a reference structure for the translational and rotational fitting. All the atoms were used in the superposition with mass weighting. We noted that for flexible molecules the rotation and the internal motions cannot be unambiguously separated.

The quasiharmonic method was originally developed for molecular systems with a single highly populated energy well and has been shown to overestimate vibrational entropy in complex systems with multiple energy minima.⁵³ Its utility has been limited by the following factors. First, the QH entropy usually does not converge in nanoseconds MD simulations.^{28,46} Second, the assumption of multivariate normal distribution may be inadequate for describing the dynamics of flexible molecules. In the present study, we employ the QH method to obtain qualitative information regarding the relative chain flexibilities of the three secondary structural forms.

Replica-Exchange Molecular Dynamics. The rugged free energy surface in proteins and nucleic acids often causes MD simulations to be trapped in local minima. The replica exchange molecular dynamics (REMD) and similar methods have been developed to improve the efficiency of barrier crossing and conformational sampling in such situations.^{54,55}

Table 1. RNA Simulations in Explicit Water^a

name	temp (K)	starting conformation	length (ns)
D276	276	duplex	5
L276A	276	hairpin	10
L276B ^b	276	hairpin	10
S276	276	single strand	5
D300	300	duplex	5
L300	300	hairpin	70
S300	300	single strand	100

^a Unless specified, all residues in the initial structures have the C3'-endo sugar pucker. ^b The sugar pucker configuration in the initial structure is as follows: U5 and U7 have the C2'-endo pucker, while U4 and U6 are in the C3'-endo pucker.

In this approach, a simulated replica system has a finite probability of escaping local minima by exchanging its configuration with another replica simulated at a higher temperature. The Metropolis criteria governing the configuration exchange ensures that the Boltzmann distribution as a function of temperature is preserved.

Performing REMD simulation in explicit solvent is possible but computationally expensive, requiring a large number of replicas to cover a small range of temperature values. In the present study, the REMD simulation is performed for RNA hairpin conformer in a continuum solvent described by a generalized Born model with a switching (GBSW) method.^{35,36} Molecular surface was used to approximate a smoothed dielectric boundary. The nucleic acids atomic radii set derived by Banavali and Roux⁴³ for the CHARMM27 parameters was used as the input atomic radii for the GBSW method. No explicit counterions were used in the REMD simulations in continuum solvent, and the salt effect was approximated by assigning a salt concentration of 0.15 M in the GBSW method. The temperature range from 270 to 550 K was covered by 19 exponentially spaced replicas. The configuration exchange between replicas at neighboring temperatures was attempted every 2 ps. The total simulation length was 8.3 ns, and conformations generated in the last 6.3 ns were used in the analysis. The integration time step was 2 fs, and the bond lengths involving hydrogen atoms were constrained by the SHAKE method.⁵⁶ The temperature was kept at the constant value by the Nose-Hoover method.⁵⁷

3. Results and Discussion

To probe the conformational dynamics of the UUUU tetraloop 5'-CGC(UUUU)GCG-3', we performed a series of MD simulations at 276 and 300 K, initiated from duplex, hairpin, and single stranded forms. The conditions of these simulations are described in Table 1. The main results of the free energies obtained by postprocessing the molecular dynamics trajectories using the MM-PB/SA method are collected in Tables 2–5 and in Table S1 of the Supporting Information.

The free energy results presented here indicate that at low temperatures the duplex and hairpin are in equilibrium. Higher temperature destabilizes the duplex, and the hairpin become the dominant form at 300 K. Increasing the salt concentration was found to stabilize the duplex form, as would be expected from increased charge screening. Our results obtained from the normal-mode analysis and quasi-

Table 2. Free Energy Differences: (a) $T = 276$ K^{a,h} and (b) $T = 300$ K^a

	duplex - hairpin	hairpin - sst	duplex - sst ^b
(a) $T = 276$ K			
$\Delta \langle E(\text{intra}) \rangle$	0.4	-1.7	-1.3
$\Delta \langle E(\text{elec}) \rangle$	709.0	51.6	760.6
$\Delta \langle E(\text{vdw}) \rangle$	3.0	-18.6	-15.6
$\Delta \langle E(\text{PB}) \rangle^e$	-711.5	-55.2	-766.7
$\Delta \langle E(\text{total_elec}) \rangle^c$	-2.5	-3.6	-6.1
$\Delta \langle G(\text{np}) \rangle$	-0.6	-1.8	-2.4
$\Delta \langle G(\text{MM-PB/SA}) \rangle^d$	0.3	-25.7	-25.4
$-T\Delta \langle S(\text{vib}) \rangle^f$	-3.4	1.6	-1.8
$-T\Delta \langle S(\text{trans}) \rangle$	6.2	0.0	6.2
$-T\Delta \langle S(\text{rot}) \rangle$	5.6	0.4	5.9
$\Delta \langle G(\text{tot}) \rangle^g$	8.6	-23.7	-15.1
(b) $T = 300$ K			
$\Delta \langle E(\text{intra}) \rangle$	1.3	6.5	7.8
$\Delta \langle E(\text{elec}) \rangle$	784.6	6.9	791.5
$\Delta \langle E(\text{vdw}) \rangle$	2.7	-14	-11.3
$\Delta \langle E(\text{PB}) \rangle^e$	-774.3	-22.3	-796.6
$\Delta \langle E(\text{total_elec}) \rangle^c$	10.3	-15.4	-5.1
$\Delta \langle G(\text{np}) \rangle$	-1.1	-1.1	-2.2
$\Delta \langle G(\text{MM-PB/SA}) \rangle^d$	13.2	-24.0	-10.8
$-T\Delta \langle S(\text{vib}) \rangle^f$	-3.7	0.3	-3.4
$-T\Delta \langle S(\text{trans}) \rangle$	6.7	0.0	6.7
$-T\Delta \langle S(\text{rot}) \rangle$	6.1	0.2	6.3
$\Delta \langle G(\text{tot}) \rangle^g$	22.3	-23.5	-1.2

^a Energies are in kcal/mol, per RNA strand. ^b The label sst stands for single stranded form. ^c The total electrostatic energy $E(\text{total_elec})$ equals the sum of solute electrostatic energy $E(\text{elec})$ and the electrostatic solvation free energy $E(\text{PB})$, i.e. $\langle E(\text{total_elec}) \rangle = \langle E(\text{elec}) \rangle + \langle E(\text{PB}) \rangle$. ^d $\langle G(\text{MM-PB/SA}) \rangle = \langle E(\text{gas}) \rangle + \langle G(\text{sol}) \rangle$. See also eqs 2 and 3 in the text. ^e Salt concentration equals to 0.15 M. ^f $-TS(\text{vib})$ is the vibrational entropy contribution as computed by the normal-mode analysis. ^g $G(\text{tot}) = G(\text{MM-PB/SA}) - TS(\text{trans}) - TS(\text{rot}) - TS(\text{vib})$. ^h The results for the hairpin simulation were calculated using the L276A trajectory described in Table 1.

Table 3. $G(\text{MM-PB/SA})$ Free Energy Differences as a Function of Salt Concentration^a

salt concn (M)	duplex - hairpin	hairpin - sst ^b	duplex - sst ^b
0.15	0.3	-25.7	-25.4
0.01	4.6	-25.3	-20.7
0	8.6	-25.1	-16.5

^a Calculated at $T = 276$ K. Units are in kcal/mol. ^b Label sst stands for single stranded RNA.

Table 4. Vibrational Entropy Contribution $-TS(\text{QH})$,^a as Calculated by Quasiharmonic Analysis with Different Trajectory Lengths at 276 K

length (ns)	duplex	hairpin	sst ^b	duplex - hairpin	hairpin - sst ^b
1	-214.5	-206.5	-238.7	-8.0	32.2
2	-232.6	-215.6	-256.5	-17.0	40.9
3	-243.6	-229.6	-268.4	-14.0	38.8
4	-250.7	-234.0	-291.8	-16.7	57.8
5	-260.0	-234.4	-303.0	-25.6	68.6

^a Units in kcal/mol. ^b Label sst stands for single stranded RNA.

harmonic analysis suggest that the duplex form exhibits larger vibrational entropy than the hairpin. The conformational behavior of the UUUU tetraloop may be compared with that of the more common UUCG tetraloop.¹³ Using the MM-

Table 5. Effect of Loop Residues U5/U7 Sugar Pucker on the $G(\text{MM-PB/SA})$ Free Energy Observed in the L276B Trajectory^a

Trajectory ^a	$E(\text{intra})$	$E(\text{elec})$	$E(\text{vdw})$	$E(\text{PB})$	$E(\text{total_elec})$	$G(\text{np})$	$G(\text{MM-PB/SA})$
2–4 ns	683.6	−150.0	44.0	−1687.9	−1837.9	12.5	−1097.8
7.5–10 ns	680.9	−153.6	35.3	−1680.8	−1834.4	12.6	−1105.6

^a During the L276B simulation, both U5 and U7 are in the C2'-endo conformation between 2 and 4 ns. Transitions to the C3'-endo pucker occur at 4 ns for the U7 residue, and at 6.75 ns for the U5 residue. The whole structure remains in the stable hairpin loop conformation from 7.2 ns until the end of the trajectory. See also Figures 6–8.

PB(GB)/SA approach, Srinivasan et al. calculated the free energy of the duplex/hairpin for a 12-nucleotide UUCG tetraloop at a single temperature (300 K) and found that the hairpin was slightly more stable than the duplex at 0.1 M salt.¹³ The solute vibrational entropy was calculated using the normal-mode analysis. The vibrational entropy change $T\Delta S(\text{vib})$ was found to favor duplex over hairpin by -8.7 kcal/mol at 300 K, which is somewhat larger than the $T\Delta S(\text{vib})$ of -3.7 kcal/mol for the duplex-hairpin conversion in the 10-nucleotide UUUU tetraloop in the present study. The overall duplex/hairpin equilibrium in RNA tetraloops appears to be not strongly sensitive to the details of the loop sequence.

Free Energy Results: Neglecting Solute Entropy Contributions. The estimation of the solute entropy contribution to conformational change remains a difficult aspect in the end-point free energy methods, such as the MM-PB/SA approach.⁵⁸ Uncertainties may arise from the simple approximations used to estimate separately the translational, rotational, and vibrational entropic contributions. In several MM-PB/SA applications, only the vibrational entropy term is estimated, usually based on the normal-mode analysis. In the following, we first analyze the free energy contributions excluding the solute entropic part and later discuss the influence of entropy on the conformational equilibrium.

Table 2 summarizes the results of free energy difference calculated at 0.15 M salt concentration using the trajectories described in Table 1 (see also Table S1, Supporting Information). In each case, the calculated free energy components have been averaged over 200 snapshots for the last 2 ns of the trajectory and reported as values per single chain of RNA.

To verify the convergence of the results, we have also calculated the free energies for the last 4 ns of the simulation and using 2000 snapshots from the trajectories (Tables S2 and S3 in the Supporting Information). While the calculated $\Delta G(\text{tot})$ can differ by up to 4 kcal/mol using different trajectory data sets, these variations in $\Delta \langle G(\text{tot}) \rangle$ are still small compared with the absolute values of $\Delta \langle G(\text{tot}) \rangle$, which indicate that the calculated free energy components (excluding entropy contribution, as discussed below) have converged during the simulations.

Neglecting the solute entropic contributions, the total free energy for each conformational state is represented by the term $G(\text{MM-PB/SA}) = E(\text{gas}) + G(\text{solv})$ in Table S1 of the Supporting Information. As shown in Table 2(a),(b), the free energy of the hairpin is -0.3 kcal/mol and -13.2 kcal/mol more favorable than that of RNA duplex at 276 and 300 K, respectively. Both duplex and the hairpin are more stable than the single stranded form at 276 and 300 K. The results indicate that (1) hairpin and duplex are the predominantly

populated states at 276 K, and (2) the duplex structure becomes less stable and the conformational equilibrium is shifted in favor of hairpin and single stranded RNA with increasing temperature.

We compare these results with NMR spectroscopy and thermodynamic measurements. Fürtig et al.⁸ monitored the temperature-dependent NMR spectra change in this RNA. They observed signals from both hairpin and duplex at 3 °C, whereas only the hairpin signal was detected at 27 °C. They also observed that between 37 °C and 67 °C the hairpin melts and forms single stranded RNA. This trend in the temperature-induced secondary structural changes is consistent with our free energy calculations using the $G(\text{MM-PB/SA})$ approach.

Analyzing the energy components in Table S1 of the Supporting Information, we found that the total electrostatic energy $E(\text{total_elec})$ is the principal factor responsible for a favorable change in the $G(\text{MM-PB/SA})$ free energy for the hairpin over the duplex and the single stranded form. As the temperature is increased from 276 to 300 K, $E(\text{total_elec})$ becomes less negative for both the duplex and the single stranded form by ~ 10 kcal/mol and ~ 9 kcal/mol, respectively, while the $E(\text{total_elec})$ of the hairpin becomes more negative by -2.7 kcal/mol. This effect is mainly due to the fact that the favorable change in the electrostatic solvation free energy $E(\text{PB})$ of the duplex form is cancelled out by the unfavorable changes in the solute electrostatic interaction energies $E(\text{elec})$ as the temperature is raised.

These results suggest that the conformational equilibrium in the RNA hairpin studied here will be sensitive to changes in the solvent dielectric properties. The free energy results calculated at different salt concentrations are presented in Table 3, which shows that the duplex formation is favored by added salt. This prediction is supported by the experimental observation in the NMR study of the conformational equilibrium of the RNA hairpin by Fürtig et al.⁸ They found out that the signal from duplex RNA appears when the NaCl concentration is increased from 10 to 120 mM. The stabilization of the duplex is attributed to the increased charge screening by added salt in the net repulsive electrostatic interaction between the two chains in a duplex RNA. As observed in earlier studies,²³ conformers with stronger gas-phase electrostatic energy are affected by the charge screening more than those conformers with weaker electrostatic interactions, as is the case with the hairpin and single stranded forms.

The thermodynamic parameters for the hairpin formation in a 12-nucleotide GGAC(UUUU)GUCC tetraloop sequence were reported by Antao and Tinoco⁶ and more recently by Proctor and co-workers.⁷ The value of ΔH° was found to be

between -38.7 and -42.7 kcal/mol. In order to estimate the ΔH° between the hairpin and the single stranded RNA from our MM-PB/SA calculations, we excluded the solute entropy and the small nonpolar contribution $\Delta G(\text{np})$ from the MM-PB/SA free energy difference, since the latter is mainly related to changes in solvent entropy (the hydrophobic effect). This yields calculated ΔH° values between -23.9 kcal/mol at 276 K and -22.9 kcal/mol at 300 K for the equilibrium between the hairpin and the single stranded form. This result, which was obtained for the 10-nucleotide CGC-(UUUU)GCG tetraloop, is in qualitative agreement with the experimental value for the 12-nucleotide RNA hairpin.

Free Energy Results: Including Solute Entropy Contributions. The results for the total free energy difference $\Delta G(\text{tot})$, which includes the solute translational, rotational entropies, and the normal modes entropy $T\langle S(\text{vib})\rangle$, are presented in Table 2 (see also Table S1 in the Supporting Information). The values for the $\Delta G(\text{tot})$ between the hairpin and the duplex are -8.6 and -22.3 kcal/mol at 276 and 300 K, respectively, while the $\Delta G(\text{tot})$ between duplex and the single stranded RNA is -15.7 kcal/mol at 276 K and -1.2 kcal/mol at 300 K. These results are qualitatively consistent with a shift in the conformational preferences favoring the hairpin and single stranded form over the duplex as the temperature is increased.

As expected, the translational/rotational entropy terms are found to oppose the duplex formation. The normal-mode analysis gives similar vibrational entropy contributions for the three secondary structural forms. At 276 K, the duplex has a slightly more favorable $T\langle S(\text{vib})\rangle$ than the hairpin and the single stranded RNA by -3.5 and -1.8 kcal/mol, respectively. This result reflects that on average the duplex RNA exhibits larger vibrational motions, possibly due to the four mismatched U-U base pairing in the middle of the double helices. The small differences in the $T\langle S(\text{vib})\rangle$ also suggest that the energy landscape for local minima is largely dependent on the topology, i.e., atom connectivity determined by bond, angles, and torsions which are similar for the three structural forms, with some modulation by the presence and the strength of the nonbonded interactions such as hydrogen bonds.

Our calculations predict a $\Delta G(\text{tot})$ of -8.6 kcal/mol for the hairpin-duplex equilibrium at 276 K in favor of the hairpin. However, according to the NMR study the two forms coexist at 3°C , which means that the free energy difference $\Delta G(\text{tot})$ is expected to be small and needs to be of the order of 0 kcal/mol at 276 K. The -8.6 kcal/mol difference between the calculated $\Delta G(\text{tot})$ and the experimental $\Delta G(\text{tot})$ suggests that certain free energy contributions are missing or calculated with insufficient accuracy. We think that this can be partially attributed to the inadequacy of the harmonic analysis in estimating the conformational entropy.

In this context we examined approximations that are used in calculation of the solute entropy. While the vibrational entropy evaluated using the harmonic analysis favors the duplex form over the hairpin by about 3.5 kcal/mol, the calculation using normal-mode analysis does not capture the changes in conformational entropy associated with transitions between energy minima. This view is qualitatively supported

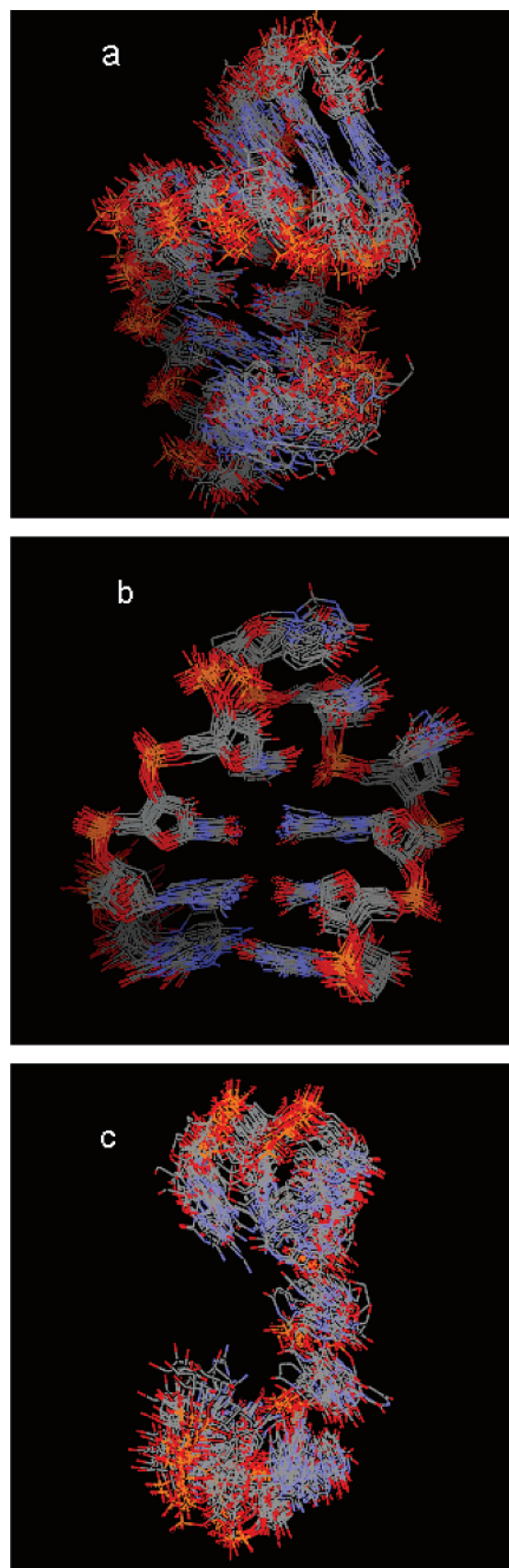


Figure 2. Superimposed snapshots from the trajectories obtained at 276 K: (a) duplex, (b) hairpin, and (c) single stranded form.

by the overlay of the superimposed RNA structures presented in Figure 2, which shows that the duplex form is considerably

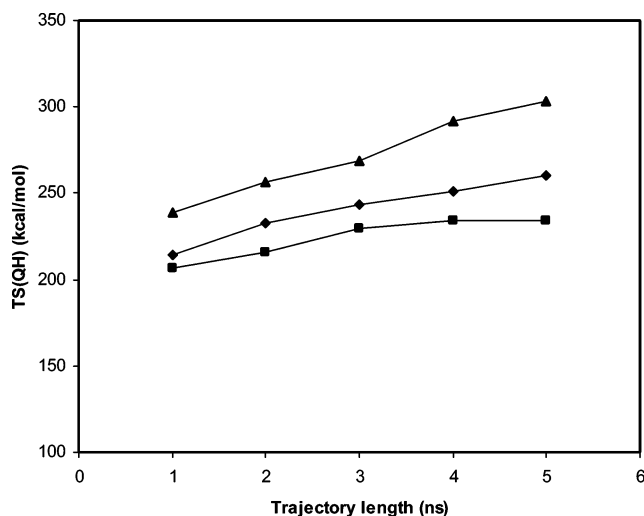


Figure 3. Quasiharmonic entropy contributions $TS(QH)$ calculated at 276 K for different lengths of the MD trajectories: single strand (triangle), duplex (diamond), and hairpin (square).

more flexible than the hairpin. This suggests that conformational entropy $\Delta S(\text{conf})$ favors the duplex formation. Figure 2 also shows that the single stranded RNA is more disordered than both the duplex RNA and the hairpin form, which suggests that the rank order of the solute conformational entropy is $S(\text{single strand}) > S(\text{duplex}) > S(\text{hairpin})$.

Interestingly, the trend in relative conformational entropy for the three secondary structural forms is reproduced by the quasiharmonic calculation of entropy: see Table 4 and Figure 3 for the results of $TS(QH)$ calculated from different trajectory lengths. Although the calculated quasiharmonic entropy did not converge within 5 ns of sampling, the rank order of the relative vibrational entropy clearly indicates that $S(\text{single strand}) > S(\text{duplex}) > S(\text{hairpin})$, which is consistent with the general picture of conformational fluctuation shown in Figure 2. The trend of relative vibrational entropy remains unchanged with the length of trajectory segment used in calculating $TS(QH)$.

However, it can be shown that the calculated $T\Delta S(QH)$ is significantly overestimated relative to expected values. We use the data set obtained with the 2 ns trajectory segment to illustrate this point. The calculated $T\Delta S(QH)$ between the duplex RNA and hairpin is -8.4 kcal/mol more negative than the -8.6 kcal/mol required to bring the calculated $\Delta G(\text{tot})$ for the duplex-hairpin equilibrium to the experimental value of 0 kcal/mol at 276 K. The quasiharmonic entropy are overestimated even more for the single stranded RNA: including the $T\Delta S(QH)$ calculated for the 2 ns trajectory segment in the free energy results predicts that the single stranded RNA is more stable than the hairpin by about -17.2 kcal/mol at 276 K, in contradiction to the experimental finding that the hairpin and the duplex RNA are the predominant forms at this temperature.

Two conclusions emerge from this result. First, the discrepancy between experimental and calculated $\Delta G(\text{tot})$ for the duplex/hairpin equilibrium may be qualitatively explained by the conformational entropy which is not accounted for by the simple harmonic approximation used

by the normal-mode analysis to calculate $T\Delta S(\text{vib})$ as shown in Table 2. Second, while the quasiharmonic analysis correctly predicted the rank order of relative conformational entropy contribution, the method significantly overestimates the entropy difference for the 10-nucleotide RNA molecule. This error could be related to the assumption of a multivariate Gaussian distribution in the quasiharmonic approximation, which does not hold in the presence of configurational transitions involving multiple energy wells.^{53,59} Clearly, more accurate treatment of the conformational entropy beyond the level of harmonic/quasiharmonic model is required for a quantitative understanding of the duplex-hairpin equilibrium.

The larger conformational fluctuations exhibited by the duplex compared to the other forms may be attributable to the dynamics of internal loop formed by the four mismatched U–U base pairs in the center of the helix. The relative flexibility in different parts of the duplex structure can be seen from the overlay of the superimposed segment structures (Figure S4, Supporting Information). The internal loop in the duplex shows more fluctuation than the helix regions in the duplex; it is also more disordered than the loop and stem regions in the hairpin form. Fluctuation in the internal loop of the duplex is also likely to be transmitted to the two ends of the double helix and enhance the bending and twisting of the helix, although MD simulations at longer time scale is needed to observe such large scale movement. Figure S4 also suggests that the stem and the loop region of the hairpin structure have similar flexibility.

Stable Hairpin Loop Conformers Observed in Explicit Solvent Simulations. The MD simulations L276A and L300, which are initiated from hairpin conformations at 276 and 300 K, respectively, converged to the same lowest energy conformer (Figures S5 and S6, Supporting Information). In both simulations, the rms deviation relative to the final structure of the L300 simulation decreases to less than 1.5 Å. In L276A simulation, the transition to the converged structure occurs at 7.7 ns, while the similar transition occurs at a much later time, ~ 50 ns, for the L300 simulation. The transitions to the converged structure resulted in reductions in the MM-PB/SA free energy between -15 kcal/mol and -20 kcal/mol relative to the initial structure, indicating that the final structure attained by the two simulations is the most stable loop conformation described by the physical model used in the present study. Since the 3D coordinates are not available for the hairpin structure from NMR studies, we cannot rule out that our simulation starting from modeled hairpin conformations may not have converged to the global free energy minima. However, the following observations suggest that the final parts of the trajectory in our explicit solvent simulations have converged to physically viable conformational states: (1) free energies and structures stabilized in the 70 ns simulation at 300 K; (2) independent simulations at 276 and at 300 K led to the same low-energy hairpin conformations; (3) some experimental results, such as ΔH° of hairpin melting and the temperature dependence of conformational equilibrium, are reproduced reasonably well by free energy calculations using the last part of the trajectories.

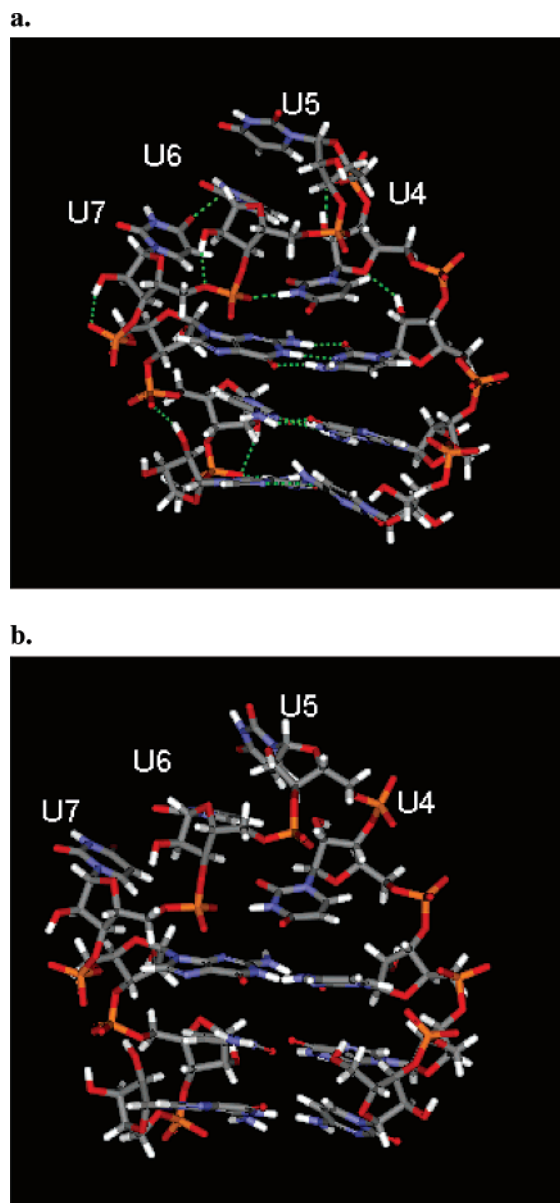


Figure 4. Final structures in the (a) L276A and (b) L300 simulations.

The final structures from the L276A and L300 trajectories are shown in Figure 4(a),(b). Several specific interactions can be identified in these structures: a weak stacking interaction between the U5 and U6, a hydrogen bond between the U4 base and the phosphate linking the U6 and U7, and hydrogen bonds involving the 2'OH groups at U4 and U6. The stacking interaction between the U5 and U6 bases is persistent for the most part of the 70 ns simulation L300 (Figure S1, Supporting Information), and the average separation distance between the U5 and U6 bases is around 4 Å. The van der Waals interaction between the U5 and U6 bases contributed about -4.0 kcal/mol to the stability of the loop. In Figure S7 of the Supporting Information, the existence of a CH...O hydrogen bond between C6-H (U7) and O5'-(U7) is highlighted (see also Figure S2 in the Supporting Information for the time dependence of this hydrogen bond). This CH...O hydrogen bond has a 73% time occupancy during the simulation (using a 3.7 Å cutoff) and is also present in the crystal structure of the 1mme hairpin loop,

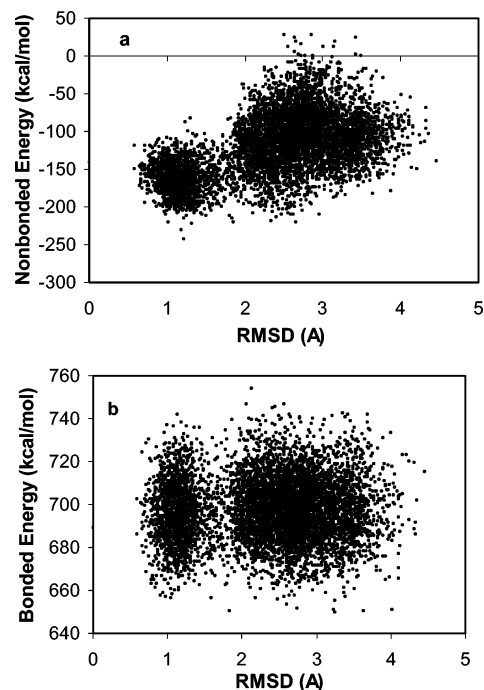


Figure 5. Free energy components in the L300 simulation: (a) total gas-phase nonbonded energy $E(\text{elec}) + E(\text{vdw})$ as a function of the rms deviation relative to the final structure and (b) total bonded energy $E(\text{intra})$ as a function of the rms deviation relative to the final structure.

from which the current UUUU hairpin is modeled. The importance of these nonbonded interactions in stabilizing the loop conformation can be seen in Figure 5(a), which shows that the conformers in the same cluster as the final structure have favorable total gas-phase electrostatic and van der Waals, $E(\text{vdw}) + E(\text{elec})$, interaction energy. In contrast to the nonbonded energy, the bonded energy term showed no correlation with the distance in rmsd space from the stable loop structure (Figure 5(b)).

Sugar Pucker Preferences. A recent NMR study⁷ on the cUUUg tetraloop found that U4 adopts the C3'-endo and U6 is 37% C2'-endo, while both U5 and U7 have predominantly the C2'-endo pucker. In the 70 ns hairpin simulation L300, which started with the four uridines in the C3'-endo conformation, conversion to the NMR sugar pucker mode did not occur, and all four uridines maintain the C3'-endo conformation throughout the simulation.

To investigate the relative stability of different sugar pucker for the loop residues, we conducted a 10 ns hairpin simulation L276B at 276 K, starting with the NMR pucker configuration in which the U4 and U6 are C3'-endo, while U5 and U7 are C2'-endo. Interestingly, in this simulation both the U5 (from 6.75 ns) and U7 (from 4 ns) are converted into the C3'-endo conformation: see Figure 6. This transition to the C3'-endo pucker is also reflected by the rmsd results for the loop residues: as seen from Figure 7, at 7.2 ns the loop conformation is converted into the stable hairpin loop structure observed in the L276A and L300 simulations. The automatic transition to the C3'-endo pucker suggests that according to the physical model used in the present study, the C3'-endo conformation is more stable than the C2'-endo conformation deduced from NMR analysis.

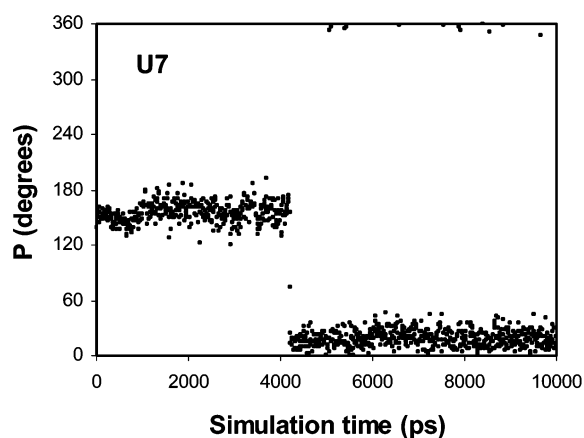
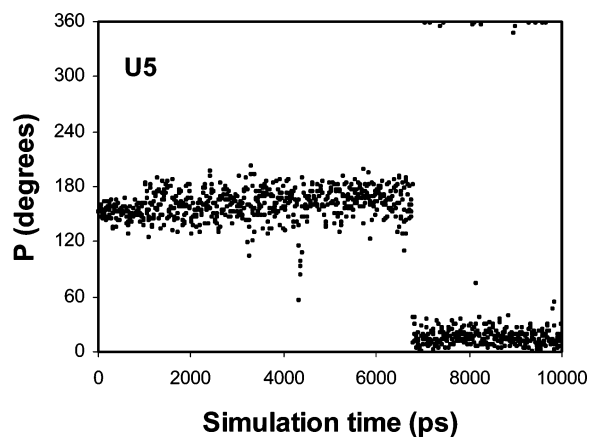


Figure 6. Results for the L276B simulation: the sugar pucker pseudorotation phase angle P for U5 and U7 as a function of simulation time.

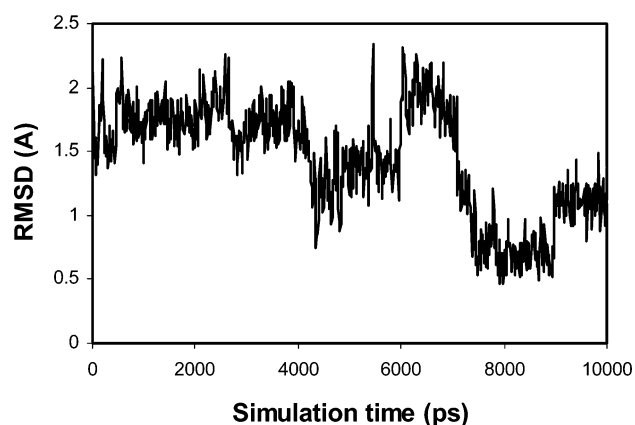


Figure 7. Results for the L276B simulation: rms deviation for the loop residue (c3-U4-U5-U6-U7-g8) relative to the last frame in the L300 trajectory as a function of simulation time. rmsd are calculated for the heavy atoms in the loop.

Two conformational substates of interest can be identified in Figure 6: the 2–4 ns time window corresponds to the C2'-endo conformer observed in the NMR study, and the 7.5–10 ns window represents the C3'-endo state. Representative structures for these two states are shown in Figure 9. These structure snapshots taken before and after the S-type \rightarrow N-type sugar pucker transition reveals that the C3'-endo conformer is stabilized by the U5–U6 and U6–U7 base stacking interactions. This base stacking is not possible to

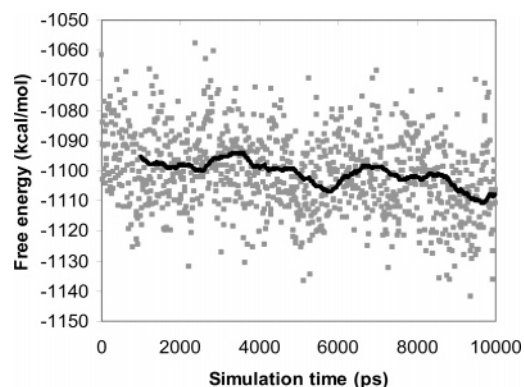


Figure 8. Results for the L276B simulation: G(MM-PB/SA) free energy as a function of simulation time.

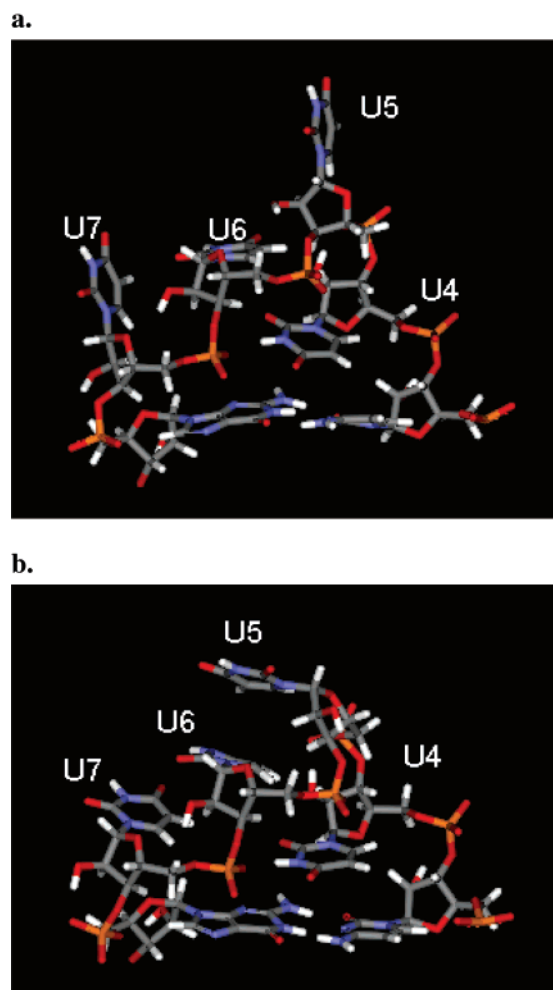


Figure 9. Results for the L276B simulation: representative structures in which both U5 and U7 are in the (a) C2'-endo and (b) C3'-endo conformations. Structures shown in (a) and (b) are from the 4 and 10 ns snapshots, respectively.

create in the C2'-endo puckering mode where the U5 and U7 bases extend out into the solvent phase.

Based on the above assignment of conformational substates, we calculated the effects of the loop sugar pucker on free energy (Table 5). The C2'-endo conformer has more favorable electrostatic solvation free energy $E(\text{PB})$, which is consistent with the fact that the U5 and U7 in the C2'-endo state extend away from the loop and thus are better

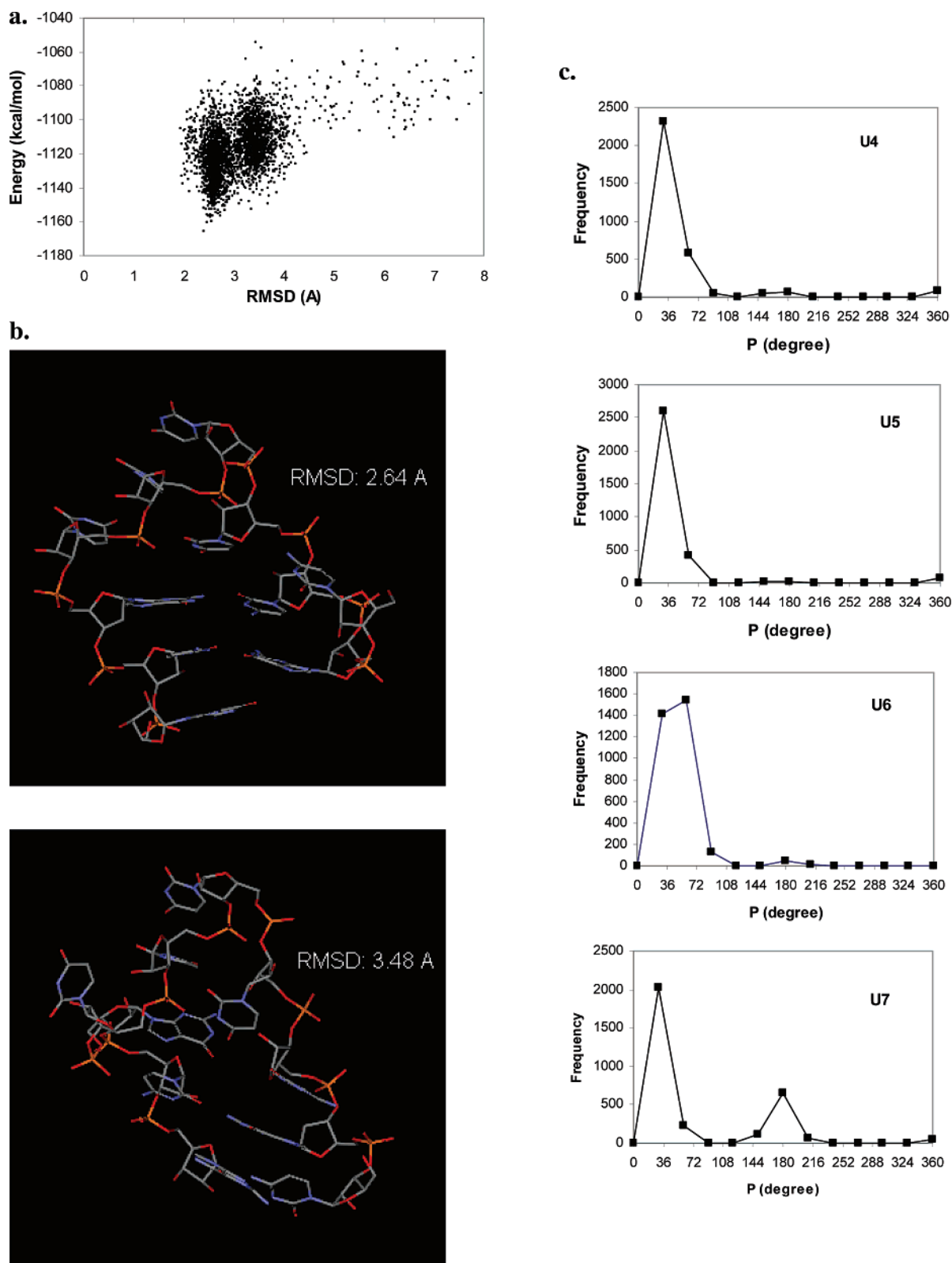


Figure 10. Results for the REMD simulation: (a) the total energy as a function of rms deviation from the starting structure, which is the stable hairpin loop conformation observed in the explicit solvent simulations, (b) representative structures for the two most populated clusters, and (c) the time occupancy of the sugar pucker modes for the four uridines. All data were collected from the last 6.3 ns of the simulation.

solvated (Figure 9). However, while the total electrostatic energy $E(\text{total_elec})$ favors the more extended conformation in Figure 9(a), this is offset by the larger unfavorable changes in the van der Waals energy and the bonded energy. The unfavorable van der Waals energy is associated with the loss

of U5–U6 and U6–U7 base stacking in the C2'-endo conformer discussed earlier. Overall, the MM-PB/SA calculation predicts that the C3'-endo is more stable than the C2'-endo structure by -8 kcal/mol (Table 5 and Figure 8), which is in discrepancy with the NMR experiments.

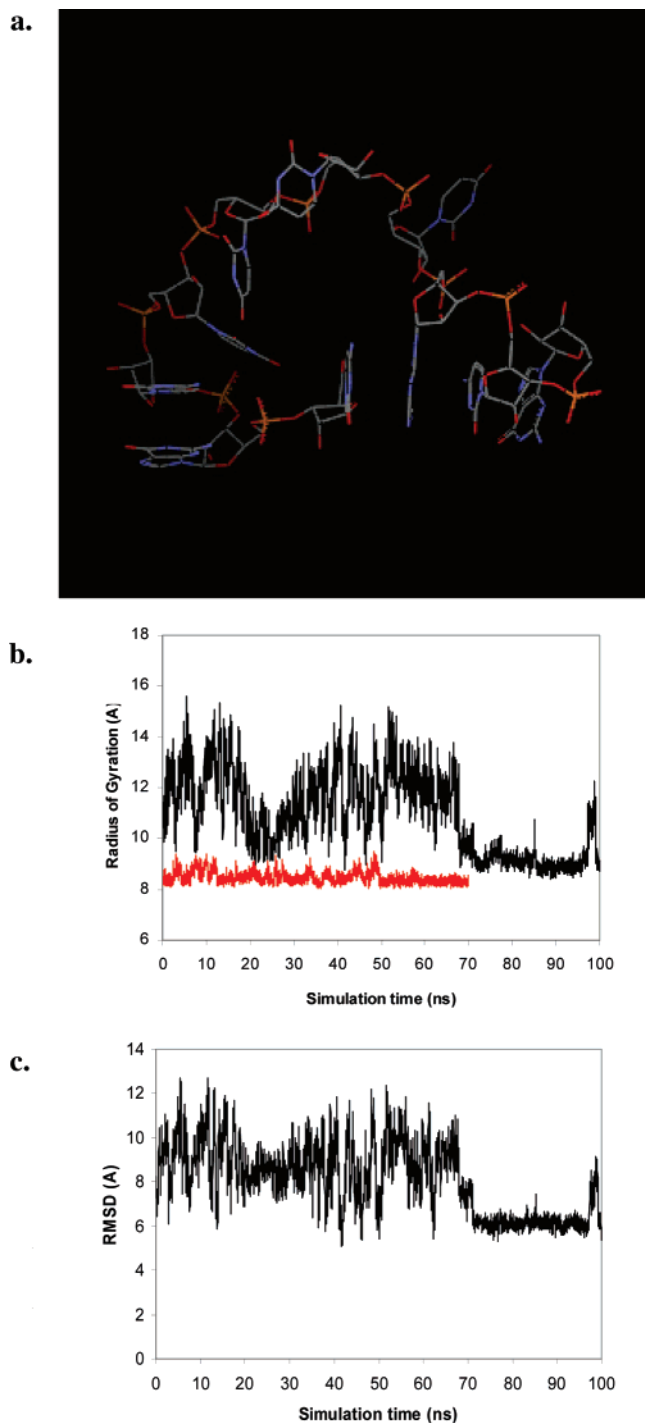


Figure 11. Results for the S300 simulation: (a) the final structure, (b) the temporal history of the radius of gyration for the S300 trajectory (black) and the L300 trajectory (red), and (c) the root-mean-squared distance (rmsd) from the hairpin loop structure as a function of the simulation time.

Although errors in the NMR data analysis cannot be completely ruled out, we think that the disagreement between our simulation and the NMR results more likely reflects some problems in the simulation method/energy model used in the explicit solvent and implicit solvent simulations in this study. Our simulation results shows that the repuckering to the experimental values was not achieved in 70 ns MD simulation in TIP3P water, although some limited repuckering is

seen from Replica-Exchange MD in the GBSW implicit solvent model (see below).

It is not clear whether the problem lies in force-field parametrization or is simply a result of limited sampling of the phase space. MacKerell⁴⁰ and co-workers reported the conformational energetics for a single deoxyribose sugar using the CHARMM27 parameter set. The energy difference between the North- and South-type sugar is <1 kcal/mol, and the barrier between the two conformations through the O4'-endo conformation is ≤ 3 kcal/mol. These values for sugar pucker in single nucleotide are consistent with estimations from ab initio calculations.²⁵

In the loop region of the RNA hairpin, where 2'-OH hydroxyl groups can interact with neighboring nucleotide, ions and water, the barrier between the N- and S-type sugar pucker should be higher those in the single deoxyribose. This may explain sampling problem in our calculations, where the repuckering to the experimental values was not achieved in 70 ns MD in TIP3P water. The fact that some limited repuckering is seen from our REMD simulations in the GBSW implicit solvent model reflects the role of solvent molecules in slowing down the conformational dynamics.

In addition to possible problems with the potential energy function used in the MD simulations, the MM-PB/SA scoring method may not be accurate or sensitive enough to correctly rank the different conformations of a few ribose moieties in the loop region: the MM-PB/SA calculation actually favors the hairpin conformer observed in MD simulation, which exhibits the non-NMR sugar pucker. This incorrect result underscore the limitation of MM-PB/SA discussed in the Introduction.

We feel that a more thorough study is warranted to clarify the situation on the stability and the energy barrier of conversion in ribose conformation in RNA hairpin, and such a study should be useful for improving force field and implicit solvent model.

Replica Exchange MD Simulations. We performed replica exchange MD simulation to explore broader conformational space for the hairpin form. The REMD simulation was initiated from the stable loop conformation obtained in the explicit solvent simulations. The results presented in Figure 10(a) reveal a funnel-like energy surface and a clear correlation between the total energy and the rmsd from initial structure. The low rmsd conformers are associated with low MM-GBSA energy, indicating that the stable hairpin conformation observed in the explicit solvent simulation corresponds to the free energy minimum. Two most populated clusters are identified in Figure 10(a), and their representative structures are shown in Figure 10(b). In the low rmsd structure (Figure 10(b)), the loop conformation is close to the initial structure, and the rmsd of 2.64 Å arises mainly from the base pair fraying at the end of the stem. Excluding the two end residues, the rmsd is 1.54 Å.

Interestingly, in the high rmsd structure in Figure 10(b), the U7 sugar exhibits the C2'-endo pucker, in agreement with the NMR sugar puckering. This high rmsd structure shows more distortions in both the loop and the stem, and as seen from Figure 10(a), it is thermodynamically less stable than the low rmsd structure by ~ 10 kcal/mol. Thus, in agreement

with our explicit solvent simulations, the REMD simulations also indicate that the C3'-endo conformation is preferred over the C2'-endo form for U7 sugar.

Figure 10(c) shows the time occupancies of the sugar pucker pseudorotation phase angle for the four uridines. While the U4, U5, and U6 sugar rings are predominantly in the C3'-endo conformation, the REMD simulation revealed more variations in the sugar pucker of the U7 ribose ring, in which the ratio between the time occupancy for the C2'-endo and that for the C3'-endo conformations is 1:2.7. It should be noted that while the REMD simulation reveals the trend in the conformational flexibility of the loop residues, the sampling is limited by the duration of the simulation length, and the results may not reflect the true equilibrium condition.

Submicrosecond Simulation of Single Stranded RNA.

A 100 ns simulation initiated from the single stranded conformation was performed at 300 K in an explicit water solvent to study a possible folding pathway for the formation of hairpin loop. As seen from Figure 11, although the RNA did not fold into the nativelike hairpin structure during 100 ns of simulation, the transitions toward more compact structures were observed. These structures have a similar radius of gyration to that of the hairpin loop. The rmsd of the final structure is 5.8 Å from the native hairpin conformation. The structure appears to be trapped into a local minima stabilized by the incorrect base stacking between the two end nucleotides. The MM-PB/SA calculation suggests that the inability to fold the RNA into hairpinlike structures is attributable to the inadequate sampling at room temperature in explicit solvent: as can be seen from Table 2b, the average $G(\text{MM-PB/SA})$ of the single stranded RNA at the end of the 100 ns simulation is 24.0 kcal/mol higher than that of the hairpin conformations. Simulations of a much longer time scale, possibly using REMD techniques, would be needed to observe folding of the hairpin at room temperature in explicit solvent.

4. Conclusions

To investigate the physical basis for the conformational equilibrium in a 10-nucleotide UUUU RNA tetraloop, we have applied the MM-PB/SA method to estimate the free energies of different RNA secondary structures in solution. The results reproduce the experimental trend in relative stabilities of duplex, hairpin loop, and single stranded RNA at two temperatures: The difference in calculated values of the $G(\text{MM-PB/SA})$ free energy between the duplex and hairpin is close to zero at low temperature. With the increase in temperature, the duplex structure was found to be destabilized, consistent with the experiments. The calculated ΔH° for the hairpin-single strand RNA conversion falls into the range of the experimental values for similar UUUU RNA tetraloops.

While the results using the gas-phase and solvation terms in the free energy calculation reproduce the experimental enthalpy and the trend of conformational equilibrium as a function of temperature, neglecting the entropy contributions yields calculated ΔG for the hairpin-duplex RNA equilibrium

larger than the experimental values at 276 K. Qualitatively, we have shown that the three secondary structural forms exhibit significantly different conformational fluctuations, and the rank order of the solute conformational entropy is $S(\text{single strand}) > S(\text{duplex}) > S(\text{hairpin})$. We used the normal-mode analysis and quasiharmonic analysis to estimate this entropic contribution to free energy. Results from both methods indicate that the helical duplex exhibits more chain flexibility than the hairpin, in agreement with inspection of the geometry fluctuations in the snapshots extracted from the trajectory. The quasiharmonic method reproduces the ranking order for the relative conformational flexibility, but the calculated entropies are overestimated. Overall, the quantitative estimation of entropy remains a very challenging task and is the major source of the uncertainties in the free energy determination using end-point methods such as the MM-PB/SA approach.

Our simulations conducted at two temperatures revealed the low-energy hairpin loop structure and a number of specific interactions that are responsible for its stability. However, the predicted sugar pucker modes in the U5 and U7 loop uridines in the low-energy hairpin conformation are different from those suggested from NMR experiments. This inability to reproduce the experimental sugar pucker mode for the two loop uridines suggests that either the experimental data are not precise enough or there are flaws in the simulation protocol adopted here. Accuracies in MD simulations are known to be limited by errors in the force-field parametrization and inadequate sampling of the configuration space. The importance of more extensive sampling was underscored by our 8.3 ns REMD simulation in the GBSW continuum solvent initiated from the hairpin loop structure.

Although the REMD simulation confirmed the stability of the low-energy loop structure observed in the explicit solvent simulations, it reveals significantly more fluctuation in the sugar pucker of the U7 ribose ring, and the results of the time occupancies for the C3'-endo and C2'-endo conformations are in better agreement with the experimental values compared to those predicted from the explicit solvent simulations. Using advanced sampling methods such as REMD at a longer time scale combined with more accurate implicit solvent models could eventually improve the theoretical results.

Acknowledgment. We thank Dr. Jianhan Chen for helpful discussions. We also thank Dr. Jayashree Srinivasan for providing the reprint of ref 13.

Supporting Information Available: Three tables and figures. The free energy results obtained at $T = 276$ K and $T = 300$ K; the free energy differences $\Delta \langle G(\text{tot}) \rangle$ calculated using different trajectory segments; the free energies calculated using different coordinate sets along MD trajectories; the distance between the U5 and U6 bases in the L300 simulation; the distance between C6 (U7) and O5' (U7) as a function of simulation time in the L300 simulation; MM-PB/SA free energy as a function of simulation time at 276 K; the superimposed structures for different parts of the

RNA; the results for the L300 and L276 trajectories: RMSD from the last frames of the trajectories and MM-PBSA calculations; the CH \cdots O hydrogen bond between C6H (U7) and O5' (U7); and the starting and final conformations of the different secondary structural forms from simulations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Uhlenbeck, O. Nucleic-acid structure-tetraloops and RNA folding. *Nature* **1990**, *346*, 613–614.
- (2) Jagath, J. R.; Matassova, N. B.; de Leeuw, E.; Wernecke, J. M.; Lentzen, G.; Rodnina, M. V.; Luirink, J.; Wintermeyer, W. Important role of the tetraloop region of 4.5S RNA in SRP binding to its receptor FtsY. *RNA* **2001**, *7*, 293–301.
- (3) Althoff, S.; Selinger, D.; Wise, J. A. Molecular evolution of SRP cycle components: Functional implications. *Nucleic Acids Res.* **1994**, *22*, 1933–1947.
- (4) Antao, V. P.; Lai, S. Y.; Tinico, I.; Jr.; A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res.* **1991**, *19*, 5901–5905.
- (5) Persson, T.; Hartmann, R. K.; Eckstein, F. Selection of hammerhead ribozyme variants with low Mg²⁺ requirement: importance of stem-loop II. *ChemBioChem* **2002**, *3*, 1066–71.
- (6) Antao, V. P.; Tinoco, I., Jr. Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic Acids Res.* **1992**, *20*, 819–824.
- (7) Proctor, D. J.; Ma, H.; Kierzek, E.; Kierzek, R.; Gruebele, M.; Bevilacqua, P. C. Folding thermodynamics and kinetics of YNMG RNA hairpins: specific incorporation of 8-bromoguanosine leads to stabilization by enhancement of the folding rate. *Biochemistry* **2004**, *43*, 14004–14014.
- (8) Fürtig, B.; Richter, C.; Wöhnert, J.; Schwalbe, H. NMR spectroscopy of RNA. *ChemBioChem* **2003**, *4*, 936–962.
- (9) Cheatham, T. E. III Simulation and modeling of nucleic acid structure, dynamics and interactions. *Curr. Opin. Struct. Biol.* **2004**, *14*, 360–367.
- (10) MacKerell, A., Jr.; Nilsson, L. *Nucleic acid simulations, Computational Biochemistry and Biophysics*; Becker, O., MacKerell, A., Jr., Roux, B., Watanabe, M., Eds.; Marcel Dekker Inc.: New York, 1999; pp 441–463.
- (11) Orozco, M.; Perez, A.; Noy, A.; Luque, F. Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.* **2003**, *32*, 350–364.
- (12) Karplus, M.; Kuriyan, J. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci.* **2005**, *102*, 6679–6685.
- (13) Srinivasan, J.; Miller, J.; Kollman, P.; Case, D. Continuum solvent studies of the stability of RNA hairpin loops and helices. *J. Biomol. Struct. Dyn.* **1998**, *16*, 671–682.
- (14) Williams, D.; Hall, K. Experimental and computational studies of the G[UUCG]C RNA tetraloop. *J. Mol. Biol.* **2000**, *297*, 1045–1061.
- (15) Sorin, E.; Engelhardt, M.; Herschlag, D.; Pande, V. RNA simulations: probing hairpin unfolding and the dynamics of a GNRA tetraloop. *J. Mol. Biol.* **2002**, *317*, 493–506.
- (16) Sarzynska, J.; Nilsson, L.; Kulinski, T. Effects of base substitutions in an RNA hairpin from molecular dynamics and free energy simulations. *Biophys. J.* **2003**, *85*, 3445–3459.
- (17) Sorin, E.; Rhee, Y.; Pande, V. Does water play a structural role in the folding of small nucleic acids? *Biophys. J.* **2005**, *88*, 2516–2524.
- (18) Sorin, E.; Rhee, Y.; Nakatani, B.; Pande, V. Insights into Nucleic Acid Conformational Dynamics from Massively Parallel Stochastic Simulations. *Biophys. J.* **2003**, *85*, 790–803.
- (19) Spackova, N.; Sponer, J. Molecular dynamics simulations of sarcin-ricin rRNA motif. *Nucleic Acids Res.* **2006**, *34*, 697–708.
- (20) Li, W.; Ma, B.; Shapiro, B. Molecular dynamics simulations of the denaturation and refolding of an RNA tetraloop. **2001**, *19*, 381–396.
- (21) Williams, D. J.; Hall, K. B. Unrestrained Stochastic Dynamics Simulations of the UUCG Tetraloop Using an Implicit Solvation Model. *Biophys. J.* **1999**, *76*, 3192–3205.
- (22) Fadrna, E.; Spackova, N.; Stefl, R.; Koca, J.; Cheatham, T. E., III; Sponer, J. Molecular dynamics simulations of Guanine quadruplex loops: advances and force field limitations. *Biophys. J.* **2004**, *87*, 227–242.
- (23) Srinivasan, J.; Cheatham, T. E., III; Cieplak, P.; Kollman, P.; Case, D. Continuum solvent studies of the stability of DNA, RNA and phosphoramidate-DNA helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409.
- (24) Jayaram, B.; Sprous, D.; Young, M. A.; Beveridge, D. L. Free Energy Analysis of the Conformational Stability of A and B Forms of DNA in Solution. *J. Am. Chem. Soc.* **1998**, *120*, 10629–10633.
- (25) Kollman, P. I.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D.; Cheatham, T., III Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889–897.
- (26) Swanson, J.; Henchman, R.; McCammon, J. Revisiting free energy calculations: A theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys. J.* **2004**, *86*, 67–74.
- (27) Zoete, V.; Meuwly, M.; Karplus, M. Study of the insulin dimerization: binding free energy calculations and per-residue free energy decomposition. *Proteins* **2005**, *61*, 79–93.
- (28) Gohlke, H.; Case, D. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem.* **2004**, *25*, 238–250.
- (29) Jayaram, B.; McConnell, K.; Dixit, S.; Beveridge, D. Free Energy Analysis of Protein–DNA Binding: The EcoRI Endonuclease–DNA Complex. *J. Comput. Phys.* **1999**, *151*, 333–357.
- (30) Spackova, N.; Cheatham, T. E., III; Ryjáček, F.; Lankas, F.; van Meervelt, L.; Hobza, P.; Sponer, J. Molecular Dynamics Simulations and Thermodynamics Analysis of DNA-Drug Complexes. Minor Groove Binding between 4',6'-Diamidino-2-phenylindole and DNA Duplexes in Solution. *J. Am. Chem. Soc.* **2003**, *125*, 1759–1769.

- (31) Gouda, H.; Kuntz, I.; Case, D.; Kollman, P. Free energy calculations for theophylline binding to an RNA aptamer: Comparison of MM-PBSA and thermodynamic integration methods. *Biopolymers* **2003**, *68*, 16–34.
- (32) Lee, M.; Duan, Y.; Kollman, P. Free energy analysis of the folding process of villin headpiece subdomain. *Proteins* **2000**, *39*, 309–316.
- (33) Cubero, E.; Luque, F.; Orozco, M. Theoretical studies of d (A:T)-based parallel-stranded DNA duplexes. *J. Am. Chem. Soc.* **2001**, *123*, 12018–12025.
- (34) Spackova, N.; Berger, I.; Spomer, J. Nanosecond molecular dynamics of zipper-like DNA duplex structures containing sheared G:A mismatch pairs. *J. Am. Chem. Soc.* **2000**, *122*, 7564–7572.
- (35) Im, W.; Lee, M. S.; Brooks, C. L., III Generalized born model with a simple smoothing function. *J. Comput. Chem.* **2003**, *24*, 1691–1702.
- (36) Lee, M. S.; Salsbury, F. R., Jr.; Brooks, C. L., III Novel generalized Born methods. *J. Chem. Phys.* **2002**, *116*, 10606–10614.
- (37) Scott, W. G.; Finch, J. T.; Klug, A. The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. *Cell* **1995**, *81*, 991–1002.
- (38) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (39) Brooks, B.; Bruccoleri, R.; Olafson, B.; States, D.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (40) Foloppe, N.; MacKerell, A. D., Jr. All atom empirical force field for nucleic acids: 1) Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comput. Chem.* **2000**, *21*, 86–104.
- (41) MacKerell, A., Jr.; Banavali, N. All-atom empirical force field for nucleic acids: 2) Application to molecular dynamics simulations of DNA and RNA in solution. *J. Comput. Chem.* **2000**, *21*, 105–120.
- (42) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (43) Banavali, N. K.; Roux, B. Atomic radii for continuum electrostatics calculations on nucleic acids. *J. Phys. Chem. B* **2002**, *106*, 11026–11035.
- (44) McQuarrie, D. A. *Statistical mechanics*; Harper & Row: New York, 1976; pp 136–137.
- (45) Harris, S. A.; Gavathiotis, E.; Searle, M. S.; Orozco, M.; Laughton, C. A. Cooperativity in drug-DNA recognition: a molecular dynamics study. *J. Am. Chem. Soc.* **2001**, *123*, 12658–12663.
- (46) Jusuf, S.; Loll, P.; Axelsen, P. Configurational entropy and cooperativity between ligand binding and dimerization in glycopeptide antibiotics. *J. Am. Chem. Soc.* **2003**, *125* (13), 3988–3994.
- (47) Karplus, M.; Kushick, J. N. Method for estimating the configurational entropy of macromolecules. *Macromolecules* **1981**, *14*, 325–332.
- (48) Levy, R. M.; Karplus, M.; Kushick, J.; Perahia, D. Evaluation of the Configurational entropy for proteins: application to molecular dynamics simulations of an alpha-helix. *Macromolecules* **1984**, *17*, 1370–1375.
- (49) Brooks, B. R.; Janezic, D.; Karplus, M. Harmonic analysis of large systems. I. Methodology. *J. Comput. Chem.* **1995**, *16*, 1522–1542.
- (50) Schlitter, J. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett.* **1993**, *215*, 617–621.
- (51) Andricioaei, I.; Karplus, M. On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.* **2001**, *115*, 6289–6292.
- (52) Di Nola, A.; Berendsen, H.; Edholm, O. Free energy determination of polypeptide conformations generated by molecular dynamics. *Macromolecules* **1984**, *17*, 2044–2050.
- (53) Chang, C.; Chen, W.; Gilson M. Evaluating the accuracy of the quasiharmonic approximation. *J. Chem. Theory Comput.* **2005**, *1*, 1017–1028.
- (54) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (55) Hansmann, U. H. E.; Okamoto, Y. New Monte Carlo algorithms for protein folding. *Curr. Opin. Struct. Biol.* **1999**, *9*, 177–183.
- (56) 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.
- (57) Nose, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **1984**, *81*, 511–519.
- (58) Lee, M. S.; Olson, M. A. Calculation of absolute protein-ligand binding affinity using path and endpoint approaches. *Biophys. J.* **2006**, *90*, 864–877.
- (59) Janezic, D.; Brooks, B. Harmonic analysis of large systems. II. Comparison of different protein models. *J. Comput. Chem.* **1995**, *16*, 1543–1553.

CT6003388

Continuous Fractional Component Monte Carlo: An Adaptive Biasing Method for Open System Atomistic Simulations

Wei Shi and Edward J. Maginn*

*Department of Chemical and Biomolecular Engineering, University of Notre Dame,
182 Fitzpatrick Hall, Notre Dame, Indiana 46556-5637*

Received January 4, 2007

Abstract: A new open system Monte Carlo procedure designed to overcome difficulties with insertion and deletion of molecules is introduced. The method utilizes gradual insertions and deletions of molecules through the use of a continuous coupling parameter and an adaptive bias potential. The method draws upon concepts from previous open system molecular dynamics and expanded ensemble Monte Carlo techniques and is applied to both the grand canonical and osmotic ensembles. It is shown to yield correct results for the volumetric properties of the Lennard-Jones fluid and water as well as the phase behavior of the CO₂-ethanol binary system.

1. Introduction and Background

The use of atomistic simulations to compute the phase behavior of fluids has reached a high level of sophistication. Accurate vapor–liquid,^{1–3} liquid–liquid,^{4,5} solid–liquid,^{6–8} and solid–vapor^{9,10} equilibria calculations have been performed for a wide range of systems. Although a large number of simulation methods have been used for this purpose, they can be split into two main categories. The first category requires that the free energy (chemical potential) be computed for the system at a given state point relative to some reference state. Given this free energy, one can locate coexistence points through the enforcement of the phase equilibria criteria, namely equality of chemical potential, temperature, and pressure. In this category are techniques such as thermodynamic integration, free energy perturbation, expanded ensembles, and histogram methods. Kofke^{11,12} has provided excellent reviews of the strengths and weaknesses of these different approaches. The second category of simulation method, and the one we are concerned with here, utilizes an open system in which a chemical potential is imposed and the composition changes in response to this chemical potential. Examples include grand canonical Monte Carlo (GCMC) as well as various open system molecular dynamics techniques. The Gibbs ensemble technique can also be grouped in this category, although the chemical potential

is not specified a priori with this technique. Common to both simulation categories is the need to *insert and delete* molecules from the system to either evaluate the chemical potential or change composition. The accuracy and precision of the simulations depend critically on the ability to carry out these insertion and deletion moves. Poor sampling can result when systems are dense and have slowly evolving dynamics, such that suitable “holes” in the system large enough to accommodate the inserted molecule appear infrequently. It also can occur when there are specific interactions between the inserted or deleted species and the rest of the molecules in the system. For example, in strongly hydrogen-bonding systems specific orientations are required for energies to be favorable during an insertion. For deletions of hydrogen-bonded systems, the status quo conformation may be so energetically favorable that it is exceedingly improbable that a deletion can be successful.

A large number of biasing techniques have been developed to overcome these difficulties. While the literature is too extensive to list all efforts, we note that within the Monte Carlo (MC) framework moves have been designed to search for favorable energetic locations in the system¹³ as well as cavities where insertions will more likely be accepted.¹⁴ Advanced configurational bias moves,^{15–17} multicanonical methods,^{18–21} and many other sophisticated sampling techniques have also been developed. Of particular relevance to the present work are the various Monte Carlo “switch”

* Corresponding author phone: (574)631-5687; fax: (574)631-8366; e-mail: ed@nd.edu.

methods in which small “impurity” molecules are inserted and deleted into the system with great efficiency and then transformed into larger “full” molecules.²² Such an approach has been refined and applied to alkanes,²³ alcohols,²⁴ and polymers.²⁵

One strategy that is particularly appealing is to accompany molecule insertions or deletions with some sort of local relaxation process, such that energetic penalties can be minimized and thus the success probability of the move increased. This is one of the strategies behind various open system molecular dynamics (MD) methods.^{5,26–31} The basic idea behind these approaches is similar to that involved in the MC switch methods, in which insertions and deletions are “staged”. In the open system MD methods, this is accomplished by making either stochastic or deterministic changes in the coupling strength between a “fractional” molecule and the rest of the system. The composition changes as molecules are continuously created (destroyed) as the fractional molecule becomes fully coupled (decoupled) with the system. In between these changes in coupling strength, dynamics are run to enable the system to accommodate the change in the fractional molecule. The advantage is that the natural forces present can allow for collective or cooperative relaxation of the system in response to the perturbation; there is no need to develop special biasing rules ahead of time. Very recently, a new method has been proposed^{32,33} that takes this idea a step further by using local energy minimization to increase the success probability of insertions and deletions.

Here we report the development and application of a method for the efficient insertion and deletion of molecules in the grand canonical (μ, V, T) and osmotic (f_1, N_2, P, T) ensembles. The method, which we call continuous fractional component Monte Carlo (CFC MC), can be generalized to other ensembles such as the Gibbs ensemble. It offers a number of attractive features for simulating the phase behavior of complex systems, including self-adapting biasing capability that requires no foreknowledge of any bias functions. CFC MC draws upon a number of previous methods, including the expanded ensemble GCMC method described by Escobedo and de Pablo,³⁴ the Monte Carlo “swatch” move developed by Siepmann and co-workers,^{23–25} the expanded ensemble molecular dynamics method of Lísal and co-workers,³¹ and Boinepalli and Attard’s grand canonical molecular dynamics approach.³⁵ In the next section, a brief summary of related methods is provided, followed by details of the CFC MC technique. Following that, results from CFC MC simulations of three test systems are presented and shown to give equivalent results to accepted methods. We also demonstrated that some of the existing open system methods fail to give the correct results for these test systems.

2. Previous Methods

2.1. Open System Molecular Dynamics. As mentioned above, the advantage of performing insertions and deletions within an MD framework is that the forces allow the system to relax in response to the perturbations brought about by the insertion or deletion move. Cooperative and collective moves, which are difficult to treat with a priori MC biasing rules, occur more or less naturally with MD. One obvious

difficulty of open system MD approaches is that standard MD algorithms require a constant number of molecules. To simulate open systems with MD, a modified Hamiltonian is required, as is a method of coupling insertion and deletion moves dynamically to the system.

Çağın and Pettitt^{26,27} appear to have been the first to develop an MD method for the grand canonical ensemble. Their grand canonical molecular dynamics (GCMD) method utilizes an extended system Hamiltonian and a dynamical variable λ that links the physical system to a particle bath, thereby enabling compositional changes. The potential for λ is a function of the chemical potential of the system. Pettitt and co-workers successfully applied the GCMD method to a Lennard-Jones fluid^{27,36} and water.^{28,36} One problem that was observed with the GCMD method is that the dynamics can become unstable due to large accelerations in the coupling parameter when strongly repulsive regions are sampled, as frequently occurs when insertions result in significant molecular overlap. To minimize this, biased insertion and deletion moves were used.

Lo and Palmer²⁹ proposed an alternative extended Hamiltonian to perform GCMD simulations. This method still makes changes in the strength of the coupling between the fractional molecule and the rest of the system deterministically. The difference is that stochastic moves are made in between composition changes, thereby switching to a new Hamiltonian at each composition. Two other major differences exist. First, Lo and Palmer did not scale the mass of the transforming molecule, while the mass was scaled by λ in the work of Pettitt and co-workers. Second, while Pettitt and co-workers used a simple linear scaling to couple the fractional molecule with the system, in simulating a Lennard-Jones system Lo and Palmer chose the following functional form

$$\phi_f(r_{jf}, \lambda) = \lambda 4\epsilon \left\{ \left[\frac{\sigma}{r_{jf} + (1 - \lambda)\zeta} \right]^{12} - \left[\frac{\sigma}{r_{jf} + (1 - \lambda)\zeta} \right]^6 \right\} \quad (1)$$

where $\phi_f(r_{jf}, \lambda)$ is the interaction potential between the transforming “fractional” molecule and the other “integer” molecules in the system, ϵ and σ are the well depth and collision diameter for the LJ potential, j denotes integer molecules, and f denotes the fractional molecule. The parameter ζ can be adjusted to optimize the performance of the algorithm. Note that the normal Lennard-Jones potential is recovered when $\lambda = 1$. This type of scaling significantly reduces the repulsive forces present when atoms are close to one another at small values of λ and thus improves stability. Biased insertions are typically not required with this type of scaling.

Shroll and Smith³⁰ extended the GCMD method in two ways. First, they introduced a bias potential to improve the probability of transitions among adjacent values of λ . Without this bias potential, the system can become “stuck” at values of λ between 0 and 1 whenever the free energy barrier for λ transitions is large. This appears to not be an issue for the simple Lennard-Jones system³⁵ but becomes a factor for more complex systems such as water. In the spirit of expanded ensemble methods,³⁷ this bias potential should be close to the free energy difference between adjacent λ states to

maintain a uniform distribution of λ values. Shroll and Smith determined their bias potential for water by first computing the free energy difference between adjacent states via thermodynamic integration. The obvious drawback of this approach is that a separate thermodynamic integration must be carried out for each state point, which mitigates the advantages of using a bias potential. However, the authors note that if the free energy is insensitive to the state points, the same bias potential can be used for multiple state points. The second modification introduced by Shroll and Smith was the use of a continuous λ that “rolls over” when it reaches the boundaries of $\lambda = 0$ or $\lambda = 1$. That is, the velocity of the coupling parameter $\dot{\lambda}$ is reassigned to the new fractional molecule whenever the bounds on λ are exceeded. In the original GCMD methods, dynamic λ transitions abruptly end whenever λ exceeds its bounds. We also note that these authors found that the use of a nonlinear scaling of the potential avoided instabilities when the strongly repulsive regions of the potential were sampled.

2.2. Hybrid Methods. All of the above techniques employ fully deterministic methods for making changes in the value of the coupling parameter λ (and thus the composition) by employing an extended Hamiltonian. As with any extended system, they require the use of “time constants” or virtual masses for the extended variables, which control the rate of change of the extended variable. It has been pointed out³⁸ that to properly compute free energies using techniques such as this there must be a dynamical adiabatic separation between the evolution of the fractional molecule coupling and the remaining degrees of freedom. In other words, the transformation must be sufficiently slow to ensure a large separation in time scales between the dynamics of the system and that of λ . This is potentially a problem with fully dynamic methods, since there is no a priori way of knowing what the relevant “time constants” should be to ensure adiabatic separation.

It is also possible to exploit the favorable properties of dynamic insertions via *hybrid approaches* in which stochastic “open system” MC moves are used to change composition while deterministic MD trajectories are also used to relax the system. As with the former methods, particle number is a continuous variable so that insertions and deletions are not done with full molecules. Because composition changes are done stochastically using Metropolis-like acceptance rules, however, these techniques can be constructed to satisfy detailed balance and thus are guaranteed to satisfy the limiting probability distribution of the ensemble. Compared to extended system approaches, these methods are relatively simple to implement.

Boinepalli and Attard³⁵ have proposed such a hybrid method for the grand canonical ensemble. For a given state of the fractional molecule, a regular microcanonical MD step is taken, followed by a stochastic attempt to increase or decrease the value of the coupling parameter for a fractional particle. This method was implemented for the Lennard-Jones system and found to give correct volumetric properties.

Lísal and co-workers³¹ have recently proposed a hybrid MD–MC method called expanded-ensemble osmotic molecular dynamics. As with the Boinepalli and Attard method,

composition changes are treated stochastically by making gradual changes in the coupling strength of a fractional molecule with the other molecules in a system. As did Shroll and Smith, these authors used a bias potential to overcome problems with the coupling parameter becoming “stuck” at values between 0 and 1. The bias must be removed in the acceptance rules to obtain proper averages. Following Shroll and Smith, thermodynamic integration was used to determine the values for the bias potential.

3. Continuous Fractional Component Monte Carlo Method

The CFC MC method described here draws on a number of features of the previous methods outlined above. It is a hybrid method in which continuous composition changes are handled stochastically, while MD and, if desired, MC moves are used to relax the system. Extended variables (or bias potentials) are used to overcome large free energy barriers for insertion or deletion moves. Unlike previous methods, these bias factors are determined in a self-adaptive manner via an interactive Wang–Landau approach.³⁹ A nonlinear potential is used to avoid problems with repulsive overlap. We implement the method in both the expanded grand canonical and expanded osmotic ensembles and then use the technique to compute volumetric properties of the Lennard-Jones fluid and the simple point charge (SPC) model of water as well as phase equilibria for CO₂/ethanol mixtures. We demonstrate the accuracy and efficiency of the method by comparing CFC MC results with previous simulation studies and our own calculations using existing methods.

3.1. Grand Canonical Ensemble. In this section, we show how CFC MC is implemented in the grand canonical ensemble. To prove the validity of any stochastic move, it is sufficient to show that it satisfies detailed balance, which we write as

$$\Pi_{mn}\alpha_{nm}p_m = \Pi_{nm}\alpha_{mn}p_n \quad (2)$$

where Π_{mn} is the one-step transition or acceptance probability of going from state m to state n , α_{mn} is the probability of attempting such a move, and p_m and p_n are the ensemble probability distribution functions for states m and n , respectively. Now consider a system at fixed temperature T , chemical potential μ , and volume V that contains two types of fully flexible molecules containing n_a atoms: N fully interacting “integer” molecules and a single “fractional” molecule whose potential couples to the integer molecules via a pseudocontinuous coupling parameter λ of the type discussed above that ranges from 0 to 1. The fractional molecule can exist in a large number of states M , and at each state j there is an expanded system variable η_j associated with it. The partition function for this system is that of an expanded grand canonical ensemble having the following form^{34,40}

$$\Psi(\mu, V, T, \eta) = \sum_{N=0}^{\infty} \sum_{j=1}^M q_t^{N+1} \frac{\exp[\beta\mu N] \exp[\eta_j]}{N!} Z(N, j, V, T) \quad (3)$$

where $\beta = 1/k_B T$. Z is the configurational integral given by

$$Z(N, j, V, T) = \int \exp[-\beta\phi(\mathbf{q})] \prod_{i=1}^{N+1} d\mathbf{q}_i \quad (4)$$

where \mathbf{q}_i is the vector of generalized coordinates for molecule i , which includes both integer molecules and the fractional molecule. The variable q_i contains kinetic energy contributions from all atoms in the system (integer and fractional) as well as the Jacobian of the transformation from atomic Cartesian coordinates to internal coordinates.^{40,41}

It is often more convenient to work with fugacities f rather than chemical potentials. It is straightforward to show that the probability distribution function for some state m having N_m integer molecules and one fractional molecule is⁴⁰

$$p_m = \frac{1}{\Psi} \frac{(f\beta V)^{N_m} q_i \Omega}{N_m! (Z^{\text{IG}}/\Omega)^{N_m}} \exp[\eta_m] \exp[-\beta\phi_m] \quad (5)$$

where η_m is the bias factor associated with the particular value of λ at state m , ϕ_m is the total potential energy of state m , Z^{IG} is the ideal gas configurational integral of a single molecule, and Ω is the volume associated with the generalized coordinates of a single molecule. Formally,

$$Z^{\text{IG}} = \int \exp[-\beta\phi_{\text{intra}}(\mathbf{q})] d^{3n_a} q \quad (6)$$

$$\Omega = \int d^{3n_a} q \quad (7)$$

where $\phi_{\text{intra}}(\mathbf{q})$ is the intramolecular contribution to the potential energy of a single molecule with generalized coordinates given by the vector \mathbf{q} . For the special case of an atomic species, $\phi_{\text{intra}} = 0$ and $n_a = 1$, so that $Z^{\text{IG}} = \Omega = V$.

Given the probability distribution function in eq 5, a set of Metropolis-like acceptance rules can be derived for making changes in the value of λ . Assume that state m consists of N_m integer molecules and one fractional molecule. If a change in the coupling parameter occurs such that $\lambda > 1$, then the existing fractional molecule is converted to an integer molecule, and a new fractional molecule is added to the system by randomly selecting a position and assigning the new fraction molecule a coupling parameter of $\lambda - 1$. The new state n contains $N_n = N_m + 1$ integer molecules and a single fractional molecule. The move is accepted with a probability reminiscent of a grand canonical ensemble insertion move

$$\Pi_{mn}^{\text{ins}} = \min\left(1, \frac{\alpha_{nm} \Omega}{\alpha_{mn} Z^{\text{IG}}} \frac{f\beta V}{N_m + 1} \exp[\eta_n - \eta_m] \exp[-\beta(\phi_n - \phi_m)]\right) \quad (8)$$

Likewise, when a transition occurs for a system of N_n integer molecules and one fractional molecule such that $\lambda < 0$, the fractional molecule is removed from the system, and one of the remaining integer molecules is randomly assigned a coupling parameter value of $1 + \lambda$. This molecule thus becomes the new fractional molecule. The transition probability for such a move from state n to state m is similar to the grand canonical deletion move probability and is given by

$$\Pi_{nm}^{\text{del}} = \min\left(1, \frac{\alpha_{mn} Z^{\text{IG}} N_n}{\alpha_{nm} \Omega f\beta V} \exp[\eta_m - \eta_n] \exp[-\beta(\phi_m - \phi_n)]\right) \quad (9)$$

For moves in which λ changes but is still between (0,1), the acceptance probability is

$$\Pi_{mn}^{\lambda} = \min(1, \exp[\eta_n - \eta_m] \exp[-\beta(\phi_{\text{inter},n} - \phi_{\text{inter},m})]) \quad (10)$$

where $\phi_{\text{inter},m}$ and $\phi_{\text{inter},n}$ are the intermolecular contributions to the potential energy of states m and n , respectively. Note that no intramolecular energies are altered by this type of move.

Finally, thermal equilibration moves at constant values of λ occur. Any MC sampling method desired can be used for these moves.^{42,43} In the present work, a hybrid Monte Carlo procedure was used⁴⁴ in which velocities are selected from a Maxwell–Boltzmann distribution, followed by a number of microcanonical MD steps. The new positions are accepted with probability

$$\Pi_{mn}^{\text{HMC}} = \min(1, \exp[-\beta(H_n - H_m)]) \quad (11)$$

where H_m and H_n are the Hamiltonians of the system in states m and n .

Note that in eqs 10 and 11 the attempt probabilities α_{nm} and α_{mn} are not included, because for these moves they are equal and thus cancel out. That is, there is no bias as to whether an attempt to increase or decrease λ is made. However, the weighting factors η do influence the likelihood of a successful change in λ and are optimized for performance, as discussed below. On the other hand, the attempt probabilities have been explicitly included in eqs 8 and 9, because moves in which the number of integer molecules changes are compound moves. The probabilities associated with these moves consist of the product of the symmetric probability of increasing or decreasing the value of λ and the asymmetric probability of adding or removing a molecule from the system. There is a great deal of flexibility in the choice of these latter probabilities. In this work, when λ exceeds unity and a new fractional molecule is added to the system, a rigid conformation is chosen from a “reservoir” of ideal gas molecules generated “on the fly” during the simulation.^{40,45} These conformations occur with probability

$$p^{\text{IG}} = \frac{\exp[-\beta\phi_{\text{intra}}(\mathbf{q})]\Omega}{Z^{\text{IG}}} \quad (12)$$

Given such a conformation, a random position and orientation of this molecule is chosen, and the new fractional molecule is inserted into the system. The attempt probability for this move is $\alpha_{mn} = p^{\text{IG}}$. For the reverse move when λ falls below zero, an existing integer molecule is chosen at random to become the new fractional molecule. The attempt probability for this move is $\alpha_{nm} = 1$. Thus the ratio of attempt probabilities in eqs 8 and 9 is

$$\frac{\alpha_{nm}}{\alpha_{mn}} = \frac{Z^{\text{IG}}}{\exp[-\beta\phi_{\text{intra}}]\Omega} \quad (13)$$

Using this method for generating conformations simplifies the acceptance ratios in eqs 8 and 9 to

$$\Pi_{mn}^{\text{ins}} = \min\left(1, \frac{f\beta V}{N_m + 1} \exp[\eta_n - \eta_m] \exp[-\beta(\phi_{\text{inter},n} - \phi_{\text{inter},m})]\right) \quad (14)$$

and

$$\Pi_{mn}^{\text{del}} = \min\left(1, \frac{N_n}{f\beta V} \exp[\eta_m - \eta_n] \exp[-\beta(\phi_{\text{inter},m} - \phi_{\text{inter},n})]\right) \quad (15)$$

where $\phi_{\text{inter},m}$ and $\phi_{\text{inter},n}$ are only the intermolecular contributions to the potential energy for states m and n , respectively. Note that the intramolecular potential energies cancel out, given that the fractional molecule's intramolecular degrees of freedom are sampled from eq 12 and the existing molecules are unperturbed by the move. Note that additional biasing moves are not used, given the gradual coupling of the fractional molecule with the system and the use of the extended system biasing factors η .

3.2. Osmotic Ensemble. The CFC MC method can also be applied to the osmotic ensemble^{46–48} in which for a c -component system with k solute molecules and $c - k$ solvent molecules, the following variables are fixed: $f_1, f_2, \dots, f_k, N_{k+1}, N_{k+2}, \dots, N_c, P, T$. The number of molecules of species $1 - k$ will fluctuate in this ensemble. Such an ensemble is particularly effective in simulating solubilities, especially for light species in low-volatility solvents. Here, we develop osmotic CFC MC acceptance rules for the binary case where species 1 is the solute and species 2 is the solvent. Generalizing to multicomponent systems is relatively straightforward.

The probability density of a system of volume V containing N_1 integer solute molecules having generalized coordinates \mathbf{q}_1 , a fractional molecule in state j at \mathbf{q}_j with coupling parameter λ and associated bias factor η_j , and N_2 solvent molecules at \mathbf{q}_2 is⁴⁶

$$\rho(N_1, V, \eta_j, \mathbf{q}_1, \mathbf{q}_j, \mathbf{q}_2; f_1, N_2, P, T) \propto \frac{V^{N_2}}{N_1!} \times \frac{(\beta f_1 V)^{N_1} (V \exp[\eta_j])}{(Z^{\text{IG}}/\Omega)^{N_1+1}} \exp[-\beta PV] \exp[-\beta(\phi(\mathbf{q}_1, \mathbf{q}_j, \mathbf{q}_2))] \quad (16)$$

In eq 16, the dependence of the potential energy on the coordinates of all species is explicitly noted.

Changes in λ that result in an increase or decrease in the number of integer molecules are accepted with the same probability as that in eqs 14 and 15, assuming the ratio of attempt probabilities is given by eq 13. Likewise, changes in λ that do not alter the number of integer molecules are accepted according to eq 10, and thermal equilibration moves are performed with hybrid Monte Carlo (HMC) and accepted according to eq 11.

Random changes in volume are performed to maintain a constant pressure. These moves occur at constant λ and $N_{\text{tot}} = N_1 + N_2 + 1$, with acceptance probability given by

$$\Pi_{mn}^{\text{vol}} = \min\left(1, \exp\left[-\beta\left\{P(V_n - V_m) + (\phi_n - \phi_m) - N_{\text{tot}} k_B T \ln \frac{V_n}{V_m}\right\}\right]\right) \quad (17)$$

4. Simulation Details

4.1. Lennard-Jones. The CFC MC method was first implemented and tested for grand canonical simulations of the Lennard-Jones fluid, in which the following cut and shifted potential was used

$$\phi_{\text{LJ,cut-shift}}(r) = \begin{cases} \phi_{\text{LJ}}(r) - \phi_{\text{LJ}}(r_c) & r \leq r_c \\ 0 & r > r_c \end{cases} \quad (18)$$

where the cutoff distance was taken as $r_c = 2.5\sigma$ and ϕ_{LJ} is given by

$$\phi_{\text{LJ}}(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^6 \right] \quad (19)$$

To minimize problems with repulsive overlap, the following scaled potential⁴⁹ was used to describe the interaction between a fractional species and the integer species

$$\phi_f(r_{if}, \lambda) = \lambda 4\epsilon \left\{ \frac{1}{\left[\xi(1 - \lambda)^2 + \left(\frac{r_{if}}{\sigma}\right)^6 \right]^2} - \frac{1}{\left[\xi(1 - \lambda)^2 + \left(\frac{r_{if}}{\sigma}\right)^6 \right]} \right\} \quad (20)$$

As with eq 1, r_{if} is the distance between an integer species i and fractional species f , while ξ is an adjustable parameter that was set to 0.5. This scaled potential has the correct behavior at the limits of $\lambda = 0$ and $\lambda = 1$ but is well-behaved at very close overlap, as often occurs when inserting a species.

Due to the relative ease with which λ transitions occur in this system, it was found that bias factors were not necessary. Simulations were carried out at three state points with and without bias factors, and the results were identical. Thus for most of the Lennard-Jones results reported here, bias factors were not used. Sixty percent of moves were HMC moves used to thermally equilibrate the system, while 40% of the moves were used to change the value of the coupling strength between the fractional atom and the rest of the system. Each HMC move involved selection of velocities from a Maxwell-Boltzmann distribution followed by 5 microcanonical MD steps. Changes in λ were made uniformly up to a maximum change of $\Delta\lambda_{\text{max}}$. The HMC time step and $\Delta\lambda_{\text{max}}$ were adjusted during equilibration to achieve 50% move acceptance rates. Equilibration consisted of 2 million MC steps, after which averages were taken over 10 million MC steps. Cubic boxes with edge lengths of 5.87, 9, and 13.2 σ were used. Reduced units⁴² were used to enable easy comparison with previous work and an equation of state. The reduced chemical potential was

$$\mu^* = \mu/\epsilon - k_B T/\epsilon \ln(\Lambda^3/\sigma^3) \quad (21)$$

where Λ is the de Broglie wavelength. When computing the average composition, only the integer number of molecules was considered.

To enable comparison with the CFC-GCMC simulations, the GCMD method of Lo and Palmer was implemented and run for the Lennard-Jones system, as was the hybrid method of Boinepalli and Attard. In addition, traditional GCMC simulations were carried out.

4.2. Water. Traditional GCMC and CFC-GCMC simulations were conducted for a fully flexible model of water in which harmonic potentials were used for bond lengths and bond angles. A flexible model was required, since hybrid Monte Carlo requires the use of reversible integrators; traditional constraint dynamics methods are not reversible. The nominal bond lengths and angles as well as partial charges and Lennard-Jones parameters were taken from the SPC water model.⁵¹ Because SPC is a rigid model, bond length and bond angle force constants were taken from the CHARMM22 force field⁵⁰ and set to 450 kcal/(mol Å²) and 55 kcal/(mol rad²), respectively.

As described below, simulations of this water model were also attempted with the method of Boinepalli and Attard, but convergence was difficult. To enable comparison between this method and the CFC approach, a “soft” water model was also simulated, in which all potential parameters were the same except the partial charges were removed from the oxygen and hydrogen atoms. By removing the partial charges, the soft water model does not hydrogen bond, and thus it is much easier to insert or delete molecules.

For the traditional GCMC simulations, 60% of the moves were HMC moves, 20% were insertion moves, and 20% were deletion moves. For CFC-GCMC, 60% of the moves were HMC moves, while 40% were random changes in λ . A typical GCMC simulation consisted of 1 million equilibration moves followed by 2 million moves for the production run.

The simulations for the “real” water model were conducted at 298 K and a fugacity of 3746.8 Pa in a cubic box with an edge length of 13.9623 Å. This corresponds to the conditions used previously by Mezei^{52,53} in cavity bias GCMC simulations of a rigid water model. “Soft” water simulations were carried out at 300 K and fugacities ranging from 100 to 300 Pa. Electrostatics were treated using an Ewald sum with a real space cutoff of 6.5 Å and tin foil boundary conditions. A switching function of the following form was used for the Lennard-Jones interactions between all atoms in the system

$$\phi_{\text{switch}}(r) = \begin{cases} \phi_{ij}(r) & r < r_{\text{on}} \\ \phi_{ij}(r) \times \frac{(r_{\text{off}}^2 - r^2)(r_{\text{off}}^2 + 2r^2 - 3r_{\text{on}}^2)}{(r_{\text{off}}^2 - r_{\text{on}}^2)^3} & r_{\text{on}} \leq r \leq r_{\text{off}} \\ 0 & r > r_{\text{off}} \end{cases} \quad (22)$$

where $\phi_{ij}(r)$ denotes the full interaction between two atoms i and j with a distance of r , and r_{on} and r_{off} were set to be 6.0 Å and 6.5 Å, respectively. The switching function

guarantees that the potential energy and force are continuous over the entire range of r . The scaled potential given in eq 20 was used for interactions between the fractional molecule and integer molecules. In addition, the partial charges on the fractional molecule were scaled as $q_f = \lambda^5 q_i$, where q_f and q_i are the partial charges on a fractional and integer atom, respectively. This nonlinear scaling was found to moderate strong electrostatic interactions that result when molecules were inserted close to an existing species. To avoid singularities, a hard core cutoff of 1.0 Å was used between fractional and integer atoms. If an insertion of a fractional molecule resulted in two atoms coming closer than 1.0 Å, then the insertion was rejected. This hard core cutoff was required for the SPC model because the hydrogen atoms are modeled with only a partial point charge. The lack of a van der Waals radius can cause singularities during insertions. If all atoms in the system have van der Waals radii, then the hard core cutoff will be unnecessary.

A reservoir of water molecules used for insertions was generated by simulating 500 ideal gas waters using HMC in parallel with the CFC-GCMC simulations. When an insertion is required, a configuration is randomly chosen from the reservoir, given a random orientation, and then randomly inserted into the system. After 100 MC steps, additional HMC moves are performed on the reservoir to ensure new conformations are generated such that the distribution in eq 12 is properly sampled. It should be noted that for the simple water model used here, the distribution in eq 12 could be sampled analytically with no need to generate a reservoir of configurations. This is not the case, however, for more complex solutes having intramolecular nonbonded interactions. It was verified that the reservoir method used here did indeed satisfy the analytic distribution for this model consistent with eq 12.

As with the Lennard-Jones case, 5 MD time steps were used for each HMC move, with the time step adjusted during equilibration to obtain roughly 50% acceptance rates. Random changes in λ were made, with the maximum change $\Delta\lambda_{\text{max}}$ adjusted during equilibration to achieve 50% acceptance of these moves. It was found that $\Delta\lambda_{\text{max}}$ was typically about 0.2, and the HMC time step was about 1.0 fs. During the production phase, success rates for λ changes averaged about 50%, while HMC moves were successful about 80% of the time.

As described below, it was found that bias factors were necessary to adequately simulate the real water system. Ideally, the bias factors should yield a flat distribution in λ values, and this criterion can be used to optimize the bias factors during the equilibration stage of the simulation. The simple updating method of Smith and Bruce⁵⁴ and the method of Wang and Landau³⁹ were both investigated. It was found that the Wang–Landau approach was superior. In this approach, the (0,1) range of λ is divided into ten adjacent bins of [0,0.1],[0.1,0.2],..., [0.9,1.0], and each bin j is assigned a bias factor η_j . Initially, all bias factors are set to $\eta_j = 0$. During the equilibration phase, each time a value of λ falls within the range of a bin j , the corresponding value of η_j is modified according to

$$\eta_j = \eta_j - \ln(v) \quad (23)$$

where v acts as a scaling factor. Initially, $\ln(v)$ was set to unity. One hundred iterations were performed, each consisting of 10 000 steps. During this phase, histograms were collected for the number of times a given λ bin was visited. After each iteration, the distribution of λ was checked, and if the minimum probability of each λ bin was greater than 1%, then v was modified according to the following expression

$$v = \sqrt{v} \quad (24)$$

The histograms were zeroed after each iteration.

4.3. CO₂ and Ethanol. To test the osmotic CFC MC method, the solubility of CO₂ in ethanol was computed. This is an ideal test system to study, since it was the basis of the 2004 Industrial Fluid Properties Simulation Challenge.⁵⁵ As a result, there are two sets of previous simulations that used the same force field against which comparison can be made. The first set of simulations was conducted by Errington and co-workers and utilized a transition matrix MC method.⁵⁶ The second simulation study was performed by Zhang and Siepmann and utilized Gibbs ensemble MC.⁵⁷ These two studies had consistent results, so agreement with these calculations is a strong indication of the validity of the CFC method.

Force field parameters were the same as that used in the two previous studies^{56,57} with the exception that the bond lengths and angles for both CO₂ and ethanol were flexible in the CFC calculations but were held rigid in the other two works. Force constants were taken from the CHARMM22 force field.⁵⁰ The use of flexible models is expected to have a minimal impact of the phase equilibria properties. A flexible model was used in the CFC simulations because HMC was used for sampling, which requires the use of reversible and symplectic integration schemes. This is nontrivial to accomplish with constrained degrees of freedom, and so fully flexible models were used. A switching function, eq 22, was used for the Lennard-Jones interactions, with $r_{\text{on}} = 14.0 \text{ \AA}$ and $r_{\text{off}} = 14.19 \text{ \AA}$. The Ewald sum method was used with a real space cutoff of 14.19 \AA and $\kappa^{-1} = 0.2857 \text{ \AA}^{-1}$.

To begin the calculations, NPT simulations were run for 216 ethanol molecules at 323 K. Ninety-eight percent of the moves were HMC moves, each consisting of 5 MD time steps. The remaining 2% of the moves were volume change moves. One million equilibration moves followed by 2 million production moves were made. An HMC time step of 1.1 fs resulted in about 31% of the HMC moves being accepted, while a maximum volume change of 260 \AA^3 resulted in 56% of the volume moves being accepted.

Once it was confirmed that the computed densities matched those in the literature, isotherms were calculated using osmotic CFC MC simulations at varying CO₂ fugacity and a fixed number of ethanol molecules, temperature, and pressure. In a typical application of the osmotic ensemble, the solvent is usually much less volatile than the solute, such that the total pressure of the system is equal to the partial pressure of the solute. The solute fugacity can be directly related to the pressure via $f_i = \phi_i P$ where ϕ_i is the solute fugacity coefficient. For the calculations here, CO₂ is

certainly much more volatile than ethanol. Instead of performing calculations at varying pressure and fugacity, isotherms were computed at constant pressure and varying fugacity. In all cases, $P \geq f_i/\phi_i$, where ϕ_i was calculated for pure CO₂ at a given T and P from the Peng–Robinson equation of state. Since the simulations were of a binary single phase system, the phase rule was not violated. However, in a two phase vapor–liquid equilibrium experiment, a third sparingly soluble component (such as helium) would need to be added to the vapor phase to independently vary pressure and CO₂ fugacity. Not surprisingly, the calculations show that the solubility of CO₂ is a very weak function of total pressure, and so isotherms computed in this manner can be compared with experiment with little error.

The attempt probabilities for translation, volume change, and λ change moves were 60%, 2%, and 38%, respectively. The Wang–Landau scheme was used to optimize the bias factors, with anywhere from 20 to 100 iterations used to obtain an acceptably flat distribution of λ values. As an example, for a simulation at 323 K and a total pressure of 20 bar with a fugacity of 10 bar, the weighting factors for each of the ten λ bins were 0.0, 0.511, 1.47, 1.784, 1.613, 1.289, 0.845, 0.397, -0.232 , and -1.194 , respectively. The resulting λ distribution was quite flat; the ratio of the most probable to least probable λ bin was roughly 1.6.

5. Results

5.1. Lennard-Jones. A series of CFC GCMC simulations was conducted for the Lennard-Jones fluid at T^* values ranging from 0.769 to 1.25 and ρ^* values ranging from 0.005 up to 0.886. The results were compared against the accurate equation of state of Johnson et al.,⁵⁸ standard GCMC simulations carried out as part of this work, and the simulation results reported by Lo and Palmer.²⁹ We also implemented the methods reported by Lo and Palmer²⁹ and Boinepalli and Attard³⁵ and conducted simulations with these algorithms. This was done to compare the efficiency and accuracy of these methods against the CFC method. The only difference between our implementation of the Boinepalli and Attard method and theirs is that they used stochastic temperature control, while we used HMC. This should not alter the results.

Figure 1 shows the isotherm calculated at $T^* = 1.0$ and is typical of the type of agreement obtained between the different methods. A complete listing of all the results is provided in Table 1. With the exception of our implementation of Lo and Palmer's method, all the results agree with each other and the equation of state within numerical accuracy. The results reported by Lo and Palmer agree with our calculations using their method at low densities, but at higher densities our implementation of their method systematically overpredicts the densities. It is possible there is a mistake in our implementation of the Lo and Palmer method, although it is unclear why our implementation agrees at low density but not high density. We are unaware of any other publications in which the Lo and Palmer method has been implemented and tested. It would be interesting to see if other groups can reproduce the original results.

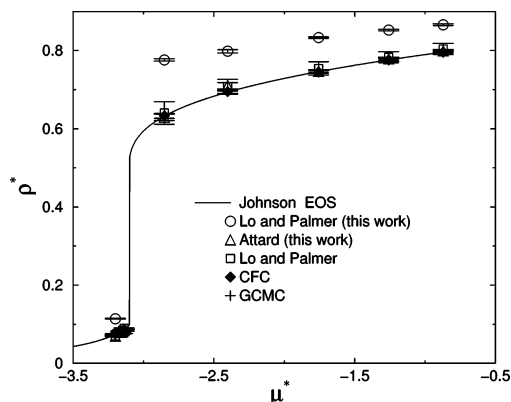


Figure 1. Computed isotherms for the Lennard-Jones fluid at $T^* = 1.0$ using various methods. The line is from the equation of state of Johnson et al.⁵⁸ The circles are from this work using the method of Lo and Palmer.²⁹ The triangles are from this work using the method of Boinepalli and Attard.³⁵ The squares are the results reported by Lo and Palmer.²⁹ The filled diamonds are from CFC GCMC simulations, while pluses are from standard GCMC simulations. All the numerical values are also shown in Table 1. With the exception of our implementation of Lo and Palmer's method, all results agree with each other and the equation of state.

In any case, these results indicate that the CFC algorithm is correct and has been properly implemented in our locally developed software package. For the Lennard-Jones system at these state points, there is no advantage to using the CFC GCMC method (or the other slow growth open system methods). Standard GCMC is sufficient for computing volumetric properties of the Lennard-Jones fluid under these conditions. At lower temperatures and higher densities, however, there may be a need to use more sophisticated techniques. These state points were not explored, because the goal was to test the method under conditions where standard techniques give reliable results to compare against.

It is interesting to investigate the role the bias potentials η_j play on the results and the performance of the method. As mentioned earlier, some test simulations were run with and without the use of the extended system bias potentials. As shown in Table 2, the final result for the Lennard-Jones fluid under these conditions is independent of whether or not the bias potentials are used. As shown below, however, the use of bias potentials for systems more complex than Lennard-Jones leads to more efficient sampling. The CFC results agree with the GCMC results regardless of whether or not a bias potential is used. We note one exception. At the lowest temperature, the GCMC and the CFC results are very close but fall just outside of the error bars. It may be that the uncertainty of these calculations has been underestimated. Figure 2 shows the distribution of λ with and without bias potentials for the Lennard-Jones system at $T^* = 0.769$, $L/\sigma = 5.87$, and $\mu^* = -0.816$. Although ultimately both calculations yield the same density, the simulations without the bias potentials have very few intermediate values of λ . The fractional molecule is almost always at states where $\lambda < 0.3$ and $\lambda > 0.6$. With the use of bias potentials, however, the fractional molecule spends roughly the same amount of

time at all values of λ . The latter situation leads to much more reliable sampling.

As an aside, it turns out that the pseudopotential in the Boinepalli and Attard method can be reduced to a specific value of the CFC bias factor for the Lennard-Jones system via the following relation

$$\exp(\eta) = \frac{\exp[\beta\mu\lambda]V^{\lambda-1}N!}{\Lambda^{3(\lambda-1)}(N+\lambda)!} \quad (25)$$

where N is the number of integer molecules present in the system. This means that the Boinepalli and Attard method and the CFC method are equivalent for the Lennard-Jones system if the bias factors are given by eq 25. The advantage of the CFC approach is that the bias factors can be adjusted in a self-adaptive way to optimize the performance of the algorithm.

5.2. Water. Water provides a more stringent test of the CFC method. For this system, hydrogen bonding and orientational order make random insertions and deletions of molecules more difficult. To begin these simulations, the “soft” water model was initially simulated at low densities. Under these conditions, the fluid should exhibit ideal gas behavior, and so algorithms can easily be checked to ensure they capture this limiting behavior. In addition to CFC GCMC calculations, standard GCMC simulations were also carried out on this system, as were simulations using our implementation of the Boinepalli and Attard method. The simulations were conducted at 300 K in a cubic box with edge lengths of 1000 Å and fugacities ranging from 100 to 300 Pa. The results are shown in Figure 3. As can be seen, all results agree. As with the case of the Lennard-Jones simulations, the use of bias factors improves sampling somewhat but is unnecessary to get accurate results. That is, simulations were performed with and without bias factors, and the results were identical. This is not surprising, given the low density and lack of hydrogen bonding in this system.

Next, simulations were carried out for the flexible SPC model of water using our implementation of Boinepalli and Attard's method and CFC GCMC. Mezei^{52,53} has conducted cavity-bias GCMC simulations on a rigid SPC water model using a cubic box with edge lengths of 13.9623 Å at $T = 298$ K and $B = -6.0$, where

$$B = \ln(\beta f V) \quad (26)$$

This corresponds to a fugacity of 3746.9 Pa. The number of water molecules at this condition was found to be 90.30 ± 0.9 ,^{52,53} which corresponds to a water density of 0.992 ± 0.0099 g/cm³.

Using our implementation of Boinepalli and Attard's method, calculations were started at the same temperature and box volume used by Mezei, but with only 60 water molecules initially in the box. This corresponds to an initial density of 0.659 g/cm³. After 3 million steps, it was observed that the number of water molecules in the system had not changed; the fractional molecule became “stuck” at an intermediate value. Figure 4 shows the composition and probability distribution of λ . It can be seen that λ lies mostly in the range of 0.3–0.5 and never exceeds 0.65 or goes lower

Table 1. Results from Simulations of the Lennard-Jones Fluid^a

T^*	μ^*	ρ^* EOS	ρ^* GCMC	ρ^* LP	ρ^* LP (pw)	ρ^* BA (pw)	ρ^* CFC
1.0	-3.2	0.0744	0.0729 (2)	0.0699 (6)	0.114 (1)	0.0698 (9)	0.075 (1)
1.0	-3.162	0.0817	0.0807 (4)	0.0808 (8)	—	0.079 (2)	0.0807 (8)
1.0	-3.150	0.0844	0.0823 (5)	0.0844 (18)	—	0.080 (2)	0.083 (1)
1.0	-3.139	0.0871	0.0857 (5)	0.0896 (10)	—	0.084 (1)	0.086 (1)
1.0	-2.852	0.632	0.638 (1)	0.640 (29)	0.776 (3)	0.629 (4)	0.633 (6)
1.0	-2.403	0.693	0.697 (1)	0.707 (19)	0.798 (4)	0.710 (4)	0.695 (5)
1.0	-1.757	0.746	0.750 (1)	0.754 (17)	0.833 (2)	0.747 (2)	0.745 (3)
1.0	-1.258	0.776	0.780 (1)	0.783 (14)	0.852 (2)	0.779 (2)	0.776 (4)
1.0	-0.87	0.796	0.799 (1)	0.804 (14)	0.866 (2)	0.799 (2)	0.796 (2)
0.769	-4.127	0.00501	0.00510 (1)	0.0050 (0)	0.00489 (9)	0.0050 (1)	0.0048 (2)
0.769	-3.797	0.00802	0.00800 (1)	0.0081 (0)	0.0078 (2)	0.0082 (2)	0.00801 (6)
0.769	-3.646	0.01004	0.01003 (2)	0.0099 (0)	0.0097 (3)	0.0100 (3)	0.00996 (8)
0.769	-3.583	0.01105	0.01101 (2)	0.0108 (0)	0.0108 (3)	0.0116 (3)	0.01100 (8)
0.769	-2.999	0.779	0.784 (1)	0.785 (20)	0.864 (3)	0.789 (3)	0.787 (9)
0.769	-2.868	0.788	0.794 (1)	0.794 (11)	0.871 (3)	0.799 (3)	0.793 (8)
0.769	-2.727	0.798	0.803 (1)	0.804 (11)	0.879 (3)	0.801 (3)	0.801 (4)
0.769	-0.816	0.886	0.890 (1)	0.895 (21)	0.934 (2)	0.889 (2)	0.886 (4)
1.1	-3.35	0.0815	0.0810 (3)	—	—	0.077 (2)	0.0803 (7)
1.15	-3.3	0.106	0.104 (1)	—	—	0.109 (4)	0.104 (1)
1.2	-3.2	0.151	0.146 (1)	—	—	0.148 (4)	0.142 (4)
1.25	-3.14	0.189	0.184 (1)	—	—	0.177 (6)	0.185 (2)
1.25	-3.03	0.244	0.241 (2)	—	—	0.249 (2)	0.242 (4)
1.25	-2.95	0.291	0.301 (3)	—	—	0.293 (4)	0.302 (5)
1.25	-2.9	0.321	0.343 (3)	—	—	—	0.332 (8)
1.25	-2.8	0.381	0.406 (4)	—	—	—	0.410 (5)
1.25	-2.76	0.403	0.432 (2)	—	—	—	0.431 (8)
1.25	-2.73	0.419	0.443 (1)	—	—	—	0.446 (3)
1.175	-2.68	0.511	0.529 (1)	—	—	0.526 (2)	0.523 (6)
1.15	-2.63	0.552	0.562 (2)	—	—	0.552 (6)	0.560 (6)
1.1	-2.55	0.610	0.616 (2)	—	—	0.611 (3)	0.619 (3)

^a Abbreviations are as follows. EOS: Equation of State of Johnson et al.;⁵⁸ GCMC: Grand Canonical Monte Carlo Simulation, present study; LP: results of Lo and Palmer;²⁹ LP (pw): results of the present work using the Lo and Palmer algorithm; BA (pw): results of the present work using the Boinepalli and Attard³⁵ algorithm; CFC: Continuous Fractional Component GCMC Calculations. A “—” indicates no results available.

Table 2. Simulation Results for the Lennard-Jones Fluids Using Standard GCMC Simulations as Well as CFC GCMC with and without Extended System Bias Potentials^a

T^*	L/σ	μ^*	ρ^* (GCMC)	ρ^* (CFC, no η)	ρ^* (CFC, with η)	η
0.769	5.87	-0.816	0.890 (1)	0.883 (3)	0.886 (4)	1.8, 3.5, 5.6, 5.9, 5.5, 4.8, 3.8, 2.7, 1.4, 0.0
1.0	9	-3.52	0.0416 (1)	0.0417 (2)	0.0418 (2)	0.0, 0.1, 0, 0, -0.3, -0.5, -0.8, -1.1, -1.5, -1.7
1.25	9	-3.03	0.241 (2)	0.242 (4)	0.242 (7)	0.7, 0.9, 0.8, 0.7, 0.5, 0.4, 0.2, 0.0, 0.8, 0.6

^a Numbers in parentheses are the standard deviations in the last digit.

than 0.1. Apparently, under the conditions investigated here, the pseudopotential used for insertions and deletions in this approach is incapable of “pushing” the fractional molecule out of the intermediate λ range. This does not indicate that the Boinepalli and Attard method is flawed; as they point out one can add an additional bias potential to the method to improve sampling. What it does indicate is that the addition or deletion of water molecules is difficult with these types of methods.

Standard GCMC as well as CFC GCMC calculations were then run at this state point, with 60 and 90 water molecules initially placed in the simulation box. This corresponds to initial densities of 0.659 g/cm³ and 0.989 g/cm³. The convergence of these simulations is shown in Figure 5. Regardless of the initial density, both methods converge to the same density of 1.03 ± 0.01 g/cm³, again indicating that

the CFC method gives equivalent results to GCMC. These densities are slightly higher than that reported by Mezei. The difference is probably due to the fact that an Ewald sum was used for the present calculations, while a charge cutoff was used by Mezei. In addition, Mezei used a rigid model for water, while the water model used here had flexible bond lengths and angles.

Although the CFC and standard GCMC simulations yield identical densities, the sampling efficiency of the CFC method is much better than standard GCMC. At the final water density, less than 0.01% of water insertion or deletion moves were accepted for the standard GCMC simulations. This means the composition changes once on average every 10 000 insertion or deletion attempts. With the CFC method, however, 2.4% of the moves involving a change in λ resulted in the creation or removal of an integer molecule, and 52%

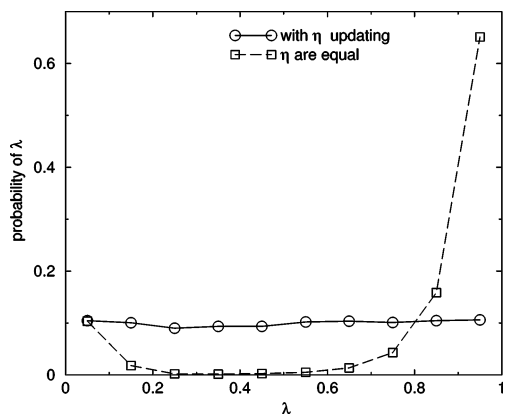


Figure 2. The probability distribution of λ during CFC GCMC simulations of the Lennard-Jones system with and without biasing potentials. The simulations are performed on a system at $T = 0.769$, $L/\sigma = 5.87$, and $\mu^* = -0.816$. The lines are to guide the eyes. Circles are for calculations using bias potentials, and squares are for unbiased simulations. The use of bias potentials promotes a more uniform λ distribution, while the simulations without bias potentials tend to become “stuck”, especially near $\lambda = 1$.

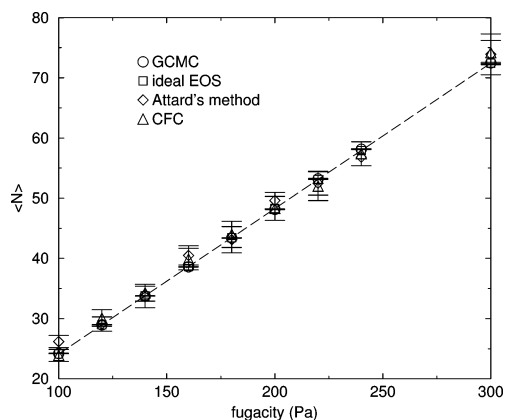


Figure 3. The average number of integer molecules as a function of fugacity for a “soft” water model at 300 K and very low density. The circles are from standard GCMC simulations, the squares are for the ideal gas, the diamonds are results from our implementation of Boinepalli and Attard’s method³⁵ extended to molecules, and the triangles are from the CFC GCMC calculations. The line is a guide for the eye.

of all attempted changes in λ were accepted. This means the composition changes once on average every 42 attempts to change λ . Figure 6 shows the instantaneous values of λ over a portion of the water simulation. Clearly, λ traverses the entire range from fully coupled to ideal gas many times during the simulation. Figure 7 shows the initial convergence characteristics of the CFC simulation starting from 90 water molecules (left) and the distribution of λ values during the final 2 million moves of the production run (right). While the λ distribution is peaked near the ends, all values are visited frequently enough to maintain good transition frequencies. Better updating schemes for the bias potentials would likely improve the distribution of λ even more.

Mezei also used cavity biased GCMC to calculate the density of the rigid SPC water model at $T = 298$ K and $B =$

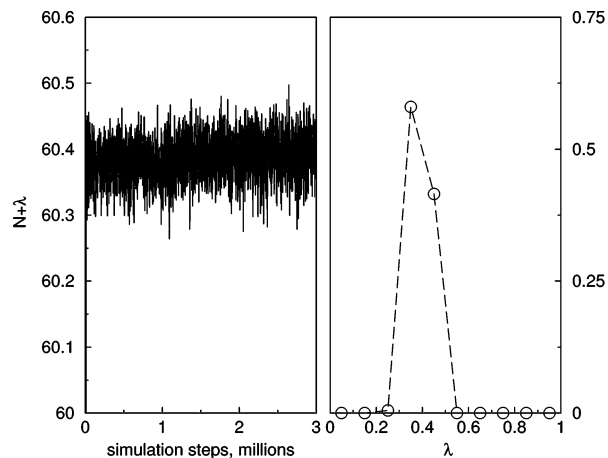


Figure 4. The real number of $N + \lambda$ fluctuation and the probability of λ for the flexible SPC water model from the method by Boinepalli and Attard.³⁵ The dashed line is to guide the eye.

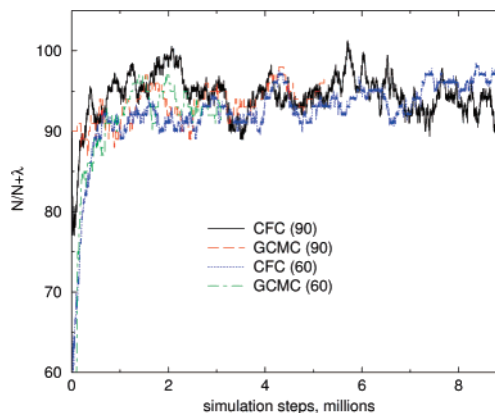


Figure 5. The real number of molecules $N + \lambda$ from the CFC method and the integer number of molecules N from GCMC simulation for the flexible SPC water model. The solid black and long-dash red lines correspond to the CFC and standard GCMC simulations starting from 90 water molecules, respectively. The dotted blue and dot-dash green lines are for the CFC and GCMC simulations starting from 60 water molecules, respectively.

-5.7 , which corresponds to $f = 5057.7$ Pa. At these conditions, the computed density was 1.006 ± 0.010 g/cm³, which is 1.4% higher than the density computed at $f = 3746.9$ Pa. Using CFC GCMC, the density was found to be 1.046 ± 0.007 g/cm³ or 1.5% higher than that calculated at the lower fugacity. Again, there is a slight difference in the absolute values, but the relative increase in density is nearly the same. Interestingly, standard unbiased GCMC becomes extremely inefficient at this state point; we were unable to obtain reliable statistics due to the vanishingly low probabilities of successful insertions and deletions.

Obviously, one can develop powerful biasing schemes that will improve the insertion and deletion success rates of GCMC, and Mezei’s cavity biasing approach is one such method. The CFC method can be thought of as another biased MC move, but one that is self-adapting in terms of the use of gradual insertions and deletions as well as bias factors that promote these gradual transitions. It is thus

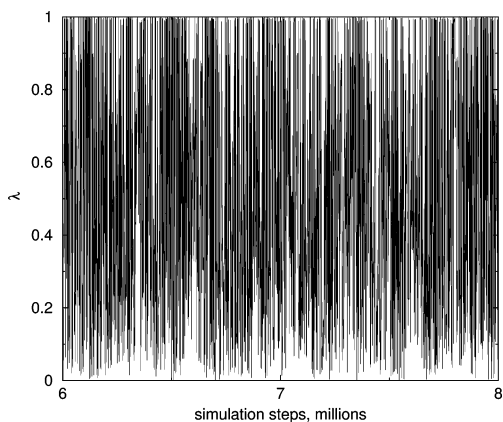


Figure 6. Instantaneous values of λ during the simulation of SPC water. For clarity, only the last 2 million steps of the simulation are shown. λ clearly traverses the entire range of values during the simulation.

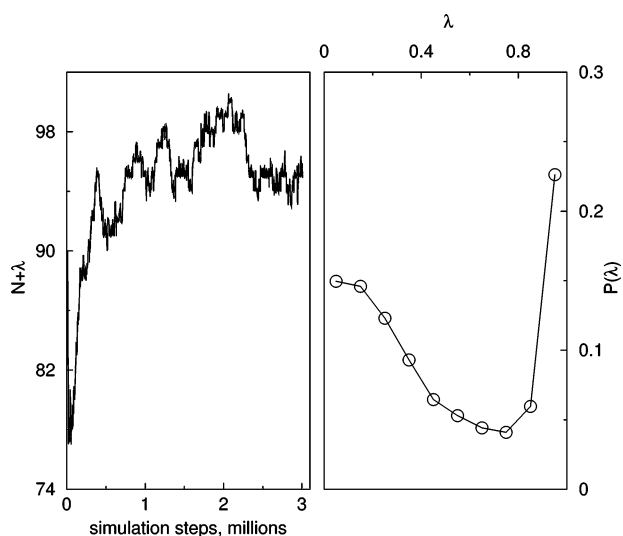


Figure 7. Left: The number of water molecules in the simulation box (integer plus fractional) as a function of simulation duration. Right: The average probability of a given value of λ , using the updating scheme of Wang and Landau³⁹ to modify the importance weighting factors η . During the last 2 million steps, a total of 798 335 attempts are made to change λ , out of which just over 51% are accepted. Out of these changes in λ , a molecule was successfully removed from the system 10 727 times, and a molecule was added to the system a total of 10 728 times.

relatively straightforward to implement and requires little a priori tuning. We also note that, as expected, the equilibrium properties do not depend on the exact values of the weighting factors used. Two different sets of weighting factors obtained from different updating schemes yielded the same density for the SPC water system. Obviously, very poor choices for weighting factors will have a negative effect on the calculations, but it appears that the method is robust enough such that a fairly wide range of weighting factors will achieve favorable results.

5.3. CO₂ and Ethanol. CFC osmotic simulations were performed to compute the solubility of CO₂ in ethanol. Isothermal–isobaric simulations were first performed on the pure ethanol system at 323 K and 0.294 bar to compare

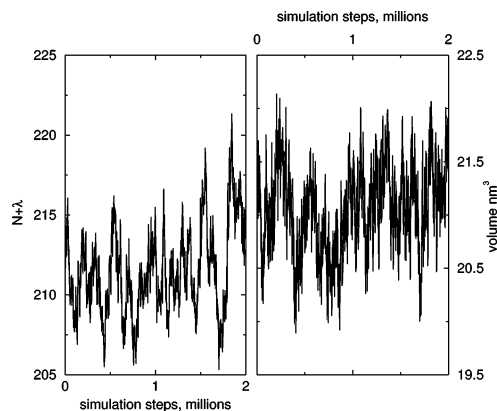


Figure 8. Left: The total number of CO₂ molecules (integer plus fractional) in ethanol at $T = 323$ K, $P = 40$ bar, and $f = 10$ bar. Right: The volume fluctuation during the simulation.

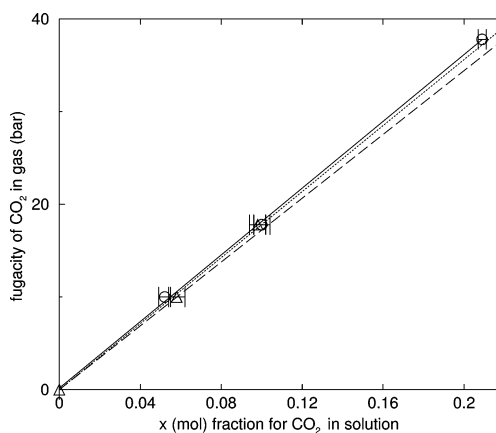


Figure 9. The fugacity of CO₂ in the gas phase versus the mole fraction of CO₂ in the solution phase at 323 K. The three circles are from CFC osmotic simulations at a total pressure of 40 bar. The triangles are from simulations at a total pressure of 20 bar. The solid line is a linear fit to the 40 bar simulation results and yields an estimated Henry's constant of 18.0 ± 0.3 MPa. A linear fit to the 20 bar results yields a Henry's constant of 18.1 ± 0.6 MPa. The dotted line represents a Henry's constant of 17.8 ± 0.6 MPa calculated by Errington et al.,⁵⁶ while the long dashed line represents a Henry's constant of 17.2 ± 0.4 MPa obtained by Zhang and Siepmann.⁵⁷

against the previous results of Errington and co-workers.⁵⁶ The pure ethanol density at these conditions was calculated as 0.751 ± 0.002 g/cm³. This is consistent with the value reported by Errington and co-workers of 0.7579 ± 0.0005 g/cm³. The small difference may be attributed to the fact that a flexible model for ethanol was used here, while Errington and co-workers used a rigid model. Also, the way in which the Lennard-Jones potential was truncated differs slightly between the two methods.

CFC osmotic simulations were then run at 323 K and varying CO₂ fugacities. The total pressure of the system was set to either 20 bar or 40 bar. An example of the convergence of the number of CO₂ solutes and system volume for a fugacity of 10 bar is shown in Figure 8. The isotherms calculated from these simulations are shown in Figure 9. Note that the solubilities calculated at total pressures of 20 and 40 bar agree within the uncertainty of the calculations,

confirming the insensitivity of the solubility to total pressure. By fitting a straight line through the data, a Henry's law constant may be obtained. For the CFC calculations the Henry's law constant was found to be 18.0 ± 0.3 MPa at 40 bar and 18.1 ± 0.6 MPa at 20 bar, respectively. This agrees very well with the results obtained from transition matrix MC calculations (17.8 ± 0.6 MPa) and Gibbs ensemble calculations (17.2 ± 0.4 MPa). The Henry's constants for these two previous calculations are shown as dotted and dashed lines in Figure 9. We note that the two previous studies^{56,57} have used low pressure and concentration of CO₂. The fact that the osmotic CFC calculations agree with these previous calculations confirms that the method yields correct results.

6. Conclusions

A new open system method called continuous fractional component Monte Carlo (CFC MC) has been presented. The method is designed to overcome difficulties with the insertion and deletion of molecules and relies upon gradual changes in a coupling parameter to increase or decrease the interaction strength between a fractional molecule and the rest of the molecules in the system. An adaptive bias potential is used to enable these transitions to occur efficiently. The bias potential is adjusted using the Wang–Landau updating scheme in such a way as to maintain a uniform distribution of fractional molecule states. In between these coupling strength moves, hybrid Monte Carlo steps are used to relax the system.

Acceptance rules were developed for the grand canonical and osmotic ensembles. The grand canonical CFC method was used to calculate the volumetric properties of the Lennard-Jones fluid and SPC water. The results agreed with accepted values and a standard implementation of grand canonical Monte Carlo. An implementation of another hybrid method³⁵ was also found to give the correct results for the Lennard-Jones fluid but failed for the SPC water case, due to an inability to make successful changes in the coupling parameter. Without the use of a bias potential, the system became “stuck” at intermediate states. It was not possible to achieve correct results for the Lennard-Jones fluid at high density with an implementation of a previously developed²⁹ grand canonical molecular dynamics procedure. The osmotic CFC method was used to compute the solubility of CO₂ in ethanol. Results from the calculations agree with two previously published simulation studies.

The CFC method has been verified by comparing calculated volumetric properties and solubilities against previous accepted results. In the future we wish to extend the CFC method to other ensembles and apply it to more challenging systems where other simulation approaches fail or become very inefficient.

Acknowledgment. Support for this work was provided by the Air Force Office of Scientific Research under contract number F49620-03-1-0212.

References

(1) Potoff, J. J.; Siepmann, J. I. *AIChE J.* **2001**, *47*, 1676.

- (2) Gao, G. T.; Wang, W. C. *Fluid Phase Equil.* **1997**, *130*, 157.
- (3) Martin, M. G. *Fluid Phase Equil.* **2006**, *248*, 50.
- (4) Crozier, P. S.; Rowley, R. L. *Fluid Phase Equil.* **2002**, *193*, 53.
- (5) Morrow, T. I.; Maginn, E. J. *J. Chem. Phys.* **2005**, *122*, 54504.
- (6) Anwar, J.; Frenkel, D.; Noro, M. G. *J. Chem. Phys.* **2003**, *118*, 728.
- (7) Agrawal, P. M.; Rice, B. M.; Thompson, D. L. *J. Chem. Phys.* **2003**, *119*, 9617.
- (8) Eike, D. M.; Maginn, E. J. *J. Chem. Phys.* **2006**, *124*, 164503.
- (9) Chen, B.; Siepmann, J. I.; Klein, M. K. *J. Phys. Chem. B* **2001**, *105*, 9840.
- (10) Zhao, X. S.; Chen, B.; Karaborni, S.; Siepmann, J. I. *J. Phys. Chem. B* **2005**, *109*, 5368.
- (11) Kofke, D. A.; Cummings, P. T. *Fluid Phase Equil.* **1998**, *150*, 41.
- (12) Kofke, D. A. *Fluid Phase Equil.* **2005**, *228*, 41.
- (13) Snurr, R. Q.; Bell, A. T.; Theodorou, D. N. *J. Phys. Chem.* **1993**, *97*, 13742.
- (14) Mezei, M. *Mol. Phys.* **1980**, *40*, 901.
- (15) Siepmann, J. I.; Frenkel, D. *Mol. Phys.* **1992**, *75*, 59–70.
- (16) Mooij, G. C. A. M.; Frenkel, D.; Smit, B. *J. Phys.: Condens. Matter* **1992**, *4*, L255–L259.
- (17) Laso, M.; de Pablo, J. J.; Suter, U. W. *J. Chem. Phys.* **1992**, *97*, 2817–2819.
- (18) Berg, B. A.; Neuhaus, T. *Phys. Rev. Lett.* **1992**, *68*, 9.
- (19) Orkoulas, G.; Panagiotopoulos, A. Z. *J. Chem. Phys.* **1999**, *110*, 1581.
- (20) Yan, Q. L.; de Pablo, J. J. *J. Chem. Phys.* **1999**, *111*, 9509.
- (21) Fenwick, M. K.; Escobedo, F. A. *J. Chem. Phys.* **2003**, *119*, 11998.
- (22) De Pablo, J. J.; Prausnitz, J. M. *Fluid Phase Equil.* **1989**, *53*, 177.
- (23) Martin, M. G.; Siepmann, J. I. *J. Am. Chem. Soc.* **1997**, *119*, 8921.
- (24) Chen, B.; Siepmann, J. I.; Klein, M. L. *J. Am. Chem. Soc.* **2002**, *124*, 12232.
- (25) Wick, C. D.; Siepmann, J. I.; Theodorou, D. N. *J. Am. Chem. Soc.* **2005**, *127*, 12338.
- (26) Çağın, T.; Pettitt, B. M. *Mol. Phys.* **1991**, *72*, 169–175.
- (27) Çağın, T.; Pettitt, B. M. *Mol. Simul.* **1991**, *6*, 5.
- (28) Ji, J.; Çağın, T.; Pettitt, B. M. *J. Chem. Phys.* **1992**, *96*, 1333–1342.
- (29) Lo, C. M.; Palmer, B. *J. Chem. Phys.* **1995**, *102*, 925–931.
- (30) Shroll, R. M.; Smith, D. E. *J. Chem. Phys.* **1999**, *111*, 9025–9033.
- (31) Lísal, M.; Smith, W. R.; Kolafa, J. *J. Phys. Chem. B* **2005**, *109*, 12956–12965.
- (32) Theodorou, D. N. *J. Chem. Phys.* **2006**, *124*, 34109.
- (33) Uhlherr, A.; Theodorou, D. N. *J. Chem. Phys.* **2006**, *125*, 84107.

- (34) Escobedo, F. A.; de Pablo, J. J. *J. Chem. Phys.* **1996**, *105*, 4391–4394.
- (35) Boinepalli, S.; Attard, P. *J. Chem. Phys.* **2003**, *119*, 12769–12775.
- (36) Lynch, G. C.; Pettitt, B. M. *J. Chem. Phys.* **1997**, *107*, 8594–8610.
- (37) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. *J. Chem. Phys.* **1992**, *96*, 1776–17837.
- (38) Rosso, L.; Minary, P.; Zhu, Z.; Tuckerman, M. E. *J. Chem. Phys.* **2002**, *116*, 4389.
- (39) Wang, F.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050–2053.
- (40) Macedonia, M. D.; Maginn, E. J. *Mol. Phys.* **1999**, *96*, 1375–1390.
- (41) June, R. L.; Bell, A. T.; Theodorou, D. N. *J. Phys. Chem.* **1990**, *94*, 8232.
- (42) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon: Oxford, 1987.
- (43) Frenkel, D.; Smit, B. *Understanding Molecular Simulation*; Academic Press: San Diego, 2002.
- (44) Duane, S.; Kennedy, A. D.; Pendleton, B. J.; Roweth, D. *Phys. Lett. B* **1987**, *195*, 216–222.
- (45) Errington, J. R.; Panagiotopoulos, A. Z. *J. Chem. Phys.* **1999**, *111*, 9731.
- (46) Spyriouni, T.; Economou, I. G.; Theodorou, D. N. *Phys. Rev. Lett.* **1998**, *80*, 4466–4469.
- (47) Banaszak, B. J.; Faller, R.; de Pablo, J. J. *J. Chem. Phys.* **2004**, *120*, 11304.
- (48) Zervopoulou, E.; Mavrantzas, V. G.; Theodorou, D. N. *J. Chem. Phys.* **2001**, *115*, 2860.
- (49) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529–539.
- (50) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- (51) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Dordrecht: The Netherlands, 1981.
- (52) Mezei, M. *Mol. Phys.* **1987**, *61*, 565–582.
- (53) Mezei, M. *Mol. Phys.* **1989**, *67*, 1207–1208.
- (54) Smith, G. R.; Bruce, A. D. *J. Phys. A* **1995**, *28*, 6623–6643.
- (55) Industrial Fluid Properties Simulation Challenge. <http://fluidproperties.org> (accessed January 4, 2007).
- (56) Cichowski, E. C.; Schmidt, T. R.; Errington, J. R. *Fluid Phase Equil.* **2005**, *236*, 58.
- (57) Zhang, L.; Siepmann, J. I. *Theor. Chem. Acc.* **2006**, *115*, 391.
- (58) Johnson, J. K.; Zollweg, J. A.; Gubbins, K. E. *Mol. Phys.* **1993**, *78*, 591–618.

CT7000039

AMBER Force Field Parameters for the Naturally Occurring Modified Nucleosides in RNA

Raviprasad Aduri, Brian T. Psciuk, Pirro Saro, Hariprakash Taniga,
H. Bernhard Schlegel, and John SantaLucia, Jr.*

Department of Chemistry, Wayne State University, Detroit, Michigan 48202

Received November 7, 2006

Abstract: Classical molecular dynamics (MD) simulations are useful for characterizing the structure and dynamics of biological macromolecules, ultimately, resulting in elucidation of biological function. The AMBER force field is widely used and has well-defined bond length, bond angle, partial charge, and van der Waals parameters for all the common amino acids and nucleotides, but it lacks parameters for many of the modifications found in nucleic acids and proteins. Presently there are 107 known naturally occurring modifications that play important roles in RNA stability, folding, and other functions. Modified nucleotides are found in almost all transfer RNAs, ribosomal RNAs of both the small and large subunits, and in many other functional RNAs. We developed force field parameters for the 107 modified nucleotides currently known to be present in RNA. The methodology used for deriving the modified nucleotide parameters is consistent with the methods used to develop the Cornell et al. force field. These parameters will improve the functionality of AMBER so that simulations can now be readily performed on diverse RNAs having post-transcriptional modifications.

1. Introduction

Ribonucleic acids (RNA) play important roles in diverse biological functions including protein synthesis, gene silencing, and in the regulation of gene expression.^{1–3} RNA is initially synthesized as a phosphodiester polymer of four nucleosides namely adenosine, guanosine, cytidine, and uridine, which are called the “common” nucleosides. In addition to the four common nucleosides, there are many modified nucleosides found in RNA.⁴ These nucleoside modifications are formed post-transcriptionally. Presently there are at least 107 modifications that have been discovered in natural RNA.^{5–8} Modified nucleosides are found in almost all tRNAs, ribosomal RNAs of both the small and large subunits of the ribosome, mRNAs, snoRNA, and other functionally important RNA molecules.⁵ Currently, the biological functions of most modifications are unknown, though some roles are beginning to be elucidated.^{9–11} The most commonly occurring modification is pseudouridine, in which the C5 of uracil is covalently attached to the sugar

C1', resulting in a C–C glycosidic bond instead of the usual C–N glycosidic bond.¹² The next most common modification found in RNA is the methylation of the 2'-O position of the ribose sugar. The lifetimes of base pairs involving certain modified nucleosides are reported to be longer than the typical Watson–Crick base pairs, making these modifications essential for the viability of extremophiles.⁵ Owing to the ubiquitous presence of the modified nucleosides in RNA, it is essential to develop accurate and reliable force field parameters for these modifications that enable the simulation of molecular dynamics of RNA with or without modifications.¹³ Stable MD simulations require uniformity in the force field parameter sets for modified nucleosides to be consistent with the present force field for the common nucleosides.

Molecular mechanics (MM) and molecular dynamics (MD) are useful for revealing dynamics and structure of biomacromolecules thereby elucidating biological function. There are several MM force fields available for performing simulations of biomolecules including CHARMM,¹⁴ AMBER,¹⁵ XPLOR,¹⁶ and others.¹⁷ Armed with an increasing amount of computational resources, researchers have successfully incorporated more accuracy and elegance to force

* Corresponding author phone: (313)577-0101; fax: (313)577-8822; e-mail: jsl@chem.wayne.edu.

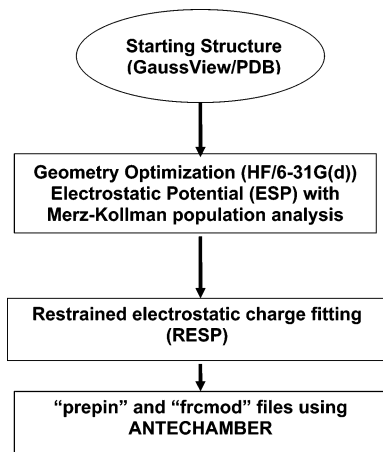


Figure 1. Flowchart of the protocol used in generating the parameters for modified nucleotides.

fields including polarizable functions,¹⁸ lone pairs, coupled stretching and bending modes, and sophisticated models of solvation and electrostatics.¹⁹ AMBER is one of the most widely used force fields in the simulation of biological molecules possessing the necessary parameters for the common nucleosides and amino acids. Recently, AMBER force field parameters were developed for phosphorothioate nucleic acids²⁰ as well as for various polyphosphates.²¹ Presently, force field parameters are available for modifications found in tRNA^{Phc} (http://pharmacy.man.ac.uk/amber/nuc/tRNA_inf.html) and some of the 2' sugar modifications.²² Some groups have reported parameters for a few modifications present in the anticodon stem loop of the tRNA.^{23–25} However, parameters for the other naturally occurring modified nucleosides are not contained within the AMBER suite.

An expanding knowledge surrounding the role of RNA in various biological processes and the presence of a large variety of modified nucleosides provide an important demand for the development of force field parameters for modified nucleosides suitable for use with the well-established AMBER force field. Herein, we report the development of force field parameters for the known 107 modified nucleosides found in natural RNA.⁴ The modified RNA force field parameters have been developed to be consistent with the Cornell et al. force field²⁶ of AMBER.

2. Methods

2.1. Parametrization Strategy. The strategic approach used for developing AMBER force field parameters for the 107 modifications in RNA is summarized in Figure 1. The parametrization protocol developed by Cornell et al.²⁶ was followed to be consistent with the AMBER force field. Atom-centered partial charges were calculated using the RESP methodology. The electronic structure calculations were carried out at the Hartree–Fock level of theory using the 6-31G(d) basis set despite improvements in computing resources that would have enabled us to perform calculations at higher levels of theory. In this way, the calculations in this work are consistent with the procedure followed in the original development of the AMBER Cornell et al. force field.²⁶ To obtain the charge constraint for the sugar moiety,

QM calculations were performed on the four common nucleosides, A, C, G, and U with both C3'endo and C2'endo ribose sugars. In both these cases, the sugar atoms among all four nucleosides were equivalenced. The phosphate group and O3' and O5' charges were obtained using dimethyl phosphate (DMP) as the model system as shown in Figure 2. RESP charge fitting was done with all the four nucleosides with either C2'endo or C3'endo sugar. C2'endo and C3'endo nucleosides were also fit together during the RESP procedure (data not shown). The modifications may play a role altering the sugar pucker, but the sugar pucker preferences for the modified nucleosides are not well understood.^{27,28} Since RNA predominantly contains a C3'endo ribose sugar conformation, and it is the conformation of the sugar used in the initial development of AMBER parameters, we decided to use the charge obtained from the RESP fitting of only the C3'endo sugar containing nucleosides. The ribose sugar charge was calculated by multiequivalencing the four natural nucleosides A, G, C, and U with C3'endo sugar conformation as described in Cieplak et al.²⁹ The charges obtained for the common nucleosides in the C3'endo conformation are given in Table 1. The charges obtained for C2'endo ribose, C3'endo ribose sugar, and 2'-O-methyl ribose are given in Table 2. The ribose sugar charges are relatively insensitive to sugar pucker conformations. The standard deviation of charges for comparison of C3'endo vs C2'endo riboses is 0.0269e which is less than the systematic error of the RESP methodology itself, and thus there is no need for separate parametrization of C3' and C2'endo sugar puckers.

2.2. Ab Initio Calculations. AMBER force field parameters were developed by performing ab initio calculations at the Hartree–Fock level of theory using the 6-31G(d) basis set using the GAUSSIAN03³⁰ suite of programs. To test that our calculations followed the Restrained ElectroStatic Potential (RESP) charge fitting methodology³¹ procedure outlined in the Cieplak et al.,²⁹ we performed computations on the four commonly occurring nucleosides A, C, G, and U. The charges reported by Cieplak et al. are in excellent agreement (with a standard deviation of 0.0362e, see Table 4) with those determined here, thereby validating our approach. The modular nature of the RESP as well as of the structure of RNA itself allowed us to split the nucleosides into separate base, sugar, and phosphate moieties resulting in the reduction of the computational burden. To account for the phosphate charge, dimethyl phosphate (DMP) was used as the model system. Nucleosides with modifications in the base moiety were modeled by replacing the sugar with a methyl group. Conversely, nucleosides with modifications in the sugar moiety were modeled by replacing the base with a methyl group (Figure 2). The RESP procedure developed by Kollman and colleagues allows a modular approach to recombine sugar and base moieties by “equivalencing”.³¹ This strategy not only reduces the number of atoms in each ab initio computation but also allows portability of parameters so that different bases and sugars can be appropriately constructed. For example, once computations for 2'-O-methyl ribose are complete, the results can be combined with a variety of bases. Conversely, a modified base can be recombined with differing sugars (e.g., ribose, deoxyribose,

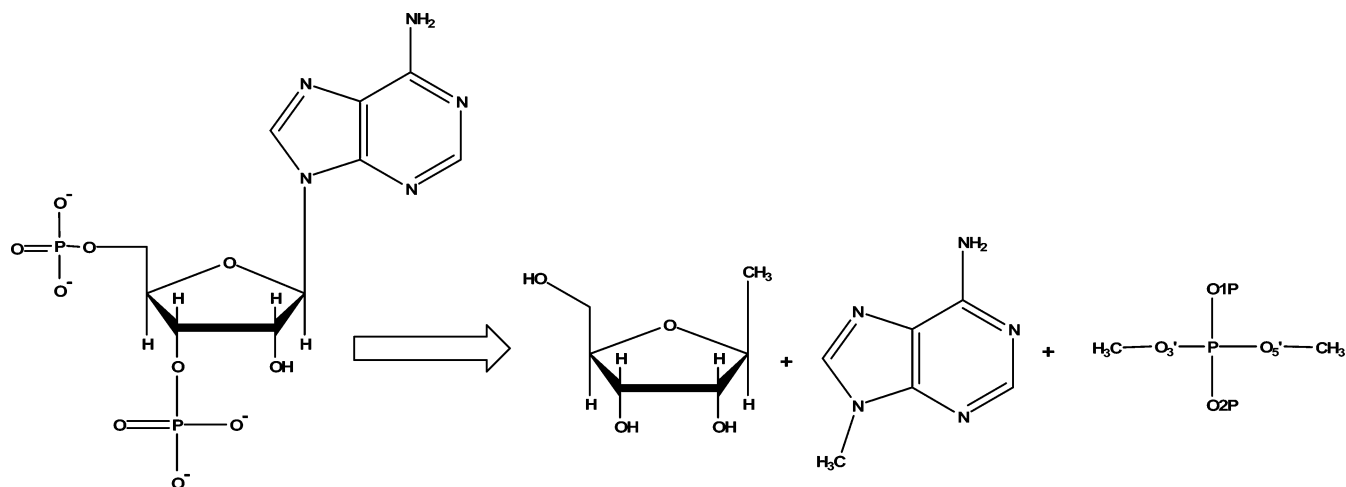


Figure 2. The charge fitting method used to generate the charges for the common nucleosides A, G, C, and U, using the modular nature of RNA to reduce the computational time. The charges for O1P, O2P, O3', O5', and P were obtained by using dimethyl phosphate (DMP) as a model system. See text for explanation.

Table 1. Charge Values Obtained in this Work for the Common Nucleosides A, G, C, and U

adenosine		guanosine		cytidine		uridine	
N9	0.0172	N9	0.0268	N1	-0.2152	N1	0.1110
C8	0.1299	C8	0.1066	C2	0.8867	C2	0.4539
N7	-0.5850	N7	-0.5575	O2	-0.6560	O2	-0.5407
C6	0.7111	C6	0.5316	N3	-0.8128	N3	-0.3681
N6	-0.9386	O6	-0.5483	C4	0.9020	C4	0.6022
C5	0.0586	C5	0.1513	N4	-0.9919	O4	-0.5652
C4	0.3050	C4	0.1563	C5	-0.5972	C5	-0.3135
N3	-0.6835	N3	-0.5959	C6	0.1262	C6	-0.2320
C2	0.5741	C2	0.7191	H5	0.2023	H5	0.1697
N1	-0.7536	N2	-0.9044	H6	0.1875	H6	0.2557
H8	0.1749	N1	-0.5287	NH1	0.4251	N3H	0.3087
H2	0.0467	H8	0.1767	NH2	0.4251		
HN1	0.4125	N2H1	0.3968				
HN2	0.4125	N2H2	0.3968				
		N1H	0.3546				

2'-O-methyl ribose, etc.). Anticipating the future discovery of new modified RNA, this strategy will allow for many nucleosides to be modeled that have not yet been found in nature or artificially synthesized. For example, deoxy pseudouridine is not found in nature, but it could be constructed from the parameters presented here for the pseudouridine along with the deoxyribose sugar parameters. The physiological pH of 7.0 was used in deciding the protonation states of all functional groups. Other protonation states observed at different pH were not considered in this study.³² In addition, only the lowest energy tautomeric state was considered. The starting geometries for ab initio calculations were obtained from the PDB database, when available. When a suitable crystal structure could not be retrieved, the structure was generated using GaussView and GAUSSIAN03. Hydrogens were added to the PDB structures using an automated feature in GaussView. Each nucleoside was manually inspected to ensure the proper valence of each heavy atom. The generic names, three-letter codes, starting

Table 2. Charge Values Obtained for the Three Common Sugars in RNA, C3'-Endo, C2'-Endo, and 2'-O-Methyl Ribose Sugars

atom name	C3'-endo	C2'-endo	2'O methyl ribose
P	1.0878	1.0825	1.0878
O1P	-0.7667	-0.7655	-0.7667
O2P	-0.7667	-0.7655	-0.7667
O5'	-0.4713	-0.5036	-0.4725
C5'	0.0635	0.0292	0.1289
C4'	0.0386	0.0625	0.1522
O4'	-0.3272	-0.3851	-0.4652
C3'	0.2125	0.2165	0.0675
O3'	-0.4890	-0.4649	-0.4878
C2'	0.0775	0.1064	0.0405
O2'	-0.5913	-0.6198	-0.3277
C1'	0.0460	0.1096	0.3686
H5'	0.0689	0.0823	0.0426
H5''	0.0689	0.0823	0.0426
H4'	0.1168	0.1215	0.0394
H3'	0.0825	0.0858	0.1460
H2'	0.0929	0.0659	0.0904
H1'	0.1643	0.1462	0.0417
OH2'	0.4101	0.4210	na ^a
CM2	na ^a	na ^a	-0.0385
HM'1	na ^a	na ^a	0.0651
HM'2	na ^a	na ^a	0.0651
HM'3	na ^a	na ^a	0.0651

^a na – not applicable.

geometries, and, where available, the RNA in which they occur for all modified nucleotides are summarized in Table 3.

2.3. Electrostatic Potential Calculations. After geometry optimization, the electrostatic surface potential (ESP) was fit using the electrostatic charge computing method developed by Merz and Kollman,³³ which uses a Connolly surface algorithm to calculate a number of shells with radii of 1.4, 1.6, 1.8, and 2.0 times the van der Waals radius of the constituent atoms in the molecule. A Levenberg–Marquardt nonlinear optimization procedure was then used to compute the set of atom-centered point charges that best reproduce

Table 3. Generic Names, Three-Letter Codes, Source of Starting Geometry, and the Occurrence of Different Modifications^a

generic name	three-letter code ^b	alternate codes ^c	source ^d	occurrence
1-methyladenosine	1MA		(1EHZ)	tRNA
2-methylthio- <i>N</i> ⁶ -hydroxynorvalyl carbamoyladenosine	26A			tRNA
2-methyladenosine	2MA		1EFW	tRNA
2'- <i>O</i> -ribosylphosphate adenosine	2RA		1YFZ	tRNA
<i>N</i> ⁶ -methyl- <i>N</i> ⁶ -threonylcarbamoyladenosine	66A			tRNA
<i>N</i> ⁶ -acetyladenosine	6AA			tRNA
<i>N</i> ⁶ -glycinylicarbamoyladenosine	6GA			tRNA
<i>N</i> ⁶ -isopentenyladenosine	6IA			tRNA
<i>N</i> ⁶ -methyladenosine	6MA			tRNA
<i>N</i> ⁶ -threonylcarbamoyladenosine	6TA			tRNA
<i>N</i> ⁶ , <i>N</i> ⁶ -dimethyladenosine	DMA	M2A		16S rRNA
<i>N</i> ⁶ -(<i>cis</i> -hydroxyisopentenyl)adenosine	HIA			tRNA
<i>N</i> ⁶ -hydroxynorvalylcarbamoyladenosine	HNA			tRNA
1,2'- <i>O</i> -dimethyladenosine	M2A			tRNA
<i>N</i> ⁶ ,2'- <i>O</i> -dimethyladenosine	MMA			
2'- <i>O</i> -methyladenosine	MRA	A2M		tRNA
<i>N</i> ⁶ , <i>N</i> ⁶ , <i>O</i> -2'-trimethyladenosine	MTA			
2-methylthio- <i>N</i> ⁶ -(<i>cis</i> -hydroxyisopentenyl) adenosine	SIA			tRNA
2-methylthio- <i>N</i> ⁶ -methyladenosine	SMA			tRNA
2-methylthio- <i>N</i> ⁶ -isopentenyladenosine	SPA	MIA	1B23	tRNA
2-methylthio- <i>N</i> ⁶ -threonyl carbamoyladenosine	STA	12A	1FIR	tRNA
2-thiocytidine	2SC			tRNA
3-methylcytidine	3MC		3MCT	tRNA
<i>N</i> ⁴ -acetylcytidine	4AC			tRNA,rRNA
<i>N</i> ⁴ -methylcytidine	4MC			
5-formylcytidine	5FC			tRNA
5-methylcytidine	5MC		1EHZ	tRNA,16S rRNA
5-hydroxymethylcytidine	HMC			
lysidine	K2C			tRNA
<i>N</i> ⁴ -acetyl-2'- <i>O</i> -methylcytidine	MAC			tRNA,rRNA
5-formyl-2'- <i>O</i> -methylcytidine	MFC			tRNA
5,2'- <i>O</i> -dimethylcytidine	MMC			tRNA
2'- <i>O</i> -methylcytidine	MRC	OMC	1EHZ	tRNA
<i>N</i> ⁴ ,2'- <i>O</i> -dimethylcytidine	M4C			rRNA
<i>N</i> ⁴ , <i>N</i> ⁴ ,2'- <i>O</i> -trimethylcytidine	MTC			rRNA
1-methylguanosine	1MG		2ASY	tRNA
<i>N</i> ² ,7-dimethylguanosine	27G			
<i>N</i> ² -methylguanosine	2MG		1EHZ	tRNA,rRNA
2'- <i>O</i> -ribosylphosphate guanosine	2RG			tRNA
7-methylguanosine	7MG	G7M	1EHZ	tRNA,rRNA
under modified hydroxywybutosine	BUG	UBG		tRNA
7-aminomethyl-7-deazaguanosine	DAG		1EFZ	tRNA
7-cyano-7-deazaguanosine	DCG			tRNA
<i>N</i> ² , <i>N</i> ² -dimethylguanosine	DMG	M2G	(1EHZ)	tRNA
4-demethylwyosine	DWG			tRNA
epoxyqueuosine	EQG			tRNA
hydroxywybutosine	HWG			tRNA
isowyosine	IWG			tRNA
<i>N</i> ² ,7,2'- <i>O</i> -trimethylguanosine	M7G			tRNA
<i>N</i> ² ,2'- <i>O</i> -dimethylguanosine	MMG			tRNA
1,2'- <i>O</i> -dimethylguanosine	M1G			tRNA
2'- <i>O</i> -methylguanosine	MRG	OMG	1EHZ	tRNA
<i>N</i> ² , <i>N</i> ² ,2'- <i>O</i> -trimethylguanosine	MTG			tRNA
<i>N</i> ² , <i>N</i> ² ,7-trimethylguanosine	N2G			
peroxywybutosine	PBG			tRNA
galactosyl-queuosine	QGG			tRNA
mannosyl-queuosine	QMG			tRNA
queuosine	QUG	QUO		tRNA

Table 3. (Continued)

generic name	three-letter code ^b	alternate codes ^c	source ^d	occurrence
archaeosine	RCG			tRNA
wybutosine	WBG	YG	1EHZ	tRNA
methylwyosine	WMG			tRNA
wyosine	WYG			tRNA
2-thiouridine	2SU	SUR		tRNA
3-(3-amino-3-carboxypropyl)uridine	3AU			tRNA
3-methyluridine	3MU			rRNA
4-thiouridine	4SU	S4U	1B23	tRNA
5-methyl-2-thiouridine	52U			tRNA
5-methylaminomethyluridine	5AU			tRNA
5-carboxymethyluridine	5CU			
5-carboxymethylaminomethyluridine	5DU			tRNA
5-hydroxyuridine	5HU			tRNA
5-methyluridine	5MU		1EHZ	tRNA
5-taurinomethyluridine	5TU			tRNA
5-carbamoylmethyluridine	BCU			tRNA
5-(carboxyhydroxymethyl)uridine methyl ester	CMU			tRNA
dihydrouridine	DHU	H2U	1EHZ	tRNA
5-methyldihydrouridine	DMU			
5-methylaminomethyl-2-thiouridine	ESU			tRNA
5-(carboxyhydroxymethyl)uridine	HCU			tRNA
5-(isopentenylaminomethyl)uridine	IAU			tRNA
5-(isopentenylaminomethyl)-2-thiouridine	ISU			tRNA
3,2'- <i>O</i> -dimethyluridine	M3U			
5-carboxymethylaminomethyl-2'- <i>O</i> -methyluridine	MAU			tRNA
5-carbamoylmethyl-2'- <i>O</i> -methyluridine	MCU			tRNA
5-methoxycarbonylmethyl-2'- <i>O</i> -methyluridine	MEU			tRNA
5-(isopentenylaminomethyl)-2'- <i>O</i> -methyluridine	MIU			tRNA
5,2'- <i>O</i> -dimethyluridine	MMU	2MU	1FIR	tRNA
2'- <i>O</i> -methyluridine	MRU			tRNA
2-thio-2'- <i>O</i> -methyluridine	MSU			tRNA
uridine 5-oxyacetic acid	OAU			tRNA
5-methoxycarbonylmethyluridine	OCU			tRNA
uridine 5-oxyacetic acid methyl ester	OEU			tRNA
5-methoxyuridine	OMU			tRNA
5-aminomethyl-2-thiouridine	SAU			tRNA
5-carboxymethylaminomethyl-2-thiouridine	SCU			tRNA
5-methylaminomethyl-2-selenouridine	SEU			tRNA
5-methoxycarbonylmethyl-2-thiouridine	SMU			tRNA
5-taurinomethyl-2-thiouridine	STU			tRNA
pseudouridine	PSU		1EHZ	tRNA,rRNA
1-methyl-3-(3-amino-3-carboxypropyl)pseudouridine	13P			28S rRNA
1-methylpseudouridine	1MP			tRNA
3-methylpseudouridine	3MP			23S rRNA
2'- <i>O</i> -methylpseudouridine	MRP			tRNA
inosine	INO			tRNA
1-methylinosine	1MI			tRNA
1,2'- <i>O</i> -dimethylinosine	MMI			tRNA
2'- <i>O</i> -methylinosine	MRI			

^a The PDB reference is given for nucleosides where available. GaussView was used for generating the starting geometry wherever the PDB source is not mentioned. ^b Three-letter code proposed in this study. ^c Alternate three-letter codes used previously. ^d Source refers to where we obtained the coordinates for starting geometries of modified nucleosides. Values in parenthesis indicate that the modification occurs in that PDB file, but it was not used in this work. If no PDB source is given or if in parentheses, then the starting geometry was generated using GaussView.

the surface charges that were derived quantum mechanically.³⁴ Because of differences in convergence criteria, the optimized geometry of the molecule may also differ slightly based on the QM program used, which would alter charge values. The grid size (i.e., the number of shells of points

and the density of points on the shells) used to compute the electrostatic potential slightly influences the atom centered point charges in the ESP calculation. It is well-known that the atomic charges derived from using a grid of electrostatic potentials computed by quantum mechanical calculations

Table 4. Comparison of Adenosine Charges Computed in This Work with the Charges Available in PARM99 of AMBER

atom name	adenosine whole nucleoside	adenosine PARM 99	adenosine modular fit
N9	0.0172	-0.0251	-0.0503
C8	0.1299	0.2006	0.1060
N7	-0.5850	-0.6073	-0.5725
C6	0.7111	0.7009	0.6394
N6	-0.9386	-0.9019	-0.8963
C5	0.0586	0.0515	0.0553
C4	0.3050	0.3053	0.4499
N3	-0.6835	-0.7615	-0.7282
C2	0.5741	0.5875	0.5587
N1	-0.7536	-0.6997	-0.7354
H8	0.1749	0.1553	0.1734
H2	0.0467	0.0473	0.0579
NH1	0.4125	0.4115	0.4122
NH2	0.4125	0.4115	0.4122
SD from PARM99	0.0362	N/A	0.0529
SD from whole nucleoside	N/A	0.0362	0.0506

depend slightly on the rotational orientation of the molecule.³⁵ This effect is due to the finite grid used for the sampling of the electrostatic potential (every 1 Å² in this study) in the ESP calculation of point charges. To overcome these charge differences due to geometrical orientation of the molecule, multiorientation charge fitting can be utilized.³⁶ This allows for sampling of many orientations of the molecule, which reduces the round-off errors in atom charges. To test the effect of multiorientation on the charge fitting, we used pseudouridine (PSU) as a model system. The R.E.D. II³⁶ code provides a good platform for fitting the charges by using a rigid-body reorientation algorithm to make multiple orientations of the molecule. The R.E.D. code allows for the random selection of three different heavy atoms, which are used to orient the molecule. Due to the small standard deviation (~0.016e, see the Supporting Information) in charge values due to the orientation effect and the laborious computations and file manipulations required to implement multiorientation on the 107 modifications we decided not to perform R.E.D. on the modified nucleosides (see below for discussion).

2.4. Restrained Electrostatic Potential Charges. RESP charge fitting was carried out as described by Cieplak et al.²⁹ The modular nature of nucleotides allowed for restraining the charge of a methyl group to replace either sugar or base moiety during the ab initio calculation. In the case of base modifications, the total methyl group charge was restrained to the total charge of the sugar (0.118186e) obtained from the common nucleoside calculations during the first stage of RESP fit. When fitting the 2'-O-methyl ribose sugar to acquire the charges for this modified sugar, the methyl group replacing the base was restrained to an equivalent and opposite charge value obtained for the normal sugar (i.e., -0.118186e). All equivalent and polar hydrogens, such as hydrogens in an amino group, were equivalenced during the first stage of the RESP fit; whereas,

the nonpolar equivalent hydrogens, as in the case of methyl group and H5' and H5'' of the sugar, were equivalenced in the second stage of the fit. We used ANTECHAMBER Ver 1.24 module of AMBER to do the RESP charge fitting.³⁷

2.5. Generating the Parameters. Figure 3 shows a schematic diagram of the protocol followed for generating the charges for the modified nucleosides. A common problem in developing force field parameters is that output text files from one program are incompatible with the input format required for the program used in the next step. For a single RESP computation on a modified nucleotide one could perform such file manipulations manually. For this project, however, performing such manual manipulations on 107 nucleosides is impractical. Thus, we developed several automated text format conversion programs to accomplish this task. Since the ab initio calculations were carried out using the modular approach, we were unable to use the NEWZMAT module of GAUSSIAN to convert the three check point files into a single PDB file for a complete nucleoside. The program gif2pdb.exe was written to convert GAUSSIAN job files (gjf) into PDB format. This program and others are available on our group home page (<http://ozone3.chem.wayne.edu>). The nucleoside coordinates were generated by combining the optimized geometry of the modified base with the C3'-endo sugar in GaussView. These Gaussian files were then converted into a single PDB file using gif2pdb.exe. The resulting PDB files were then used to generate the "ANTECHAMBER" format files using ANTECHAMBER Ver 1.24. Once the ANTECHAMBER files were generated, the charges obtained from the RESP fit were input into the ANTECHAMBER files accordingly. To reduce the development of new atom types, we used the Generalized Amber Force Field (GAFF)³⁸ to assign the atom types for the modified nucleosides. GAFF contains atom types for all atoms present in the modified nucleosides studied except selenium. In the selenium case, we temporarily decided to assign atom type "SS" to selenium, since the chemical nature of selenium closely resembles sulfur. SS originally represented a thione functional group which is similar in character to the C=Se group found in the modified base 5-methylaminomethyl-2-selenouridine (SEU). The bond lengths, bond angles, and dihedral values used for selenium were similar to atom type "SS". We are in the process of determining the force constants, equilibrium distances, and equilibrium angles for selenium. Once these parameters are available, there may be a need to introduce a new atom type for selenium in GAFF. Once the atom types were assigned, the preparatory file "prepin" and force field file "frcmod" were generated using ANTECHAMBER V.1.24.

3. Naming Convention

We were unable to find a literature consensus in the naming convention used for the modified nucleosides found in RNA. For example, 5,6-dihydrouridine can be found as H2U³⁹ or DHU.⁴⁰ Consequently, we were compelled to develop a consistent three-letter code indicating the nature of the modification as clearly as possible without conflicting with amino acid or other names. In this naming

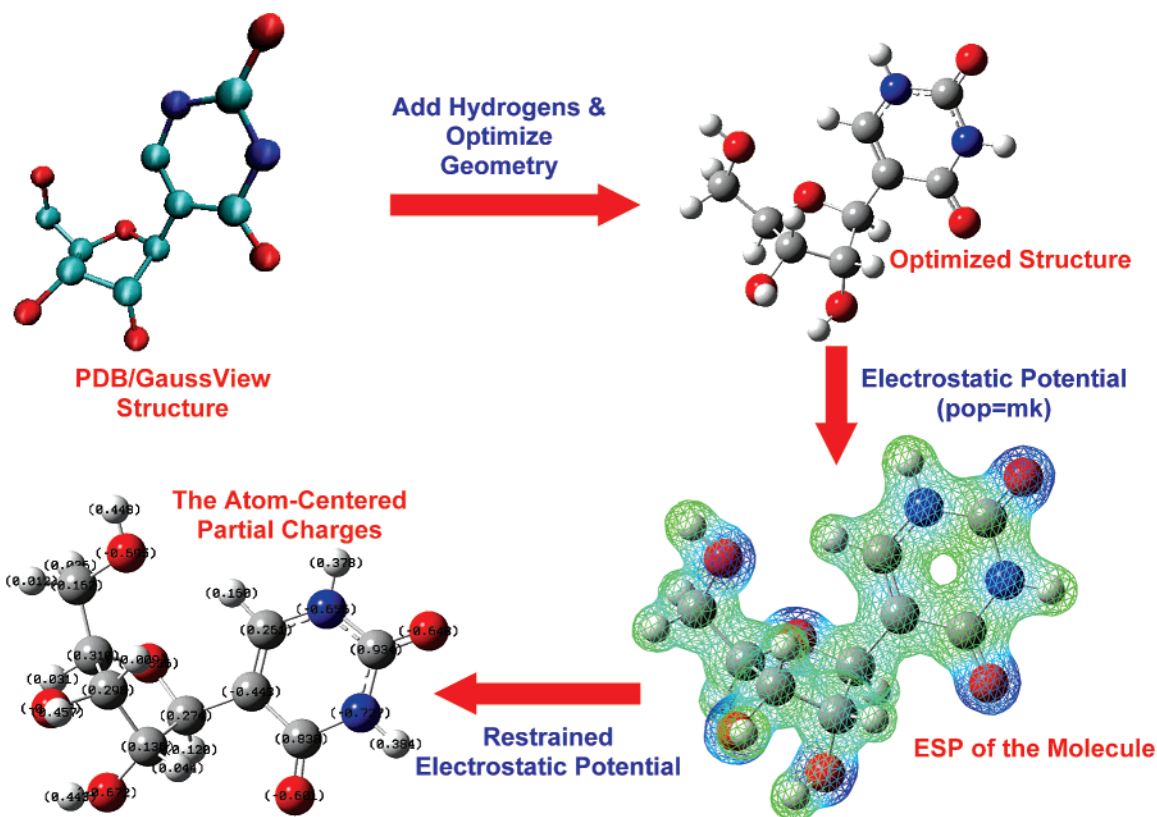


Figure 3. Protocol for the determination of atom-centered partial charges. The starting structures were obtained either from a PDB file or created using GaussView. Hydrogen atoms were added using GaussView. Geometry optimization was done using Gaussian03. The electrostatic potential was computed using Merz–Kollman population analysis, and charges were produced by fitting the ESP using RESP as explained in the text.

convention, the last letter signifies the closest common nucleoside associated with the modification (i.e., the transcribed base encoded in the genomic DNA). For example, wybutosine is named WBG and not Y base, which would conflict with the IUPAC nomenclature for a pyrimidine.⁴¹ Other examples are shown in Table 3. The pseudouridine modification uses “P” as the last letter, and modifications involving inosine were given the letter “I”. Additionally, we verified that the three-letter codes used for modified nucleotides did not interfere with any of the letter codes that were already used in AMBER. In the present naming convention, the nature of modification is explicitly used to form the three-letter code when ever possible. For example, 1MA stands for 1-methyladenosine, whereas, 5FC is the code for 5-formylcytidine, and MRX was used to indicate the presence of a 2'-O-methyl group on the ribose sugar (e.g. MRA, MRP). We also avoided using A, C, G, and U as the starting letter to escape confusion with the one-letter codes that are still used for the common nucleosides, particularly for sequence alignment algorithms. Thus, the presence of a character other than A, C, G, or U indicates that the three characters in a sequence denote a single modified nucleotide. The generic names along with their three-letter codes for all the 107 modifications are given in Table 3. We hope our naming convention will be widely adopted by the community.

4. Web Site for AMBER Parameters for Modified Nucleosides

Optimized geometries, electrostatic potentials, RESP input and output files, and format conversion executables are available on our Web site <http://ozone3.chem.wayne.edu>. The Web site also contains the “prepin” and “frcmod” files and the protocols needed to implement the modified nucleoside parameters into AMBER. The optimized geometries of the modifications allow for the opportunity to reproduce the charges obtained in the present study. The modifications are classified according to their closest common nucleotide. For example, 1-methyladenine will be found under the “adenosine modifications” section. The Web site allows users to download parameters for one modification at a time or download parameters for all 107 modifications at once in a compressed file. The Web site also includes other information regarding each particular modification and links to the McCloskey group “RNA Modification database” Web site (<http://library.med.utah.edu/RNAMods>).⁴ Apart from the force field parameters for the modified nucleosides, the Web site also contains the monomer optimized geometries for all the 107 modification. The modified nucleoside parameters have also been made available on the AMBER contributed parameters Web site (<http://pharmacy.man.ac.uk/amber>).

5. Results and Discussion

The functional form of the AMBER force field is given in eq 1

$$E = \sum_{\text{bonds}} K_r(r - r_0)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi)] + \sum_{\text{nonbonded}} 4\epsilon_{ij} \left(\left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \left[\frac{q_i q_j}{\epsilon R_{ij}} \right] \right) \quad (1)$$

The total internal energy of a molecule is decomposed into energy components representing bond stretching, angle bending, torsional angle twisting, Lennard-Jones potential, and nonbonded coulomb electrostatic terms. The present study is focused on developing the atom-centered partial charges necessary to compute the electrostatic term in eq 1 for the 107 naturally occurring modified nucleosides found in RNA. The force constants and equilibrium distances, bond angles, and dihedral terms were generated using the GAFF.³⁸ As these modifications may not occur at the 5' or 3' termini of RNA, we did not develop the parameters for the 5' or 3' terminal modifications. We validated the parameters by conducting molecular dynamics simulations on tRNA^{Phe}, which contains 14 modified bases. Figure 4 depicts example chemical structures of some of the modified nucleosides for which AMBER parameters were developed.

Charges obtained for the common nucleosides, A, C, G, and U in C3'-endo conformation are shown in Table 1. These charges are in good agreement with the AMBER force field parameters in PARM99 of AMBER as shown in Table 4. Similar agreement was observed for cytidine, guanosine, and uridine (data not shown). We cannot reproduce the charges exactly because the optimized geometry and orientation of the structures used to generate PARM99 are not available. The ribose sugar charge was obtained by equivalencing the four natural nucleosides. Although C1' and H1' atoms were not equivalenced in Cieplak et al.,²⁹ we did not see any major changes in the charges with or without C1' and H1' equivalencing. To provide sugar charges that are uniform among all the modified nucleotides, we decided to equivalent the C1' and H1' atoms along with all other sugar atoms (data not shown). To confirm that our modular fit reproduces the charges generated on whole nucleosides, we compared the results of a QM calculation on a whole adenosine nucleoside versus an adenosine with a methyl replacing the ribose. Table 4 shows the comparison between charges obtained with RESP on the nucleoside vs the methylated free base. Charges from the two methods agree with a standard deviation of 0.0506e for adenosine (0.0594e in the case of guanosine), suggesting that our modular approach is a faithful way of obtaining the charges for these large molecules. As mentioned above, the advantage of using modular approach is to combine different kinds of sugars with different kinds of modified bases thereby avoiding expensive computational calculations. In addition, the largest deviations are observed on the quaternary carbons C8, C6, and C4, which are well-known to be difficult to determine accurately.²⁰ When we compared our charges generated for adenosine with the charges reported in PARM99, there was good agreement with a standard deviation of 0.0362e. Once the modular approach was tested, it was used to produce the atom centered partial

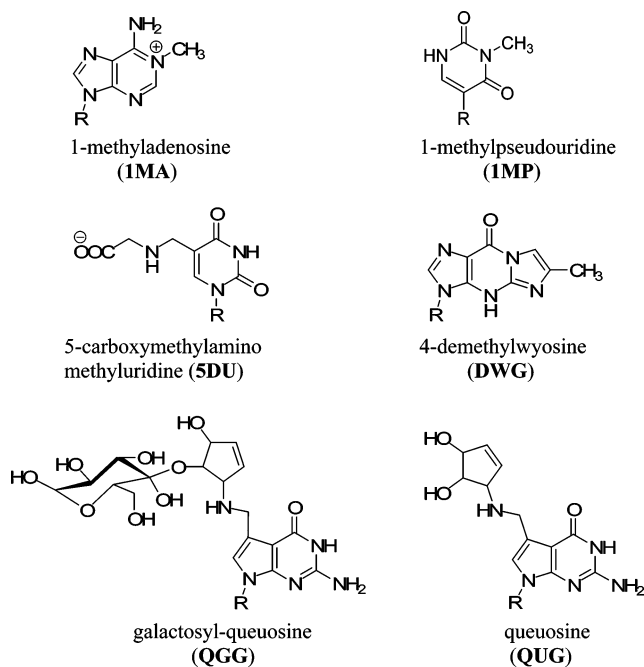


Figure 4. Examples of modified nucleosides present in RNA. These modifications range from simple methylation, as in the case of 1MA, to more complex carbohydrate containing compounds such as QGG. The generic names and their three-letter codes are also given. Only the hydrogen atoms on the polar atoms are shown for clarity. In each case, the only lowest energy tautomer was considered for the protonation that exists at pH 7.

Table 5. Charges Obtained for Pseudouridine, Inosine, and 5-Methylcytosine

	pseudouridine		inosine		5-methylcytosine
C5	-0.2218	N9	-0.0112	N1	-0.0674
C4	0.6913	C8	0.0627	C2	0.7939
O4	-0.5851	N7	-0.5341	O2	-0.6289
N3	-0.4208	C5	0.1198	N3	-0.7268
C2	0.5871	C6	0.5805	C4	0.6304
O2	-0.5729	O6	-0.5538	N4	-0.8933
N1	-0.3019	N1	-0.5208	H41	0.4095
C6	-0.1208	C2	0.3594	H42	0.4095
H6	0.2061	N3	-0.6184	C5	-0.0510
HN1	0.3084	C4	0.3503	C6	-0.1962
HN3	0.3121	H2	0.1223	H6	0.2158
		H1	0.3461	C10	-0.2707
		H8	0.1791	H20	0.0856
				H21	0.0856
				H22	0.0856

charges for all 107 naturally occurring modified nucleotides found in RNA. Table 5 contains the charges obtained for pseudouridine, inosine, and 5-methylcytosine.

To confirm the effect of multiorientation on charge derivation, we used R.E.D. II to apply multiorientation methodology and generate the atom-centered partial charges. Pseudouridine was used with the 5' and 3' oxygens capped with hydrogen to reduce the computational burden. As there is no literature available on the optimum number of orientations necessary to get reproducible charges, we decided to perform 4, 8, 12, and 20 different orientations, which are

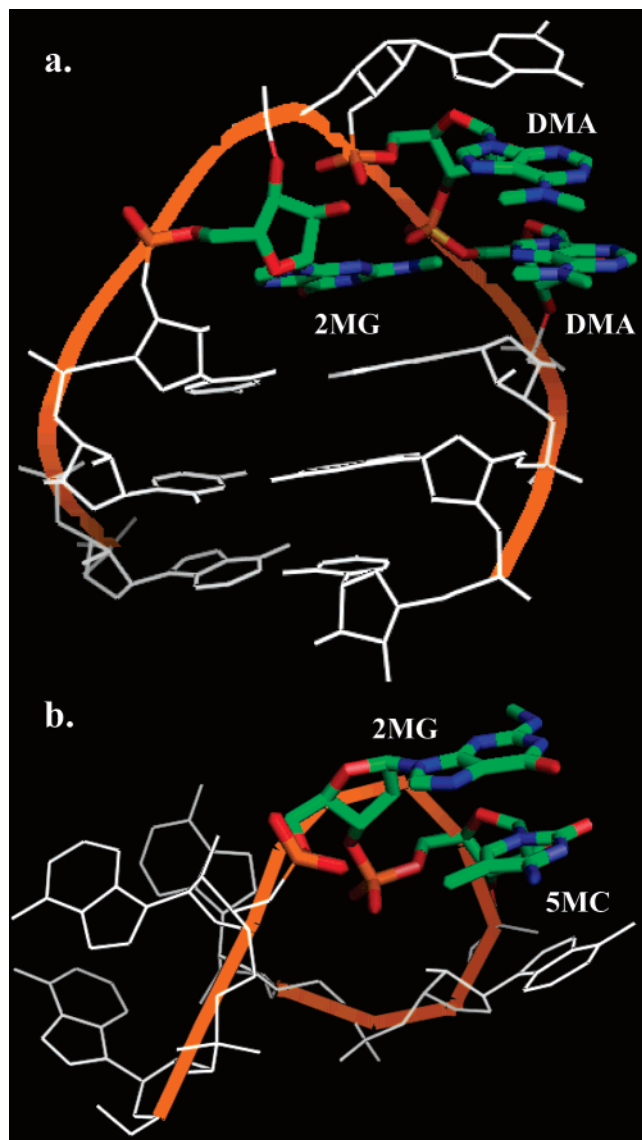


Figure 5. The effect of modifications in the stability and functioning of 16S rRNA of 30S ribosome (1J5E). (a) The dimethylated adenines (DMA) in the “dimethyl A loop” of 16S rRNA help in the stabilization of the loop through stacking interactions and forms a hydrophobic pocket with 2MG. (b) The methylated 966 and 967 positions of 16S rRNA increase the surface area for stacking and also form a van der Waals contact with the hydrophobic portion of Arg-128 of S9 protein (not shown).

shown as RED_4, RED_8, RED_12, and RED_20 in Supporting Information Table 1. R.E.D. was also performed on the orientation obtained from the GAUSSIAN calculation, which is given as RED_1 in Supporting Information Table 2. The orientations used in RED_4 were retained in the RED_8 case and so on. We did not observe significant changes between the charge values in comparing each case. The standard deviation between the RED_20 to RED_1 is 0.0149e. Thus, the change observed in the charge values from a single orientation to multiple orientations is insignificant. To overcome the multiorientation effect on the charges, we decided to increase the grid size (the number of shells of points as well as the density of points per shell) in calculating

the electrostatic potential. Different grid sizes were tested with Merz–Kollman charge fitting methodology as well as the CHELP-G method.⁴² We calculated electrostatic potential with four different options. In the first case, we computed the electrostatic potential using four shells and with a density of one point per every square angstrom (MK). We increased the density of points to four per \AA^2 in the second case (MK 4,4). We kept the density of ESP points the same and increased the number of shells from four to eight in the third case (MK 4,8) and used eight shells with a density of eight points per \AA^2 (MK 8,8). No significant changes in charge values are observed as the number of shells or the density of ESP points are increased using Merz–Kollman charge fitting methodology. We also used the CHELP-G method with four shells and density of one point per \AA^2 (ChelpG) and with eight shells and a density of eight points per \AA^2 (ChelpG 8,8). We did not see any major fluctuations in the charge values from using MK vs ChelpG methods. The results are summarized in Supporting Information Table 2. Since the goal of the present study is to develop parameters for modified nucleosides that are consistent with Cornell et al.²⁶ force field, the protocol outlined in Cieplak et al.²⁹ (i.e., calculating ESP with four layers and a density of one point per \AA^2) is sufficient to produce the atom-centered partial charges for the modified nucleosides present in RNA. Using the given optimized geometries (available at <http://ozone3.chem.wayne.edu>) to perform QM and RESP calculations as described, the charges reported herein can be readily reproduced.

5.1. Testing and Verifying the accuracy of Parameters.

Once the parameters for the 107 known modifications in RNA were computed, they were incorporated into AMBER to test the stability of MD trajectories of RNA with modified nucleotides. A molecular dynamics simulation of yeast tRNA^{Phe} containing 14 different modifications⁴⁰ was carried out using the crystal structure 1EHZ.pdb for the starting coordinates. The parameters for all the 14 modified nucleosides were successfully incorporated into LEAP, which properly generated the topology and coordinate files for this highly modified tRNA. SANDER¹⁵ was then used to do the energy minimization and molecular dynamics with two different methods: (a) use of implicit solvent with generalized-Born electrostatics⁴³ and (b) use of explicit solvent with particle-mesh Ewald electrostatics.^{44,45} To test the significance of the presence of modifications in the stability and functioning of tRNA^{Phe}, we wanted to study the MD of tRNA^{Phe} without modifications. To build the unmodified version of tRNA^{Phe}, the modified nucleosides were replaced with their respective common nucleosides (e.g., DHU with uridine) using the RNA-123 software suite developed in our lab for the analysis of RNA structures as well as 3D structure prediction of RNA. We performed 1 ns generalized-Born simulations on both the modified as well as the unmodified tRNA^{Phe}. In the case of tRNA^{Phe} with all 14 modifications, the structure remains stable throughout a 1 ns simulation using generalized-Born implicit solvent dynamics (data not shown), implying that the parameters developed can be reliably used in AMBER for simulating RNA with modifications. Further studies on these two systems using explicit

solvent conditions as well as crystallographic conditions will definitely help in understanding the role of modifications in the stability and functioning of tRNA. We are in the process of studying the effects of these modifications in the stability of tRNA^{Phe} by performing long time-scale AMBER molecular dynamics with explicit solvation on the structure with modifications and on the corresponding structure lacking modifications. Simulations of tRNA^{Phe} in the crystalline environment with the periodic boundary conditions present in the crystal are also being done.

Thus far, several modifications have been successfully incorporated into RNA-123. With the availability of the geometries for these modified nucleosides, we were able to model all 12 known modifications in *E. coli* 16S rRNA into *T. thermophilus* 30S ribosome crystal structure (1J5E). Interestingly, all of the modifications were accommodated in the published PDB structure without any steric conflicts. Further, the placement of the modifications suggests functional roles for them in increased stacking or formation of hydrophilic pockets for protein binding (Figure 5). The PDB coordinates for the modified 16S rRNA are available at our group home page <http://ozone3.chem.wayne.edu>.

5.2. Implementation of the Modified Nucleotide Parameters in Other Force Fields. In addition to using the charges obtained from our study in generating parameters for AMBER, we have also used some of these charge values in CNS,¹⁶ which is based on the CHARMM force field.¹⁴ Parameter files for CNS can be created using programs such as PRODRG and XPLO-2D,⁴⁶ but these files do not contain charge values. We introduced the charge values into the parameter files of pseudouridine for CNS. These parameter files, having the charge values from this study, were used in NMR structure calculations for the 1060 hairpin loop of human 18S rRNA, which contains a single pseudouridine residue.⁴⁷

5. Conclusions

We have successfully developed and implemented AMBER force field parameters for the 107 naturally occurring modified nucleosides present in RNA. As the evidence for the versatile functions of RNA in the cell is expanding, it is becoming apparent that modified nucleosides play important roles in achieving these functions. The availability of force field parameters for modified nucleosides enhances the functionality of AMBER and thereby will contribute to understanding how modified nucleosides participate in the function and structural stabilization of RNA. The modified nucleoside parameters described herein allow for AMBER MD simulations and molecular mechanics for all modified RNAs. Further the modular approach allows for many new combinations of base and/or sugar modified nucleotides to be readily computed.

Acknowledgment. The authors would like to thank Dr. Thomas Cheatham, III and Prof. David Case for useful discussions and suggestions. We thank Dr. Jason Sonnenberg for help with the GAUSSIAN calculations. We would also like to thank Larry Clos, Frederick Sijenyi, and Marcus Wood for critical analysis of the manuscript. We also thank

Wayne State University Grid Computing Facility for the computational time and support. This work was supported by NIH grant GM073179.

Supporting Information Available: Effects of multiorientation on atom-centered partial charges and the influence of the number of shells and the density of points used in calculating the electrostatic potential. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Noller, H. F. RNA structure: reading the ribosome. *Science* **2005**, *309*, 1508–14.
- (2) Sen, G. L.; Blau, H. M. A brief history of RNAi: the silence of the genes. *FASEB J.* **2006**, *20*, 1293–9.
- (3) Krutzfeldt, J.; Stoffel, M. MicroRNAs: a new class of regulatory genes affecting metabolism. *Cell Metab.* **2006**, *4*, 9–12.
- (4) Rozenski, J.; Crain, P. F.; McCloskey, J. A.; The RNA Modification Database: 1999 update. *Nucleic Acids Res.* **1999**, *27*, 196–7.
- (5) Henri Grosjean, R. B. *Modification and Editing of RNA*; ASM Press: 1998.
- (6) Murphy, F. V. t.; Ramakrishnan, V.; Malkiewicz, A.; Agris, P. F. The role of modifications in codon discrimination by tRNA(Lys)UUU. *Nat. Struct. Mol. Biol.* **2004**, *11*, 1186–91.
- (7) Sumita, M.; Desaulniers, J. P.; Chang, Y. C.; Chui, H. M.; Clos, L., 2nd; Chow, C. S. Effects of nucleotide substitution and modification on the stability and structure of helix 69 from 28S rRNA. *RNA* **2005**, *11*, 1420–9.
- (8) Agris, P. F. Decoding the genome: a modified view. *Nucleic Acids Res.* **2004**, *32*, 223–38.
- (9) Blount, K. F.; Uhlenbeck, O. C. The structure-function dilemma of the hammerhead ribozyme. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 415–40.
- (10) Bruno, L. *Conserved ribosomal RNA modification and their putative roles in ribosome biogenesis and translation*; Springer Berlin/Heidelberg: 2005; Vol. 12, pp 263–284.
- (11) Alexandrov, A.; Chernyakov, I.; Gu, W.; Hiley, S. L.; Hughes, T. R.; Grayhack, E. J.; Phizicky, E. M. Rapid tRNA decay can result from lack of nonessential modifications. *Mol. Cell* **2006**, *21*, 87–96.
- (12) Lane, B. G.; Ofengand, J.; Gray, M. W. Pseudouridine and O²-methylated nucleosides. Significance of their selective occurrence in rRNA domains that function in ribosome-catalyzed synthesis of the peptide bonds in proteins. *Biochimie* **1995**, *77*, 7–15.
- (13) Auffinger, P.; Westhof, E. RNA solvation: A molecular dynamics simulation perspective. *Biopolymers* **2000**, *56*, 266–274.
- (14) MacKerell, A. D., Jr.; Bashford, D.; Bellott, R. L.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S. G.; J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C. M.; S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B. S.; M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J. Y.; D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

- (15) Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–88.
- (16) Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1998**, *54*, 905–21.
- (17) Scott, W. R. P.; Hunenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Kruger, P.; van Gunsteren, W. F. The GROMOS Biomolecular Simulation Program Package. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.
- (18) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X.; Murphy, R. B.; Zhou, R.; Halgren, T. A. Development of a polarizable force field for proteins via ab initio quantum chemistry: first generation model and gas phase tests. *J. Comput. Chem.* **2002**, *23*, 1515–31.
- (19) Halgren, T. A.; Damm, W.; Polarizable, force fields. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236–42.
- (20) Lind, K. E.; Sherlin, L. D.; Mohan, V.; Griffee, R. H.; Ferguson, D. M. Parameterization and simulation of the physical properties of phosphorothioate nucleic acids. In *Molecular Modeling of Nucleic Acids*; Leontis, N. B.; SantaLucia, J., Eds.; ACS Symposium Series 682; American Chemical Society: Washington, DC, 1998; pp 41–54.
- (21) Meagher, K. L.; Redman, L. T.; Carlson, H. A. Development of polyphosphate parameters for use with the AMBER force field. *J. Comput. Chem.* **2003**, *24*, 1016–25.
- (22) Lind, K. E.; Mohan, V.; Manoharan, M.; Ferguson, D. M.; Structural, characteristics of 2'-O-(2-methoxyethyl)-modified nucleic acids from molecular dynamics simulations. *Nucleic Acids Res.* **1998**, *26*, 3694–3699.
- (23) Stuart, J. W.; Koshlap, K. M.; Guenther, R.; Agris, P. F.; Naturally-occurring modification restricts the anticodon domain conformational space of tRNA(Phe). *J. Mol. Biol.* **2003**, *334*, 901–18.
- (24) McCrate, N. E.; Varner, M. E.; Kim, K. I.; Nagan, M. C.; Molecular, dynamics simulations of human tRNA^{Lys}, 3 UUU: the role of modified bases in mRNA recognition. *Nucleic Acids Res.* **2006**, *34*, 5361–8.
- (25) Auffinger, P.; Louise-May, S.; Westhof, E.; Molecular, dynamics simulations of solvated yeast tRNA(Asp). *Biophys. J.* **1999**, *76*, 50–74.
- (26) 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 Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–97.
- (27) Durant, P. C.; Bajji, A. C.; Sundaram, M.; Kumar, R. K.; Davis, D. R.; Structural, effects of hypermodified nucleosides in Escherichia, coli and human tRNA^{Lys}, anticodon loop: The effect of nucleosides s2U, mcm5U, mcm5s2U, mmm5s2U, t6A, and ms2t6A. *Biochemistry* **2005**, *44*, 8078–89.
- (28) Sierzputowska, G. H.; Sochacka, E.; Malkiewicz, A.; Kuo, K.; Gehrke, C. W.; Agris, P. F. Chemistry and structure of modified uridines in the anticodon, wobble position of transfer RNA are determined by thiolation. *J. Am. Chem. Soc.* **1987**, *109*, 7171–7177.
- (29) Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. Application of the multimolecule and multiconformational RESP methodology to biopolymers: charge derivation for DNA, RNA, and proteins. *J. Comput. Chem.* **1995**, *16*, 1357–77.
- (30) Frisch, M. J. T.; G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Scalmani, G.; Kudin, K. N.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; 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.; Li, X.; Hratchian, H. P.; Peralta, J. E.; Izmaylov, A. F.; Brothers, E.; Staroverov, V.; Kobayashi, R.; Normand, J.; Burant, J. C.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Chen, W.; Wong, M. W.; Pople, J. A. *Gaussian DV, Revision E.05*; Gaussian, Inc.: Wallingford, CT, 2006.
- (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*, 10269–80.
- (32) Steenken, S. Purine bases, nucleosides, and nucleotides: aqueous solution redox chemistry and transformation reactions of their radical cations and e- and OH adducts. *Chem. Rev. (Washington, DC, U.S.)* **1989**, *89*, 503–20.
- (33) Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. Atomic charges derived from semiempirical methods. *J. Comput. Chem.* **1990**, *11*, 431–9.
- (34) U. Chandra Singh; Kollman, P. A. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.* **1984**, *5*, 129–145.
- (35) Woods, R. J.; Pell, W. M. K.; Moffat, S. H.; Smith, V. H., Jr. Derivation of net atomic charges from molecular electrostatic potentials. *J. Comput. Chem.* **1990**, *11*, 297–310.
- (36) Pigache, A.; Cieplak, P.; Dupradeau, F.-Y. In *Automatic, highly reproducible and effective RESP and ESP charge Derivation: Application to the development of programs RED and X RED*, 227th ACS National Meeting, Anaheim, CA, March 28–April 1, 2004; Anaheim, CA, March 28–April 1 2004.
- (37) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247–60.
- (38) 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*, 1157–74.

- (39) Sankaranarayanan, R.; Dock-Bregeon, A. C.; Romby, P.; Caillet, J.; Springer, M.; Rees, B.; Ehresmann, C.; Ehresmann, B.; Moras, D. The structure of threonyl-tRNA synthetase-tRNA(Thr) complex enlightens its repressor activity and reveals an essential zinc ion in the active site. *Cell* **1999**, *97*, 371–81.
- (40) Shi, H.; Moore, P. B. The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: a classic structure revisited. *RNA* **2000**, *6*, 1091–1105.
- (41) NC-IUB, Nomenclature for incompletely specified bases in nucleic acid sequences Recommendations 1984. *Eur. J. Biochem.* **1985**, *150*, 1–5.
- (42) Breneman, C. M.; Wiberg, K. B. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comput. Chem.* **1990**, *11*, 361–373.
- (43) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- (44) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An $N \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (45) York, D. M.; Darden, T. A.; Pedersen, L. G. The effect of long-range electrostatic interactions in simulations of macromolecular crystals: A comparison of the Ewald, and truncated list methods. *J. Chem. Phys.* **1993**, *99*, 8345–8348.
- (46) Kleywegt, G. J. *XPLOR2D, 051002*; Department of Cell and Molecular Biology, Uppsala University: Uppsala, SWEDEN.
- (47) Spahn, C. M.; Beckmann, R.; Eswar, N.; Penczek, P. A.; Sali, A.; Blobel, G.; Frank, J.; Structure of the 80S ribosome from *Saccharomyces cerevisiae*—tRNA-ribosome and subunit-subunit interactions. *Cell* **2001**, *107*, 373–86.

CT600329W

On the Use of Elevated Temperature in Simulations To Study Protein Unfolding Mechanisms

Ting Wang* and Rebecca C. Wade

*Molecular and Cellular Modeling Group, EML Research,
Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany*

Received March 15, 2007

Abstract: In protein unfolding simulations, elevated temperature, significantly exceeding the melting temperature T_m , provides an important means to accelerate unfolding to a computationally accessible time range. This procedure is based on the assumption that protein thermal unfolding has Arrhenius behavior and therefore that increasing temperature does not alter the protein unfolding pathways. However, in nature, proteins can show non-Arrhenius behavior and, in practice, overly fast unfolding in high-temperature simulations can result in difficulties in identifying unfolding intermediates and distinguishing their relative stabilities. In this paper, we describe simulations of two WW domains, small protein domains that have a three-stranded β -sheet structure. Simulations were carried out at several temperatures ranging from 300 K to 500 K, starting from folded structures. The results demonstrate the temperature dependence of the unfolding pathways, showing that to obtain unfolding pathways corresponding to those observed in experiments, the elevation of the simulation temperature has to be controlled. Based on trajectory analysis, we proposed a qualitative criterion for judging when an elevated temperature is acceptable or not, namely, that the temperature must be such that the native folded state is sampled substantially before protein unfolding begins. While depending on force field parameters and protein fold complexity, this criterion can be quantified to obtain the upper bound of an “acceptable elevated temperature”, which was observed to be dependent on the thermostabilities of the two WW domain proteins.

Introduction

In principle, molecular dynamics simulations to study protein unfolding and folding mechanisms should be conducted at the protein melting temperature (T_m) where the protein reaches folding/unfolding equilibrium. In practice, due to the long time scale ($> 10 \mu\text{s}$ for the fastest folding protein) of equilibrium unfolding and the limitations of computer power, highly elevated temperature (much above T_m) has been used in most unfolding simulations^{1–8} to enable unfolding in an accessible computational time ($< 100 \text{ ns}$). The basis for the use of elevated temperature is the

assumption that protein thermal unfolding shows Arrhenius behavior. That is, as stated in Beck and Daggett’s review,⁹ “increasing temperature does not alter the pathway of unfolding, only the rate”. However, proteins can show non-Arrhenius behavior.^{10–14} It has been observed that increasing temperature can slow down folding^{11,12} and change intermediate states.^{13,14} Two explanations for non-Arrhenius behavior are kinetic traps and the temperature dependence of hydrophobic interactions. For simple two-state folding, the temperature dependence of the hydrophobic effect is the main reason.¹⁰ On the other hand, at very high temperature, protein structures can unfold very rapidly, and a large number of intramolecular interactions can be disrupted simultaneously. As a result, it is not possible to detect any order in the unfolding events or an unfolding pathway or funnel. In particular, when long-lived intermediates are expected, high

* Corresponding author phone: (530)754-5625; fax: (530)754-9658; e-mail: twang@ucdavis.edu. Current address: Genome Center and Bioinformatics Program, 431 East Health Science Drive, University of California, Davis, CA 95616-8816.

temperature can shorten their lifetime and they can thereby go undetected.

A range of simulation temperatures can be used as in the replica exchange method (REMD) to fully sample the free energy landscape.^{15–17} Folding/unfolding pathways can be assessed at an appropriate sampling temperature. Furthermore, if sampling is complete, then T_m can be found by considering heat capacity, and protein kinetics and unfolding pathways can be computed from the shapes and heights of free energy barriers between minima. The problem is that such complete sampling is hard to achieve for real systems. Another problem is that the T_m computed from REMD has often been found to be higher than experimental values due to force fields being parametrized at room temperature.¹⁷ This can lead to overestimation of protein stability.

Consequently, the following questions arise: To what extent can temperature be elevated to accelerate unfolding without changing unfolding pathways or to reproduce the unfolding pathways observed at T_m ? In other words, what is the ‘acceptable elevated temperature’ for simulating the unfolding of a given protein? And what criteria can be used to estimate whether a trajectory at an elevated temperature will give insights into the early stages of the unfolding process of the protein itself? When unfolding simulations are used to study the relative stabilities of two proteins, is the acceptable elevated temperature for each protein a good measure for relative protein stability? To address these questions, and being motivated by our study of the relative stabilities of WW domains and mutants by molecular dynamics simulations,¹⁸ we studied the unfolding of two WW domains: one from formin binding protein 28 (FBP28) and one from Yes kinase-associated protein 65 (YAP65). Both of these WW domains consist of ca. 40 residues and have a three-stranded β -sheet. FBP28 has been studied extensively by both experiment and computation. It is the only WW domain that has been observed to exhibit biphasic folding/unfolding kinetics with a stable intermediate state.^{14,17,19,20} Nguyen et al. reported that FBP28 exhibits folding/unfolding kinetics that are tuneable by temperature, being a two-state folder with T_m of 337 K and exhibiting a stable intermediate below T_m at 312.5 K.¹⁴ The origin of the biphasic kinetics has been attributed to a unique hydrophobic packing that is absent in other WW domains.^{19,20} However, in contradicting studies, other authors have found that FBP28 exhibits two-state behavior at a range of temperatures below T_m .^{8,21} A stable intermediate state that consists of only the first and the second native β -strands has been detected in both experimental and computational works.^{14,17,19,20} Despite the controversy, consensus has been reached on the relative stabilities of the two β -hairpins: the first is more persistent in unfolding processes.^{8,22}

The thermal unfolding of YAP65 has been studied experimentally and has been reported to be a two-state process at and below $T_m = 323$ K¹² with transition states alterable by temperature.^{12,23} Compared with FBP28, YAP65 has been less studied by simulation and its atomic level unfolding details are less clear. One of the few simulations predicted that the unfolding of YAP65 begins with the loss of the third strand.²⁴

Our simulations demonstrate that the trajectories of the WW domains generated at different temperatures show different unfolding pathways. Analysis of the trajectories suggests criteria for determining suitable elevated temperatures for unfolding simulations. The qualitative criterion is that the simulation temperature be such that the majority of independent simulations starting in the native folded state sample the native state substantially before the unfolding process begins. This can only happen if the temperature is low enough for the native state to be a local energy minimum. We call this temperature the ‘acceptable elevated temperature’. It is dependent on the stability of the protein. We further propose a procedure for quantitatively applying this criterion based on computation of the native contact percentage and the radius of gyration of the protein in simulations at room temperature and elevated temperature.

Methods

Molecular Dynamics Simulation. Two WW domains, FBP28 and YAP65, were simulated by using the AMBER8 program²⁵ with the TIP3P explicit water model and the AMBER ff03 force field.²⁶ The first structure from the NMR ensemble (PDB entry 1e01) of FBP28 was used; this has 37 residues. The coordinates of the NMR structure of YAP65 were provided by Dr. Maria Macias; this has 39 residues. The simulation protocol was as described in ref 18.

The simulations were carried out at several temperatures ranging from 300 K to 500 K starting from the native folded states. At each temperature, three to five different simulations were conducted by varying the speed of heating of the system at the beginning of the simulations. The length of the simulations was chosen according to the simulation temperature. It was 20 ns for simulations at 300 K. Simulations at higher temperatures were run for at least 12 ns. They were stopped either after unfolding was completed or after 72 ns if no unfolding occurred. The coordinates were saved every 10 ps.

Trajectory Analysis. The unfolding process was monitored by computing the variation in the percentage of native contacts (Q_N) present during the simulations. A contact was defined to be present when the van der Waals spheres of any backbone atoms from two different residues were within 1 Å of each other. The van der Waals parameters were taken from the AMBER force field.²⁶ This definition yielded 30 native contacts in FBP28 and 46 native contacts in YAP65. The greater number of native contacts in YAP65 is due to contacts in and between the N- and C-termini of the domain.

Each simulation trajectory was also analyzed by plotting the two-dimensional population histograms defined by percentage of native contacts (Q_N) (X -axis) and the radius of gyration (R_g) (Y -axis). The Z -values were plotted as the negative logarithm of the population in the bins defined by the Q_N and R_g coordinates. A more negative Z -value corresponds to a state with higher population and thereby with lower free energy. The bin size of the percentage of native contacts was 2% (less than 1 contact for both FBP28 and YAP65). The bin size of the radius of gyration was 0.075 Å. Because such small bin sizes were used, different members of the population in the same x,y coordinate bin had almost identical conformations.

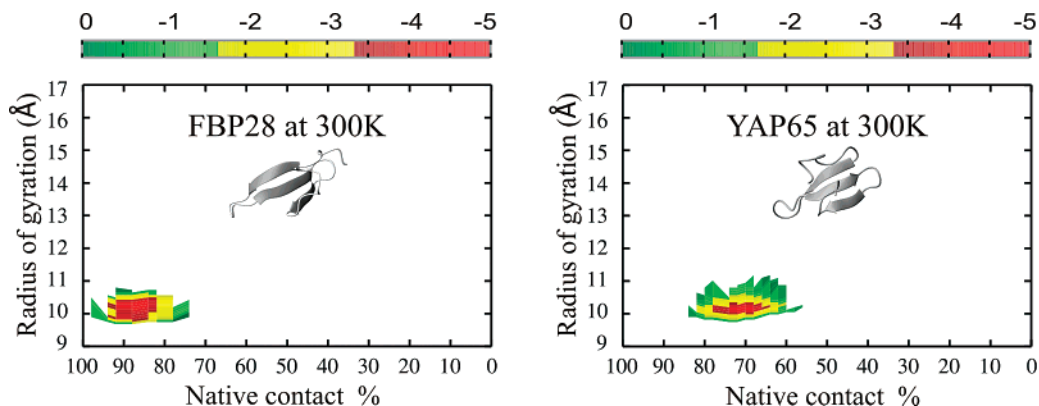


Figure 1. FBP28 and YAP65 simulated at 300 K for 20 ns. Shown are the two-dimensional population histograms defined by the percentage of native contacts (Q_N) and radius of gyration (Rg). The simulations sampled very narrow regions around the proteins' native starting structures, which were defined as the native state regions of $Q_N = 82\text{--}100\%$, $R_g = 9.4\text{--}10.4 \text{ \AA}$ for FBP28 and $Q_N = 64\text{--}100\%$, $R_g = 9.5\text{--}10.5 \text{ \AA}$ for YAP65.

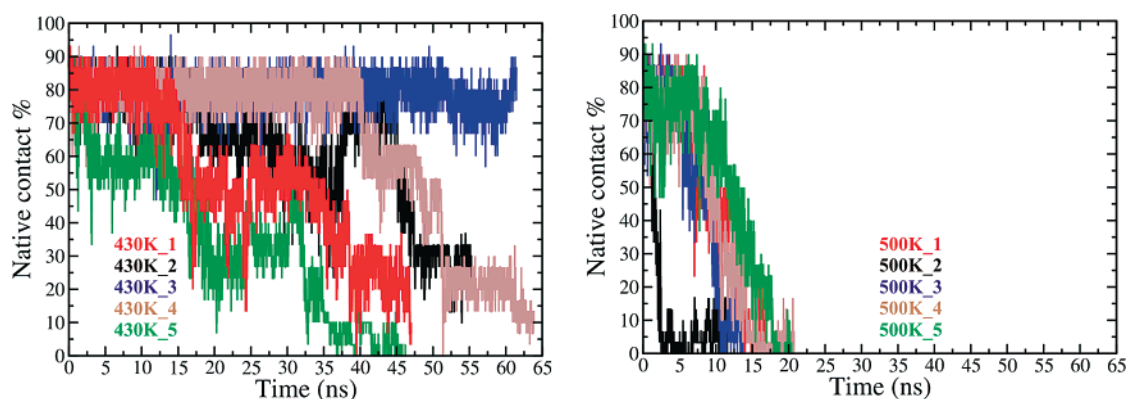


Figure 2. Time development of percentages of native contacts in the simulations of FBP28 at 430 K and 500 K. Five simulations were run at each temperature. At 430 K, four (430K_1, 430K_2, 430K_4, and 430K_5) of the five simulations left less than 30% native contacts at the ends (47 ns, 55 ns, 64 ns, and 46 ns), while the third simulation (430K_3) still maintained more than 80% native contact at the end (61 ns). At 500 K, less than 10% native contacts were left at the end of all the five simulations (of 17 ns, 12 ns, 13 ns, 21 ns, and 21 ns duration).

Most of the simulation trajectories were also projected onto the two-dimensional population histograms defined by percentages of native contacts between the β_1 and β_2 strands (Q_{12}) and the β_2 and β_3 strands (Q_{23}).

Results and Discussion

Simulations at 300 K Define Native State Regions. Both the WW protein domains maintained their folded states during the 20 ns long simulations at 300 K. Figure 1 shows the two-dimensional population histograms defined by percentage of native contacts (Q_N) and radius of gyration (Rg). Both protein domains sampled very narrow regions around their native starting states, peaking at $Q_N = 88\%$, $R_g = 10.2 \text{ \AA}$ for FBP28 and at $Q_N = 70\%$, $R_g = 10.1 \text{ \AA}$ for YAP65. The fraction of native contacts maintained in YAP65 is lower due to the lower stability of native contacts in and between the N- and C-termini that are absent in the FBP28 structure. We then used the Q_N with the highest population (88% for FBP28 and 70% for YAP65) and the Rg of the starting structure (9.9 \AA for FBP28 and 10.0 \AA for YAP65) to define a native state region where Q_N should not be smaller than the peak Q_N minus 6% and Rg should stay within 0.5 \AA of the starting Rg. The 6% window was chosen based on

the population distribution of FBP28, beyond which the population showed a large drop from 268 structures with 82% native contacts to 0 structures with 80% native contacts and 79 structures with 78% native contacts. As a result, the native state region is at $Q_N = 82\text{--}100\%$, $R_g = 9.4\text{--}10.4 \text{ \AA}$ for FBP28 and at $Q_N = 64\text{--}100\%$, $R_g = 9.5\text{--}10.5 \text{ \AA}$ for YAP65. The native state region accounts for 85% of structures in the trajectory for FBP28 and 92% for YAP65.

FBP WW Domain Unfolds by Different Pathways at 430 K and 500 K. At 430 K, complete unfolding of FBP28 occurred in four out of five simulations. Figure 2 shows the time development of the percentages of native contacts in the five simulations. In the first simulation (referred as 430K_1), before the percentage of native contacts (Q_N) fell to 20%, there was a stage from 17 to 30 ns with a stable Q_N of around 50–60%, indicating a stable intermediate on the unfolding path. The two-dimensional population histogram in Figure 3 shows a second intermediate apart from the native state. This implies two energy minima separated by a clear energy barrier. This result is in agreement with the reported three-state behavior of FBP28^{14,17,19,20} at temperatures below T_m .¹⁴ The intermediate retains the β_1 and β_2 strands but has no β_3 strand. This is shown in Figure 4 in a two-dimensional

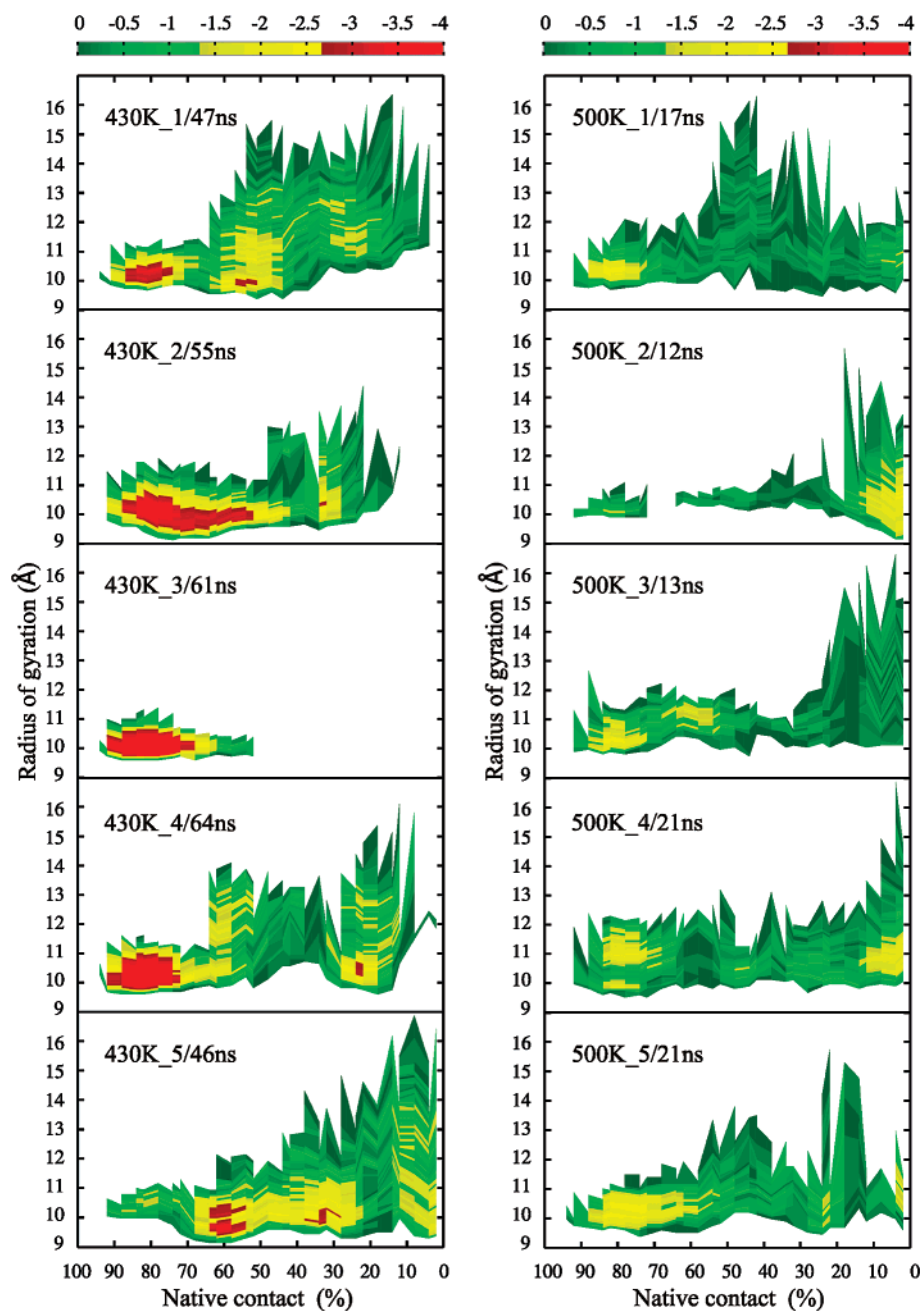


Figure 3. Two-dimensional population histograms of FBP28 simulated at 430 K and 500 K. The X-axis and the Y-axis are defined by the percentage of native contacts (Q_N) and radius of gyration (Rg). Five simulations were run at each temperature. In 430K_1, one minimum is in the native state region and the second minimum at $Q_N = 50\text{--}60\%$ and $R_g = \text{ca. } 10 \text{ \AA}$, indicating a stable intermediate state before complete unfolding. In 430K_2, there is a single and broad minimum that spans from the native state region to the region of $Q_N = 50\%$ before complete unfolding. In 430K_3, the protein did not unfold. In 430K_4, the only minimum is in the native state region. In 430K_5, the native state region was sampled very briefly. At 500 K, the protein unfolded in all three simulations without significantly sampling its native state region or intermediate region to unfolding.

population histogram defined by percentages of native contacts between $\beta 1$ and $\beta 2$ (Q_{12}) and $\beta 2$ and $\beta 3$ (Q_{23}). In addition, the small radius of gyration of the intermediate state (R_g around 10 \AA) implies that although native contacts between $\beta 2$ and $\beta 3$ were lost, non-native contacts were formed. In fact, the $\beta 3$ strand makes a U-turn in these structures. These non-native contacts were also observed in the experiments¹⁴ by Nguyen and co-workers and the simulations by Mu and co-workers.¹⁷ In addition, Mu and co-workers' computations showed that the intermediate structure has a much lower free energy of dimerization than

the native structure and is likely to be the initial structure to form amyloid fibrils.^{17,21}

Among the other three unfolding trajectories at 430 K (referred to as 430K_2, 430K_4, and 430K_5), 430K_4 exhibits two-state kinetics with the native states ($Q_N = 70\text{--}90\%$) and the unfolded states ($Q_N < 30\%$) being the only well populated states in the free energy landscape defined by Q_N and R_g as shown in Figure 3. However, in Figure 4 showing the two-dimensional population histogram defined by Q_{12} and Q_{23} , the states maintaining the $\beta 1\text{--}\beta 2$ contacts but with the $\beta 2\text{--}\beta 3$ contacts lost formed a second intermediate

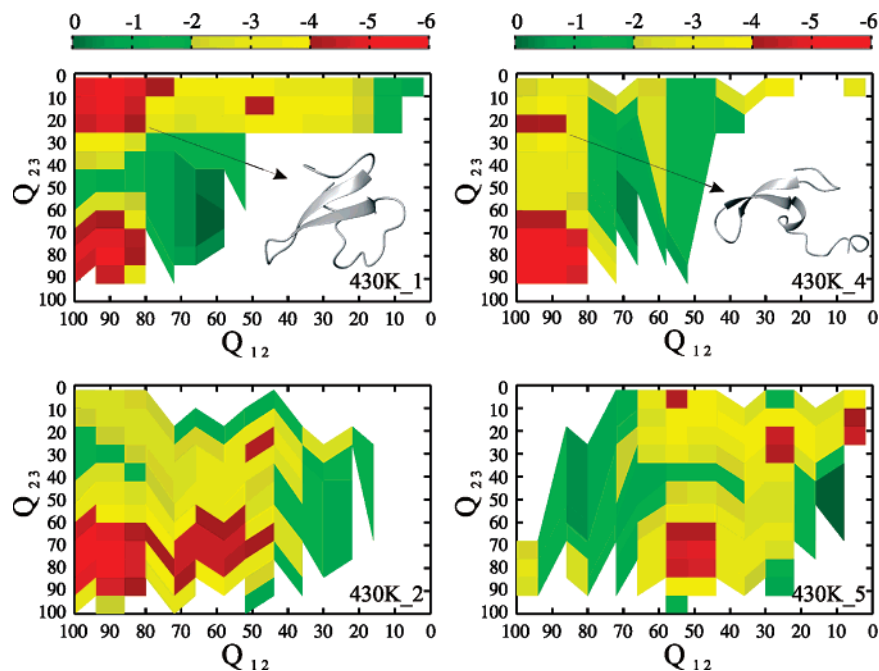


Figure 4. Two-dimensional population histogram defined by native contact percentages between $\beta 1$ and $\beta 2$ (Q_{12}) and $\beta 2$ and $\beta 3$ (Q_{23}) for FBP28 in the four unfolding trajectories. In 430K_1, the minimum around $Q_{12} = 90\%$, $Q_{23} = 10\%$ indicates a stable state that maintained the native contacts between $\beta 1$ and $\beta 2$ but lost the native contacts between $\beta 2$ and $\beta 3$. This corresponds to the intermediate state revealed in Figure 3. A similar state was also well populated in 430K_4, corresponding the minimum around $Q_{12} = 90\%$ and $Q_{23} = 20\%$. In 430K_2, the minima span almost the same regions for Q_{12} and Q_{23} at around 60–100%, which corresponds to the single and broad minimum shown in Figure 3. In 430K_5, no minimum is observable in the native state region.

separated from the native state but apparently without a high-energy barrier between these two states. In trajectory 430K_2, the protein sampled a region broad in Q_N (85–50%) and narrow in R_g (around 10 Å) where the native structure is still registered but with shorter β -strands before unfolding (Figures 3 and 4). This indicates a more two-state-like unfolding process. In trajectory 430K_5, the native state was not stable at all, and the well populated states before complete unfolding had ca. 60% of the native contacts.

At 500 K, FBP28 unfolded rapidly in all five simulations (see Figures 2 and 3). The native state region was sampled briefly before complete unfolding took place. No intermediate state was detectable. These trajectories thus suggest two-state unfolding behavior. Note that the trajectories were terminated after unfolding was completed, and therefore the relative sampling of the unfolded states (at $Q_N \sim 0\%$) in the different trajectories and compared to the native state as shown in Figure 3 is not meaningful.

The different results obtained for FBP28 at 430 K and 500 K demonstrate the temperature-dependence of the simulated unfolding of FBP28. At a temperature of 430 K, three-state behavior was observed, whereas it was absent at 500 K. We consider the temperature of 430 K an “acceptable” elevated temperature for studying the unfolding mechanism of FBP28. It should be clear that the five trajectories vary at 430 K, 1 showing no unfolding, 1 showing three-state behavior, 1 showing two-state behavior, and 2 showing intermediate behavior. This is the single molecule vs molecular ensemble problem in molecular dynamics simulations. The ensemble behavior observed in experiments may be reached by running a large number of simulations.

Nevertheless, at the temperature of 430 K, even in the trajectories showing two-state behavior, the higher stability of the first β -hairpin was clearly observed. This property of the FBP28 WW domain was hard to detect at the temperature of 500 K.

Criteria for Acceptable Elevated Temperature. Herein, we propose a criterion for judging an elevated temperature based on the analysis of the unfolding trajectories. That is, the majority of the simulations starting from the native state should sample the native state substantially, resulting in a deep and narrow minimum around the native state region in a two-dimensional population histogram defined by native contact percentage and radius of gyration. This can only happen if the temperature is low enough for the native state to be a local energy minimum. Accordingly, as any intermediate on an unfolding pathway is assumed to be less stable than a native state, a lower temperature should be used to detect intermediate states. If the native state is lost too quickly in a trajectory, the simulation temperature is probably too high to be acceptable. As a result, the acceptable elevated temperature of a protein is dependent on its stability.

To quantify this criterion, we computed the number of structures saved during the 10 FBP28 WW domain trajectories simulated that sampled the native state region. The number of structures is 668, 866, 3522, 1931, and 61, respectively, in the five simulations at 430 K and 241, 56, 165, 91, and 255, respectively, in the five simulations at 500 K. We then suggest that, for the simulated systems, an elevated temperature is acceptable only when it leads to sampling of the native state region in at least 500 saved structures. Because the structures were saved every 10 ps,

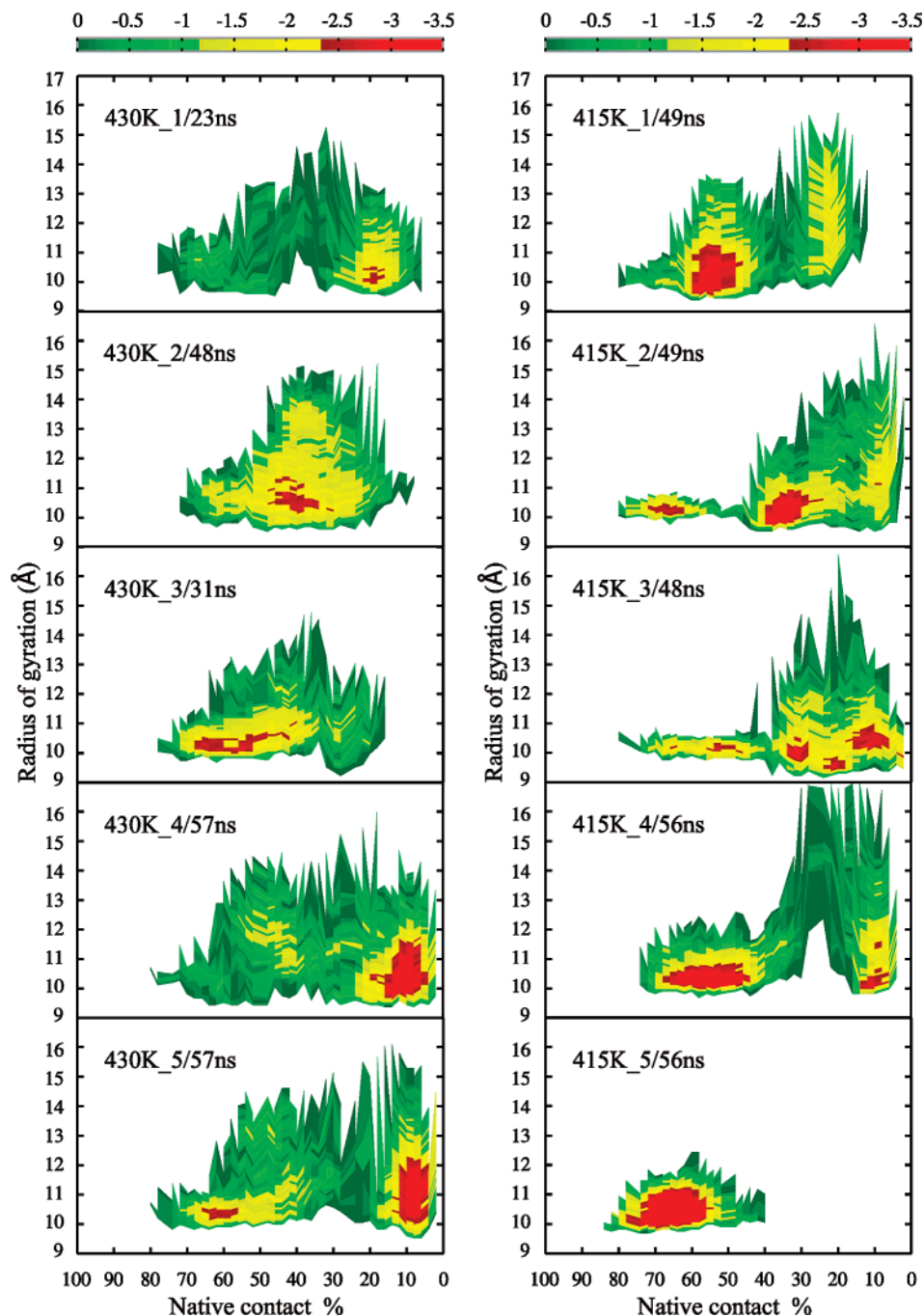


Figure 5. Two-dimensional population histogram of YAP65 simulated at 430 K and 415 K. The X-axis and the Y-axis are defined by the percentage of native contacts (Q_N) and the radius of gyration (R_g). At 430 K, the protein achieved complete unfolding in all five simulations but the native state region ($Q_N=64-100\%$, $R_g=9.5-10.5$ Å) was sampled very briefly (far less than 5 ns (see text)). According to the criterion described in the text, 430 K is not an acceptable elevated temperature for unfolding YAP65. At 415 K, sampling of the native state region was substantially increased and reached the time of 5 ns in one (415K_1) of the four unfolding trajectories, indicating that the temperature of 415 K is acceptable.

this implies a total of 5 ns of sampling of the native state. When an unfolding trajectory satisfies this criterion at an elevated temperature of T , it means that the trajectory may reveal unfolding mechanisms corresponding to those obtained in experiments, and furthermore, temperature T is considered as an “acceptable elevated temperature”. But it does not mean that all simulation trajectories obtained at this temperature will satisfy this criterion. Once this happens, as in trajectory 430K_5, it may indicate T as being the upper limit of the elevated temperature. In other words, it indicates that a

temperature higher than T should not be used. In addition, it is worth noting that this empirical quantitative criterion is likely to be highly dependent on the force field used, and it may vary significantly for a different force field.

We then applied this criterion to simulations of the YAP65 WW domain. The unfolding simulations were first performed at 430 K, and complete unfolding occurred in all five simulations, lasting for 23 ns, 48 ns, 31 ns, 57 ns, and 57 ns, respectively. The two-dimensional population histograms defined by percentage of native contacts (Q_N) and radius of

gyration (Rg) are shown in Figure 5. We can see that the minimum corresponding to the native state of YAP65 is notable in 430K_3 and 430K_5 but absent in the other three simulations (430K_1, 430K_2, and 430K_4). The native state region was sampled 65, 47, 371, 77, and 241 times, respectively. By the criterion proposed above, the temperature of 430 K is not acceptable for studying the unfolding mechanisms of YAP65, and a lower temperature should be used. We then set up simulations at 370, 400, and 415 K. No unfolding was observed in the one simulation at 370 K and the three simulations at 400 K in simulation times of 72 ns. Unfolding was observed at 415 K. Five simulations were carried out at 415 K, lasting for 49 ns, 49 ns, 48 ns, 56 ns, and 56 ns, respectively. Unfolding was completed in four (415K_1, 415K_2, 415K_3, and 415K_4) of the five simulations. In Figure 5, we can see that the shape of the landscape defined by Q_N and Rg appears to differ in simulations at 415 K and at 430 K. Overall, the sampling of the native state region was substantially increased when the temperature decreased from 430 K to 415 K, reaching 162, 453, 233, 506, and 2353 times, respectively. Moreover, trajectory 415K_4 satisfied the criterion of a total of 5 ns sampling of the native state region before complete unfolding. This indicates that the temperature of 415 K is acceptable for simulating the unfolding of YAP65 and may be the upper limit of an elevated temperature as the other three unfolding trajectories (415K_1, 415K_2, and 415K_3) did not satisfy the criterion. The lower acceptable temperature required for YAP65 than FBP28 is consistent with the lower stability of YAP65. In addition, trajectory 415K_4 showed a clear two-state unfolding, and the unfolding began with the loss of the third strand, which is consistent with experiment^{12,23} and our previous simulation.²⁴

The acceptable temperature for YAP65 is found to be 15 K lower than that of FBP28 in these simulations. This is remarkably consistent with the 14 K lower T_m of YAP65 compared to FBP28, considering that the simulation temperature is elevated about 100 K above the T_m values. As well as being dependent on the force field used, it is worth noting that the acceptable temperature should also depend on the fold complexity of a protein and the level of detail used to define the unfolding pathway of a protein. In this study, the unfolding pathways of the WW domains are defined at the residue–residue contact level because the WW domains have a simple structure consisting of a single secondary structure element, the β strands. For other proteins, such as chymotrypsin inhibitor 2 (CI2)^{6,7} and barnase,² that have complex structures consisting of both α and β elements, the acceptable temperature may vary with different levels of detail of the unfolding pathways, e.g., whether contacts are followed between secondary structure elements or between residues. More details require a lower simulation temperature.

Our observations for simulating protein unfolding also have parallels to observations of protein folding.²⁷ Marianayagam and co-workers²⁷ found that short folding simulations can only give the correct folding mechanism when started from an equilibrated denatured state ensemble, i.e. realistic sampling of the unfolded state is necessary prior to

folding, just as sampling of the folded native state is necessary prior to unfolding in this work.

Conclusions

Standard molecular dynamics simulations of two WW domains have been carried out with an explicit solvent model starting from the native folded state at several temperatures ranging from 300 to 500 K. The simulated trajectories have durations of 12–72 ns and exhibit temperature-dependence consistent with experimental data. For the FBP28 WW domain and the given simulation model and force field, 430 K appears to be an acceptable elevated temperature to accelerate unfolding and allow observation of the unfolding pathways observed in experiments around the T_m of 337 K. The native folded state of the FBP28 WW domain was observed to be a significant local minimum only when the temperature did not exceed the acceptable elevated temperature. Based on this observation and the assumption that a lower temperature is needed to observe intermediate states, we propose a qualitative criterion for accepting an elevated temperature as that the native folded state must be sampled substantially before the unfolding starts. This can only happen when the temperature is low enough for the native state to be a local energy minimum. We further quantified this criterion and applied it to the YAP65 WW domain and found a lower acceptable elevated temperature of 415 K, which is in good agreement with its 14 K lower T_m than the FBP28 WW domain.

Acknowledgment. We gratefully acknowledge the financial support of the Klaus Tschira Foundation. We thank Maria Macias (Institut de Recerca Biomedica, Barcelona) for providing coordinates and for discussions on experiments to study the unfolding of WW domains.

References

- (1) Pande, V. S.; Rokhsar, D. S. Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G. *Proc. Natl. Acad. Sci.* **1999**, *96*, 9062–9067.
- (2) Wong, K. B.; Clarke, J.; Bond, C. J.; Neira, J. I.; Freund, S. M. et al. Towards a complete description of the structural and dynamic properties of the denatured state of barnase and the role of residual structure in folding. *J. Mol. Biol.* **2000**, *296*, 1257–1282.
- (3) Ma, B.; Nussinov, R. Molecular dynamics simulations of the unfolding of beta(2)-microglobulin. *Protein Eng.* **2003**, *16*, 561–575.
- (4) Sham, Y. Y.; Ma, B.; Tsai, C. J.; Nussinov, R. Thermal unfolding molecular dynamics simulation of Escherichia coli. *Proteins* **2002**, *46*, 308–320.
- (5) Jemth, P.; Gianni, S.; Day, R.; Li, B.; Johnson, C. M. et al. Demonstration of a low-energy on-pathway intermediate in a fast-folding. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6450–6455.
- (6) Day, R.; Bennion, B. J.; Ham, S.; Daggett, V. Increasing Temperature accelerates protein unfolding without changing the pathway of unfolding. *J. Mol. Biol.* **2002**, *322*, 189–203.
- (7) Day, R.; Daggett, V. Ensemble versus single-molecule protein unfolding. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13445–13450.

- (8) Petrovich, M.; Jonsson, A. L.; Ferguson, N.; Daggett, V.; Fersht, A. R. Phi-analysis at the experimental limits: mechanism of beta-hairpin formation. *J. Mol. Biol.* **2006**, *360*, 865–881.
- (9) Beck, D. A. C.; Daggett, V. Methods for molecular dynamics simulations of protein folding/unfolding. *Methods* **2004**, *34*, 112–120.
- (10) Chan, H. S.; Dill, K. A. Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics. *Proteins* **1998**, *30*, 2–33.
- (11) Matagne, A.; Jamin, M.; Chung, E. W.; Robinson, C. V.; Radford, S. E. et al. Thermal unfolding of an intermediate is associated with non-Arrhenius. *J. Mol. Biol.* **2000**, *297*, 193–210.
- (12) Crane, J. C.; Koepf, E. K.; Kelly, J. W.; Gruebele, M. Mapping the transition state of the WW domain beta-sheet. *J. Mol. Biol.* **2000**, *298*, 283–292.
- (13) Khan, F.; Chuang, J. I.; Gianni, S.; Fersht, A. R. The kinetic pathway of folding of barnase. *J. Mol. Biol.* **2003**, *333*, 169–186.
- (14) Nguyen, H.; Jaeger, M.; Moretto, A.; Gruebele, M.; Kelly, J. W. Tuning the free energy landscape of a WW domain by temperature, mutation, and truncation. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3948–3953.
- (15) Roe, D. R.; Hornak, V.; Simmerling, C. Folding cooperativity in a three-stranded beta-sheet model. *J. Mol. Biol.* **2005**, *352*, 370–381.
- (16) Nguyen, P. H.; Stock, G.; Mittag, E.; Hu, C. K.; Li, M. S. Free energy landscape and folding mechanism of a beta-hairpin in explicit water: A replica exchange molecular dynamics study. *Proteins* **2005**, *61*, 795–808.
- (17) Mu, Y.; Nordenskiöld, L.; Tam, J. P. Folding, misfolding, and amyloid protofibril formation of WW domain FBP28. *Biophys. J.* **2006**, *90*, 3983–3992.
- (18) Wang, T.; Wade, R. C. Force field effects on a beta-sheet protein domain structure in thermal unfolding simulations. *J. Chem. Theory Comput.* **2006**, *2*, 140–148.
- (19) Karanicolas, J.; Brooks, C. L., III. The structural basis for biphasic kinetics in the folding of the WW domain from a formin-binding protein: lessons for protein design? *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3954–3959.
- (20) Karanicolas, J.; Brooks, C. L., III. Integrating folding kinetics and protein function: biphasic kinetics and dual binding specificity in a WW domain. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 3432–3437.
- (21) Ferguson, N.; Berriman, J.; Petrovich, M.; Sharpe, T. D.; Finch, J. T. et al. Rapid amyloid fiber formation from the fast-folding WW domain FBP28. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9814–9819.
- (22) Ferguson, N.; Pires, J. R.; Toepert, F.; Johnson, C. M.; Pan, Y. P. et al. Using flexible loop mimetics to extend Phi-value analysis to secondary structure interactions. *PNAS* **2001**, *98*, 13008–13013.
- (23) Koepf, E. K.; Petrassi, H. M.; Sudol, M.; Kelly, J. W. WW: An isolated three-stranded antiparallel beta-sheet domain that unfolds and refolds reversibly; evidence for a structured hydrophobic cluster in urea and GdnHCl and a disordered thermal unfolded state. *Protein Sci.* **1999**, *8*, 841–853.
- (24) Ibragimova, G. T.; Wade, R. C. Stability of the beta-sheet of the WW domain: A molecular dynamics simulation study. *Biophys. J.* **1999**, *77*, 2191–2198.
- (25) Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R. et al. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
- (26) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999–2012.
- (27) Marianayagam, N. J.; Fawzi, N. L.; Head-Gordon, T. Protein folding by distributed computing and the denatured state ensemble. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16684–16689.

CT700063C

Solvatochromic Shifts of the $n \rightarrow \pi^*$ Transition of Acetone from Steam Vapor to Ambient Aqueous Solution: A Combined Configuration Interaction QM/MM Simulation Study Incorporating Solvent Polarization

Yen-lin Lin and Jiali Gao*

Department of Chemistry and Supercomputing Institute, Digital Technology Center, University of Minnesota, 207 Pleasant Street, SE, Minneapolis, Minnesota 55455

Received March 10, 2007

Abstract: A hybrid quantum mechanical and molecular mechanical potential is used in Monte Carlo simulations to examine the solvent effects on the electronic excitation energy of the $n \rightarrow \pi^*$ transition of acetone in ambient and supercritical water fluid, in which the temperature is in the range of 25–500 °C with pressures of 1–2763 atm. In the present study, the acetone molecule is described by the AM1 Hamiltonian, and the water molecules are treated classically. Two sets of calculations are performed. The first involves the TIP4P model for water, and the second employs a polarizable model, POL2, for the solvent. The first calculation yields the excitation energy by using the static ground-state solvent charge distribution obtained from QM-CI/MM calculations. The latter takes into account the effect of solvent polarization following the solute electronic excitation. The trend of the computed $n \rightarrow \pi^*$ blue-shifts for acetone as function of the fluid density is in good agreement with experimental results. The present simulations of acetone in the supercritical, near supercritical, dense-liquid, and ambient water fluids reveal that the solvatochromic shifts are dominated by the electrostatic interactions between acetone and water molecules during the solute excitation. Additionally, the solvent charge redistribution following the solute electronic excitation has a small correlation (0 to -37 cm^{-1}) to the total solvatochromic shift and decreases linearly with water density. Both the solvatochromic shift and solvent polarization correction are more obvious in the ambient water than in the supercritical water because the solvent stabilization of the ground state over the excited state is more significant in the former condition.

Introduction

The unusual properties of supercritical water (SCW) fluid have attracted considerable interests^{1–6} as a viable medium for green chemical oxidations.^{7–9} At or near supercritical conditions, organic species and molecular oxygen are completely miscible,^{2–7,10,11} whereas electrolytes are nearly insoluble.^{7,12} Thus, it offers a tremendous opportunity to develop alternative technologies for the destruction of chemical warfare agents and organic wastes by complete oxidation.^{7,9} An advantage of performing chemical oxidations

in supercritical fluid water ($T_c = 374 \text{ °C}$, $P_c = 217.7 \text{ atm}$, and $\rho_c = 3.22 \text{ g}\cdot\text{cm}^{-3}$) or near supercritical conditions is that the reaction conditions can be optimized by varying the density of the medium with a change in pressure. The variations of the fluid density, ranging from steam vapor to dense aqueous solution, also provide an interesting medium for investigating solute–solvent interactions.^{13–15} The present study is aimed at an understanding of the change of solute and solvent interactions over the entire spectra of solvent density and the unusual behaviors of solvation near supercritical conditions. We focus our study on the continuous change of solvatochromic shifts of the chromophore acetone

* Corresponding author e-mail: gao@chem.umn.edu.

as a probe solute in fluid water by using a combined quantum mechanical configuration interaction and molecular mechanical potential in statistical mechanical Monte Carlo simulations.

Solvatochromic shifts of organic chromophores have been used extensively as a probe to investigate solute–solvent interactions in solution.^{16–25} Based on the change in electronic absorption spectra of organic dye molecules, solvent polarity scales have been established including the popular $E_T(30)$ scale based on Reichardt's betaine dye.^{16,26,27} One group of chromophores containing carbonyl, thiocarbonyl, and azo functional groups are often used, which have characteristically weak n \rightarrow π^* absorption bands.¹⁶ Typically, a blue-shift in the absorption spectrum is observed in going from a low dielectric solvent to a more polar medium, although dispersion red-shifts are also found in nonpolar solvents such as carbon tetrachloride and hexane.^{16,28–30}

Continuum solvation models coupled with electronic structure calculations have been widely used to model solvatochromic shifts.^{18,20,31–34} The ZINDO program and its associated methods developed by Zerner have been applied to a variety of chromophores with remarkable success.¹⁸ Cossi and Barone evaluated the n \rightarrow π^* transition of acetone in various polar and nonpolar solvents using the polarizable continuum solvation model,³¹ while a number of other groups have also studied this system using different techniques.^{20,33–40} Although excellent agreement with experiment can be obtained, a shortcoming of the continuum solvation approach is a lack of treating specific hydrogen bonding interactions. Zerner showed that only when one or two explicit water molecules are included, would the computed spectral shifts for a series of pyrimidine and pyrazine compounds be in accord with experiment.¹⁸ On the other hand, combined QM/MM simulations even at the level of configuration interaction with single excitations (CIS) only can yield reasonable results.²² Of course, the latter computations are much more time-consuming as it requires configurational averaging over millions of solvent configurations. Avoiding explicit electronic structure calculations, Warshel and co-workers used the partial charges derived for the ground and excited states along with an atom-centered polarizable dipole model to determine the solvent effects on vertical excitation energy.⁴¹ This approach has been used by Blair et al.⁴² and by DeBolt and Kollman et al.⁴³ in the analysis of excited-state energy relaxation. Previously, our group described a combined QM-CI/MM approach in Monte Carlo simulations, which has been applied to a number of systems, including acetone in a variety of solvents.^{22–24,44,45} Later, the method was extended by incorporating a consistent treatment of the instantaneous electronic polarization between the solute and solvent in response to solute excitation.^{23,46} Thompson and Schenter also presented a combined QM-CI/MM-pol model that includes polarization effects in the MM region and have applied it to study both ground and excited states.^{47,48} In addition, Martin et al. presented a strategy using the mean-field approximation combining the QM/MM method to calculate the solvent shift of acetone in the ambient water.⁴⁰ A combined QM/MM strategy has also been implemented in CASSCF calculations.²⁵

Bennett and Johnston carried out a most comprehensive experimental study and measured the entire range of solvatochromic shifts of the n \rightarrow π^* absorption band of acetone in vapor, fluid, and liquid water.¹ The experimental results showed that the spectral shifts can be divided into three regions. First, there is an initial phase of rapid increase in spectral shift, relative to the excitation energy of the isolated chromophore acetone in the gas phase, in the low-density steam region. This is followed by a plateau region near supercritical fluid conditions. Finally, as the fluid density increases toward the ambient value, the absorption energy increases quickly again. The existence of a plateau region near supercritical conditions has been proposed as a feature due to solvent clustering.^{1,49,50} In a separate study, Takebayashi et al., who utilized NMR spectroscopy and Monte Carlo simulations, found similar features, which were attributed to the variations in solute–solvent hydrogen bonding as the temperature and water density changes.^{51,52} On the theoretical side, a number of molecular dynamics and Monte Carlo simulations have been reported, primarily focusing on solute–solvent interactions at or near the supercritical fluid region at a few selected states.^{15,53–55} These studies provided support to solvent clustering at supercritical conditions. Recently a classical Monte Carlo simulation of acetone in water followed by cluster calculations with semiempirical and time-dependent density functional theory has been reported at the supercritical point,³⁷ and the change of solvatochromic shifts of an organic chromophore in the entire range of solvent densities has not been demonstrated computationally.

In this work, we aim to assess the solvent effects on the n \rightarrow π^* blue-shift of acetone in the full region from steam vapor to supercritical conditions to ambient water. The computed n \rightarrow π^* solvatochromic shifts of acetone in water fluids at various temperatures and solvent densities are compared with experimental values.¹ To evaluate the contributions of different molecular interactions to the acetone n \rightarrow π^* blue-shift in these fluid states, a decomposition analysis of the energies was computed based on the method our group developed previously.^{23,46} To this end, statistical Monte Carlo simulations using a hybrid quantum-mechanical-configuration interaction and molecular mechanical (QM-CI/MM) method have been carried out to explore the solvent effects in electronic spectroscopy. The effects of the solvent polarization in response to the solute electronic excitation is evaluated by using a polarizable MM solvent model.^{23,46,56} The results of the calculations reveal the factors governing the solvatochromic shifts of acetone at different water densities and temperatures, where a polarization correlation term from the instantaneous polarization of the solvent molecules following the solute excitation was also estimated. In the following, we first present the theoretical background and computational details. This is followed by results and discussion. Finally, the main findings are summarized in the conclusion.

Methods

We use a combined quantum mechanical-configuration interaction and molecular mechanical (QM-CI/MM) potential

in statistical mechanical Monte Carlo simulations to investigate the solvatochromic shifts in the $n \rightarrow \pi^*$ transition of acetone in fluid water.^{22,23} In this hybrid system, the solute chromophore (i.e., acetone) is treated by a CI wave function, and the solvent molecules are represented classically by empirical potential functions.^{22,47} Except for a few cases, most applications of combined QM/MM potentials make use of effective pairwise potentials for the MM region, in which the partial atomic charges on the solvent atoms are fixed at the same values both for the ground and excited states of the solute. This fixed-charge approach ignores the instantaneous charge polarization of the solvent due to solute electronic excitation, i.e., the interaction of the QM and solvent-induced dipoles, and the change in solvent configurations. In the present work, we also use a polarizable solvent model for water. Thus, the mutual “QM” solute and “MM” solvent polarization interactions are explicitly treated.^{23,46–48} This is of particular interest in the present study because we examine the solvation of acetone by fluid water that covers the entire range of solvent densities, ranging from steam vapor to supercritical fluid to the dense liquid at the ambient condition. Previously we have implemented a polarizable combined QM/MM method for the study of electronic absorption in polar solvents,²³ and the approach is similar to another study described by Thompson and Schenter.^{47,48} Here, we investigate solvation effects in supercritical fluids by including the instantaneous polarization of solvent molecules in response to the solute excited-state wave function.

Energy Decomposition. The total ground-state energy of the QM/MM hybrid system with the utilization of a polarizable solvent model can be written as follows^{23,46,48}

$$E_{\text{tot}}^g = \langle \Phi_{\text{CI}}^g | \hat{H}_X^o + \hat{H}_{X_s}^{\text{stat}}(\{q_s\}) + \hat{H}_{X_s}^{\text{pol}}(\{\mu_s^g\}) | \Phi_{\text{CI}}^g \rangle + E_{X_s}^{\text{vdW}} + E_{ss}^{\text{pair}} - \frac{1}{2} \sum_s \mu_s^g \cdot F_s^o + \frac{1}{2} \sum_s \mu_s^g \cdot F_s^{\text{qm}}(\{\Phi_{\text{CI}}^g\}) \quad (1)$$

where the superscript “g” signifies quantities for the solute in the ground state, Φ_{CI}^g is the ground state CI wave function of the solute, \hat{H}_X^o is the Hamiltonian of the isolated solute (X), $\hat{H}_{X_s}^{\text{stat}}(\{q_s\})$ is the electrostatic interaction Hamiltonian between the QM system and the MM permanent charges (q_s), and $\hat{H}_{X_s}^{\text{pol}}(\{\mu_s^g\})$ is the interaction between the QM solute and the MM induced dipoles (μ_s^g) in the solute ground state. The remaining terms do not involve the electronic degrees of freedom, except the last term due to the fact that energy terms are not additive in a polarizable force field.⁴⁶ In eq 1, $E_{X_s}^{\text{vdW}}$ is the van der Waals interaction between the solute and solvent atoms, E_{ss}^{pair} is the solvent pair interaction consisting of both Lennard-Jones and Coulomb terms, F_s^o is the static electrostatic field from the MM system from its permanent charges, and $F_s^{\text{qm}}(\{\Phi_{\text{CI}}^g\})$ is the electrostatic field generated by the solute wave function. The induced dipoles of MM atoms s (μ_s^g) in eq 1 are determined self-consistently by an iterative procedure using eq 2^{46,48}

$$\mu_s^g = \alpha_s \left[F_s^o + F_s^{\text{qm}}(\{\Phi_{\text{CI}}^g\}) + \sum_{t \neq s} \nabla_s \nabla_t \left(\frac{1}{r_{st}} \right) \cdot \mu_t^g \right] \quad (2)$$

where the subscripts (s, t) refer to MM atoms, α_s is the atomic

polarizability of atom s , r_{st} is the distance between atoms s and t , and μ_t^g ($t \neq s$) are induced dipoles of all the other solvent. The value of $\{\mu_s^g\}$ is a function of the permanent charges of the MM atom s , all other solvent-induced dipoles (μ_t^g , $t \neq s$) in the MM region, and the instantaneous external field from the QM system, $F_s^{\text{qm}}(\{\Phi_{\text{CI}}^g\})$, which is derived from the molecular wave function of the solute. Since Φ_{CI}^g and $\{\mu_s^g\}$ are dependent on each other, they must be solved self-consistently. We have employed a triple-iterative procedure to achieve the convergences of both the solute wave function and the solvent induced dipole, and of the overall mutually polarized system,^{46,48} and the computational details have been described in refs 46 and 48. In short, we first use a set of induced solvent dipoles, which are kept frozen, along with the solvent permanent point charges to optimize the solute wave function. Then, the electric field of the solute molecule is included in eq 2 to optimize the solvent induced dipoles $\{\mu_s^g\}$. The new set of $\{\mu_s^g\}$ is again used to obtain an updated Φ_{CI}^g . This process continues until the total energy of the entire system in eq 1 is fully converged.

For the excited state of the solute, a similar energy expression can be obtained^{46,48}

$$E_{\text{tot}}^e = \langle \Phi_{\text{CI}}^e | \hat{H}_X^o + \hat{H}_{X_s}^{\text{stat}}(\{q_s\}) + \hat{H}_{X_s}^{\text{pol}}(\{\mu_s^e\}) | \Phi_{\text{CI}}^e \rangle + E_{X_s}^{\text{vdW}} + E_{ss}^{\text{pair}} - \frac{1}{2} \sum_s \mu_s^e \cdot F_s^o + \frac{1}{2} \sum_s \mu_s^e \cdot F_s^{\text{qm}}(\{\Phi_{\text{CI}}^e\}) \quad (3)$$

where the superscript “e” indicates excited-state quantities, Φ_{CI}^e is the excited state CI wave function of the molecules in the QM region, and μ_s^e is the induced dipole of the solvent atom s in the MM region optimized in response to the presence of the QM solute in its electronically excited state. In eq 3, the solvent polarization is assumed to be instantaneous in response to the solute electronic excitation. In general, a similar triple-iterative procedure as that for the ground state can be used, but this is very time-consuming to optimize the excited-state wave function. Fortunately, it is typically not necessary. In the present work, a simplified procedure is adopted to solve the coupled QM- and MM-SCF calculations in eq 3.^{46,48} we use the excited-state electric field of the solute, determined by the optimized ground-state reference wave function, to determine the solvent dipoles $\{\mu_s^e\}$. Thus, we do not further optimize Φ_{CI}^e . This is based on the Franck–Condon principle that the solvent and solute nuclei remain fixed in the Franck–Condon transition and the solvent’s configuration can be approximated by that in the ground state.²⁸ The small perturbation of $\{\mu_s^e\}$ by optimized Φ_{CI}^e is ignored because this is of third-order effects. Consequently, the QM/MM polarization term in eq 3 could be approximately defined as follows

$$\langle \Phi_{\text{CI}}^e | \hat{H}_{X_s}^{\text{pol}}(\{\mu_s^e\}) | \Phi_{\text{CI}}^e \rangle \approx \langle \Phi_{\text{CI}}^e | \hat{H}_{X_s}^{\text{pol}}(\{\mu_s^g\}) | \Phi_{\text{CI}}^e \rangle + \langle \Phi_{\text{CI}}^e | \hat{H}_{X_s}^{\text{pol}}(\{\Delta\nu_s\}) | \Phi_{\text{CI}}^e \rangle \quad (4)$$

where $\Delta\nu_s = \mu_s^e - \mu_s^g$. In eq 4, the last term can be expressed classically for the interaction between the solvent-induced dipole with the QM electric field.

$$\langle \Phi_{\text{CI}}^e | \hat{H}_{\text{Xs}}^{\text{pol}}(\{\Delta\nu_s\}) | \Phi_{\text{CI}}^e \rangle = - \sum_s \Delta\nu_s \cdot F_s^{\text{qm}}(\{\Phi_{\text{CI}}^e\}) \quad (5)$$

We employ eqs 4 and 5 to rewrite eq 3 as

$$E_{\text{tot}}^e = \langle \Phi_{\text{CI}}^e | \hat{H}_{\text{X}}^o + \hat{H}_{\text{Xs}}^{\text{stat}}(\{q_s\}) + \hat{H}_{\text{Xs}}^{\text{pol}}(\{\mu_s^g\}) | \Phi_{\text{CI}}^e \rangle + E_{\text{Xs}}^{\text{vdW}} + E_{\text{ss}}^{\text{pair}} - \frac{1}{2} \sum_s \mu_s^e \cdot F_s^o + \frac{1}{2} \sum_s \mu_s^e \cdot F_s^{\text{qm}}(\{\Phi_{\text{CI}}^e\}) - \sum_s \Delta\nu_s \cdot F_s^{\text{qm}}(\{\Phi_{\text{CI}}^e\}) \quad (6)$$

The difference between eqs 3 and 6 is that the former involves fully iterative QM-CI and solvent-polarization SCF calculations, whereas the latter only requires the MM-SCF iteration to obtain μ_s^e . In eq 6, excited-state energies in the CI calculations are determined by using the ground-state, solvent-induced dipoles.^{46,48} Therefore, the transition energy of the solute from the ground state to the excited state in solution can be obtained by subtracting eq 1 from eq 6

$$\Delta E_{\text{tot}}^{\text{g} \rightarrow \text{e}} = \langle \Phi_{\text{CI}}^e | \hat{H}_{\text{X}}^o + \hat{H}_{\text{Xs}}^{\text{stat}}(\{q_s\}) + \hat{H}_{\text{Xs}}^{\text{pol}}(\{\mu_s^g\}) | \Phi_{\text{CI}}^e \rangle - \langle \Phi_{\text{CI}}^g | \hat{H}_{\text{X}}^o + \hat{H}_{\text{Xs}}^{\text{stat}}(\{q_s\}) + \hat{H}_{\text{Xs}}^{\text{pol}}(\{\mu_s^g\}) | \Phi_{\text{CI}}^g \rangle - \frac{1}{2} \sum_s \Delta\nu_s \cdot F_s^o + \frac{1}{2} \sum_s [\mu_s^e \cdot F_s^{\text{qm}}(\{\Phi_{\text{CI}}^e\}) - \mu_s^g \cdot F_s^{\text{qm}}(\{\Phi_{\text{CI}}^g\})] - \sum_s \Delta\nu_s \cdot F_s^{\text{qm}}(\{\Phi_{\text{CI}}^e\}) \quad (7)$$

A further approximation of eq 7 is that we assume that the van der Waals terms for the solute in the ground state and the excited state are the same. Implicitly, we ignore the dispersion effects between solute and solvent in absorption spectral calculations.^{29,30}

Explicit Simulation Studies. The excitation energy of a chromophore in solution as defined by eq 7 can be partitioned into two components as follows^{46,57}

$$\Delta E_{\text{tot}}^{\text{g} \rightarrow \text{e}} = \Delta E_{\text{stat}}^{\text{g} \rightarrow \text{e}} + \Delta E_{\text{pol}}^{\text{g} \rightarrow \text{e}} \quad (8)$$

where

$$\Delta E_{\text{stat}}^{\text{g} \rightarrow \text{e}} = \langle \Phi_{\text{CI}}^e | \hat{H}_{\text{X}}^o + \hat{H}_{\text{Xs}}^{\text{stat}}(\{q_s\}) + \hat{H}_{\text{Xs}}^{\text{pol}}(\{\mu_s^g\}) | \Phi_{\text{CI}}^e \rangle - \langle \Phi_{\text{CI}}^g | \hat{H}_{\text{X}}^o + \hat{H}_{\text{Xs}}^{\text{stat}}(\{q_s\}) + \hat{H}_{\text{Xs}}^{\text{pol}}(\{\mu_s^g\}) | \Phi_{\text{CI}}^g \rangle \quad (9)$$

and

$$\Delta E_{\text{pol}}^{\text{g} \rightarrow \text{e}} = - \frac{1}{2} \sum_s \Delta\nu_s \cdot F_s^o + \frac{1}{2} \sum_s [\mu_s^e \cdot F_s^{\text{qm}}(\{\Phi_{\text{CI}}^e\}) - \mu_s^g \cdot F_s^{\text{qm}}(\{\Phi_{\text{CI}}^g\})] - \sum_s \Delta\nu_s \cdot F_s^{\text{qm}}(\{\Phi_{\text{CI}}^e\}) \quad (10)$$

In eq 9, $\Delta E_{\text{stat}}^{\text{g} \rightarrow \text{e}}$ represents the vertical excitation energy of the solute in the presence of the total electric field of the solvent that is equilibrated to the ground-state charge distribution of the solute. The remaining contributing terms in eq 10, $\Delta E_{\text{pol}}^{\text{g} \rightarrow \text{e}}$, indicate the correlation effects resulting from the instantaneous polarization of the solvent molecules by the solute excitation. The solvatochromic shift, $\Delta\nu$, is defined as the difference between the excitation energies of the chromophore in solution and in the gas phase

$$\Delta\nu = \langle \Delta E_{\text{tot}}^{\text{g} \rightarrow \text{e}} \rangle - \Delta E_{\text{gas}}^{\text{g} \rightarrow \text{e}} \quad (11)$$

where the bracket indicates an ensemble average over the Monte Carlo or molecular dynamics simulations. Making use of the energy partition in eq 7, we can formally separate the overall solvatochromic shift into two terms: (1) the spectral shift due to the solvent potential equilibrated to the ground state of the solute, $\langle \Delta \Delta E_{\text{stat}}^{\text{g} \rightarrow \text{e}} \rangle$ and (2) the subsequent energy change of the solvent dipole due to the solute electronic excitation. Thus,

$$\Delta\nu = \langle \Delta \Delta E_{\text{stat}}^{\text{g} \rightarrow \text{e}} \rangle + \Delta E_{\text{pol}}^{\text{g} \rightarrow \text{e}} \quad (12)$$

where

$$\langle \Delta \Delta E_{\text{stat}}^{\text{g} \rightarrow \text{e}} \rangle = \langle \Delta E_{\text{stat}}^{\text{g} \rightarrow \text{e}} \rangle - \Delta E_{\text{gas}}^{\text{g} \rightarrow \text{e}} \quad (13)$$

As described in our previous works,^{23,46} the $\langle \Delta \Delta E_{\text{stat}}^{\text{g} \rightarrow \text{e}} \rangle$ term can be further decomposed into two components

$$\langle \Delta \Delta E_{\text{stat}}^{\text{g} \rightarrow \text{e}} \rangle = \langle \Delta E_{\text{Xs}}^{\text{g} \rightarrow \text{e}} \rangle + \langle \Delta \Delta E_{\text{X}}^{\text{g} \rightarrow \text{e}} \rangle \quad (14)$$

where $\Delta E_{\text{Xs}}^{\text{g} \rightarrow \text{e}}$ describes the energy change of the solute–solvent interaction due to the solute electronic excitation, and $\Delta \Delta E_{\text{X}}^{\text{g} \rightarrow \text{e}}$ depicts the difference between the excitation energy of the solute in the gas phase ($\Delta E_{\text{X,gas}}^{\text{g} \rightarrow \text{e}}$) and that in solution ($\Delta E_{\text{X}}^{\text{g} \rightarrow \text{e}}$):

$$\langle \Delta \Delta E_{\text{X}}^{\text{g} \rightarrow \text{e}} \rangle = \langle \Delta E_{\text{X}}^{\text{g} \rightarrow \text{e}} \rangle - \Delta E_{\text{X,gas}}^{\text{g} \rightarrow \text{e}} \quad (15)$$

The energy decomposition scheme of eqs 7 and 12 provides us with a convenient, approximate procedure for estimating the instantaneously mutual polarization effects upon solute electronic excitation. First, we carry out Monte Carlo or molecular dynamics simulations using an effective, pairwise potential for the solvent such as the four-point charge TIP4P models. Since polarization effects for the ground-state configurations have been included in the potential in an average sense, on average, the computed excitation energy using such a nonpolarizable model corresponds to the energy difference of eq 7, which is written for a polarizable solvent model, averaged over the Monte Carlo trajectories. Then, we switch the solvent potential to a polarizable model and use the configurations generated in the first step that employs a nonpolarizable, effective potential to determine the ensemble average of the effects (or energy contribution) of instantaneous polarization of the solvent in response to the solute excitation. This average yields the energy terms in eq 10.

Computational Details

All QM/MM calculations in statistical mechanical Monte Carlo simulations were performed using the MCQUB/MCQUM programs,^{58,59} in which the quantum mechanical energies were calculated using the MOPAC program.⁶⁰ Monte Carlo simulations were carried out for a cubic box containing 396 water molecules and one acetone molecule with periodic boundary conditions. The isothermal isobaric (NPT) ensemble was employed at temperatures of 25, 50, 100, 200, 300, 400, 450, and 500 °C and pressures in the range of 1–2763 atm. These results in bulk conditions of a

reduced density (ρ_r) range from 0.05 to 3.10. A total of 29 unique conditions were included with various temperature and pressure conditions. The size has been shown to sufficiently describe thermodynamic and spectroscopic properties of solutes in SCW and the ambient water, especially at high-temperature regions where the fluid density is low.^{13,15,51–55} The intermolecular interaction among water molecules was spherically truncated at 9 Å. The spherical cutoff distances of the solvent–solute interaction employed in these calculations were about one-half of the edge of each unit box, ranging from 10.07 to 41.94 Å. This is reasonable since the solute molecule is not charged or having significant charge separations. Nevertheless, it might be advisable to include long-range electrostatic effects in these simulations since the ability of solvent dielectric screening effects may be different in such a large density range. In all QM-CI/MM calculations, the acetone structure was held rigid at the AM1 geometry optimized in the gas phase.⁶¹ The electronic excited-state calculations were performed by configuration interaction that includes a total of 100 configurations from an active space of 6 electrons in 5 orbitals, and these combinations have been shown to yield excellent results for acetone even though the model was not originally developed for spectroscopy.^{22,30} The van der Waals parameters for the QM atoms were determined in a previous study.⁶²

Two separate calculations were executed. First, the combined QM-CI/MM potential with the pairwise four-point charge TIP4P water model⁶³ was utilized to yield the average values for $\langle \Delta E_{\text{stat}}^{\text{g} \rightarrow \text{e}} \rangle$. The TIP4P model has been verified to adequately describe the properties of SCW for the present purposes.^{15,54} In particular, a series of Monte Carlo simulations has been carried out at 400 °C and pressures ranging from 350 to 2000 atm. By analysis of reduced parameters, it was suggested that the TIP4P model may slightly underestimate the supercritical temperature by 30 to 50 degrees.¹⁵ In the second set of QM-CI/MM simulations, the polarizable POL2 model⁵⁶ was adopted for the MM solvent to give the polarization correlation energy, $\Delta E_{\text{pol}}^{\text{g} \rightarrow \text{e}}$. In this step, only the single-point energies were evaluated based on the configurations generated in the first set of simulation. Each Monte Carlo simulation in the first computational step involves at least 4×10^6 configurations of equilibration, followed by 4×10^6 configurations for data averaging. The Owicki-Scheraga preferential sampling technique was used to enhance the statistics near the solute, such that solvent moves are made proportional to $1/(R^2 + W)$, where $W = 350$ Å.⁶⁴ The averages for $\Delta E_{\text{pol}}^{\text{g} \rightarrow \text{e}}$ in the latter calculations were equilibrated for at least 4×10^6 configurations, followed by single-point energy evaluations with a total of 50 structures to obtain the instantaneous polarization response by the solvent. Note that all spectral shifts correspond to Franck–Condon excitation, in which solute and solvent electronic polarization are assumed to be instantaneous in the excited state at the solvent nucleus positions equilibrated to the ground-state electronic structures.

Results and Discussion

Solvatochromic Shifts. The total solvatochromic shift ($\Delta\nu$) for $n \rightarrow \pi^*$ excitation of acetone in water fluid calculated

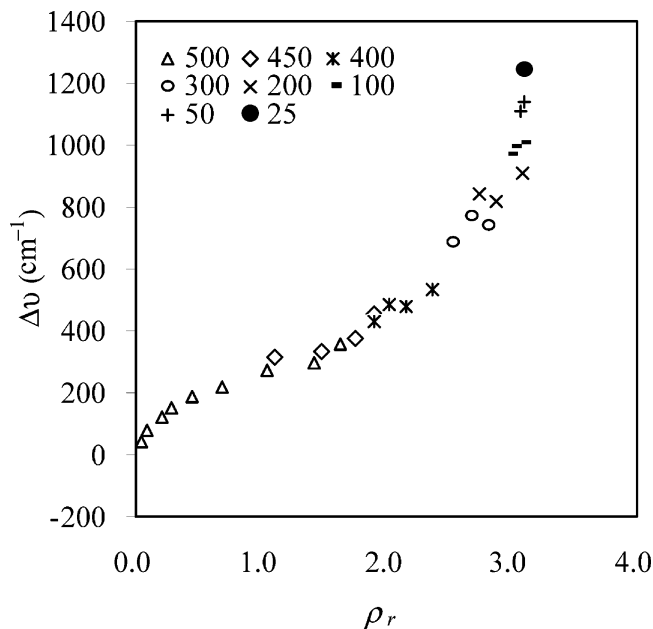


Figure 1. Computed solvatochromic shift ($\Delta\nu$) in the $n \rightarrow \pi^*$ excitation of acetone in water fluid as a function of reduced density (ρ_r). Temperatures in degrees Celsius used for different simulations are given in the upper left-hand corner.

by the QM-CI/MM method is plotted as a function of reduced density (ρ_r) of the fluid in the range from 0.02 to 3.11 (Figure 1). The reduced density of 3.11 corresponds to simulations at 25 °C and 1 atm. The theoretical results show that the initial increase in the reduced density (ρ_r) from 0.02 to 0.7 is accompanied by a rapidly rising blue-shift in $\Delta\nu$. This is followed by a slowly rising plateau region in the reduced density range of 0.7–1.5. In the third stage, the increase of $\Delta\nu$ becomes markedly steeper at higher reduced densities from 1.5 to 3.1. The trend of $\Delta\nu$ obtained in the calculations is in excellent agreement with the experimental data reported by Bennett and Johnston.¹ The three distinctive regions in Figure 1 can be categorized as (1) the *gaseous steam phase* corresponding to a temperature of 500 K and pressures of 49 → 454 atm used in the Monte Carlo simulation, (2) the *supercritical fluid region* ($T = 400$ – 500 K and $P = 454$ – 987 atm), and (3) the *dense-liquid phase* ($T = 25$ – 400 K and $P = 1$ – 2763 atm). The plateau in the UV-absorption energy in the SCW region has been attributed to the effect of solvent clustering near the solute, which plays an important role in determining the chemical reactivity of organic solute in SCW.^{65–67} The experimental observation is nicely reproduced here,¹ and we shall present structural analysis in the following section.

The quality of the present study is best illustrated by the computed $n \rightarrow \pi^*$ spectral shift ($\Delta\nu$) for acetone in ambient water, which is 1245 cm^{-1} . For comparison, the experimental value is 1560 cm^{-1} ,^{28,68} and the difference corresponds to an energy difference of only 0.9 kcal/mol. In an early study, the computed spectral shift is somewhat greater at 1690 cm^{-1} ,²² the slight difference between with the present simulations may be a reflection of the difference in size and length of different simulations, but the trends within each individual set of calculations should be reasonable. Solvatochromic shifts of acetone in various solvents have been

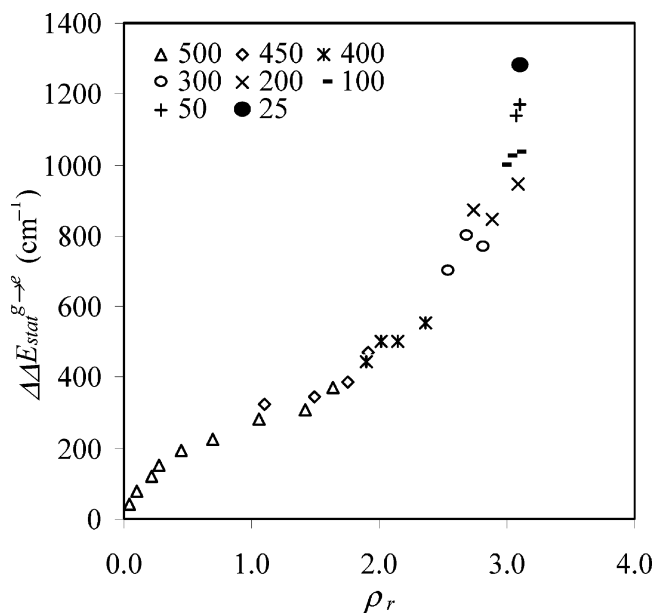


Figure 2. Electrostatic stabilization energy ($\Delta\Delta E_{\text{stat}}^{g \rightarrow e}$) due to ground-state solvent charge distribution for the $n \rightarrow \pi^*$ excitation of acetone in water fluid as a function of fluid reduced density (ρ_r). Temperatures in degrees Celsius used for different simulations are given in the upper left-hand corner.

extensively investigated in ambient conditions.^{22,30,32,39,48,69–72} It provides a prototypical system for studying solvation effects. The excellent agreement between the theoretical results and the experimental data indicates that the Monte Carlo simulations combined with the AM1 Hamiltonian for the QM atoms employed in the present work are adequate for analyses of the solvation structure and solvation energies of an organic solute in water fluids spanning the entire density ranges from vapor to supercritical fluids to dense liquid.

Energy Decompositions. To gain insight into the origin of the observed absorption spectral shifts and the possibility of solvent clustering near supercritical fluid conditions, we decomposed the total solvatochromic shifts into specific terms. If there is stable solvent cluster formation near the supercritical point, one would expect to find a relatively large and invariant condition from the $\Delta E_{\text{pol}}^{g \rightarrow e}$ term because the solvent polarization effects depend on the size of the cluster. To conveniently analyze the solvatochromic shift ($\Delta\nu$) of the acetone excitation in water fluid, the water reduced density (ρ_r) in this work is divided into four regions: the vapor phase ($\rho_r < 0.7$), the supercritical and near supercritical fluid region ($0.7 < \rho_r < 1.9$), the dense-liquid region ($\rho_r > 1.9$), and the ambient water state ($\rho_r = 3.11$).¹ In these four regions, the values of $\Delta\nu$ obtained from the hybrid QM-CI/MM calculations are in the range of 48–219 cm^{-1} , 273–452 cm^{-1} , 485–1142 cm^{-1} , and 1245 cm^{-1} , respectively (Figure 1).

$\Delta\nu$ can be decomposed into $\Delta\Delta E_{\text{stat}}^{g \rightarrow e}$ and $\Delta E_{\text{pol}}^{g \rightarrow e}$ terms (eq 11). The first term represents the electrostatic stabilization of the ground state over the excited state due to solvation, of which the excitation energy in solution is obtained using the solvent charge distribution in the ground state of the solute. The second term is the polarization correlation energy due to the instantaneous solvent polarization following the

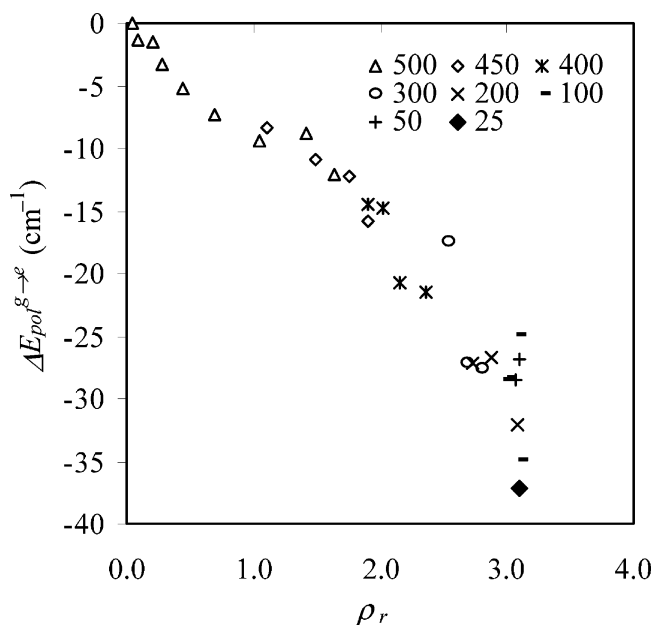


Figure 3. Computed solvent polarization contributions ($\Delta E_{\text{pol}}^{g \rightarrow e}$) to the overall spectral shifts for the $n \rightarrow \pi^*$ excitation of acetone in water fluid as a function of fluid reduced density (ρ_r). Temperatures in degrees Celsius used for different simulations are given.

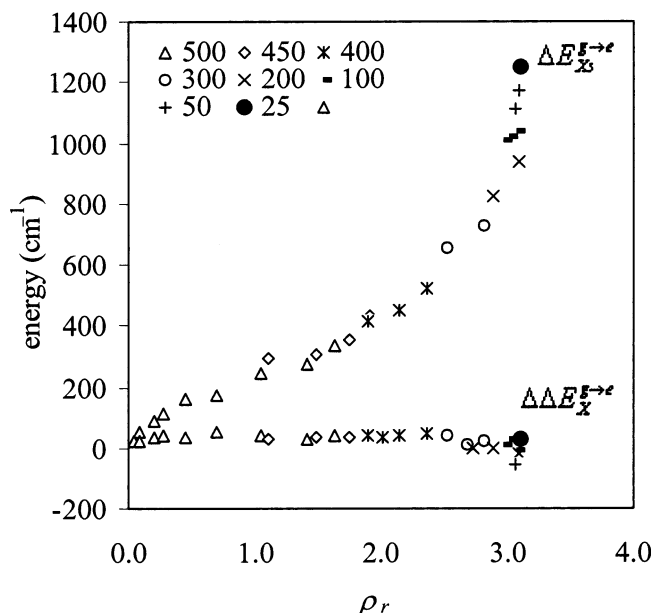


Figure 4. Decomposition of the ground-state electrostatic energy term in Figure 2 into the change in net solute–solvent interaction energy ($\Delta\Delta E_{X_s}^{g \rightarrow e}$) and the intrinsic excitation energy of the solute ($\Delta\Delta E_X^{g \rightarrow e}$) for the $n \rightarrow \pi^*$ transition of acetone in water fluid. Temperatures in degrees Celsius used for different simulations are indicated.

solute electronic excitation. The decomposition results of $\Delta\nu$ show that in the vapor, supercritical and near supercritical fluid, dense-liquid, and the ambient water regions, $\Delta\Delta E_{\text{stat}}^{g \rightarrow e}$ contributes to the blue-shifts $\Delta\nu$ by 40–226 cm^{-1} , 282–468 cm^{-1} , 500–1169 cm^{-1} , and 1282 cm^{-1} , respectively (Figure 2). On the other hand, $\Delta E_{\text{pol}}^{g \rightarrow e}$ contributes a small red-shift to $\Delta\nu$ in ranges of -0.02 to -7 cm^{-1} , -8 to -16 cm^{-1} , -15 to -35 cm^{-1} , and -37 cm^{-1} , respectively (Figure

Table 1. Theoretical Results of Ground-State Dipole Moment in Water Fluid ($\langle\mu^g\rangle$),^a Ground-State Induced Dipole Moment ($\Delta\mu_{\text{ind}}^g$),^b Excited-State Dipole Moment in Water Fluid ($\langle\mu^e\rangle$), and Excited-State Induced Dipole Moment ($\Delta\mu_{\text{ind}}^e$)^c for Acetone in the Supercritical ($\rho_r < 0.7$), Near-Critical ($0.7 < \rho_r < 1.9$) and Dense-Liquid Regions ($\rho_r > 1.9$) and Ambient Water ($\rho_r = 3.1$)

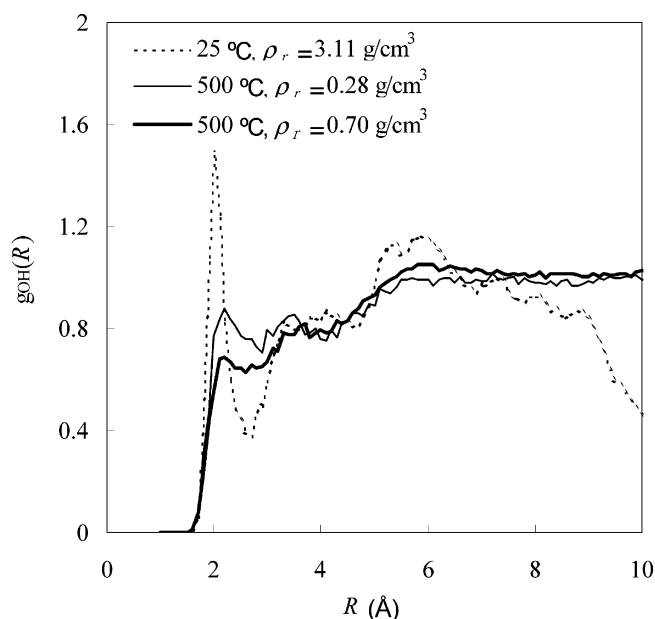
	$\langle\mu^g\rangle$ (D)	$\Delta\mu_{\text{ind}}^g$ (D)	$\langle\mu^e\rangle$ (D)	$\Delta\mu_{\text{ind}}^e$ (D)
supercritical region	2.94–3.17	0.02–0.25	2.89–3.06	0.02–0.19
near-critical region	3.21–3.41	0.29–0.49	3.15–3.36	0.28–0.49
dense-liquid region	3.42–3.98	0.51–1.06	3.34–3.95	0.47–1.08
ambient water	4.07	1.15	3.87	1.00

^a Ensemble average of AM1 ground-state dipole moment in water fluid. ^b $\Delta\mu_{\text{ind}}^g = \langle\mu^g\rangle - \mu_{\text{gas}}^g$, where μ_{gas}^g , a value of 2.91 D, is the ground-state dipole moment of acetone using the optimized AM1 geometry in the gas phase. ^c $\Delta\mu_{\text{ind}}^e = \langle\mu^e\rangle - \mu_{\text{gas}}^e$, where μ_{gas}^e , a value of 2.87 D, is the excited-state dipole moment of acetone in the gas phase.

3). Clearly, inclusion of the solvent instantaneous polarization effects leads to stabilization of the electronic excited state, giving rise to a red-shift in the absorption energy, and the effect increases as the solvent density increases. However, it only makes a small correction to the total solvatochromic shift, suggesting that the energy input required to reorient solvent dipoles following the solute excitation is small.

To further understand the solute–solvent interactions and the solute intrinsic energy contributing to the electrostatic stabilization over the excitation due to solvation ($\Delta\Delta E_{\text{stat}}^{\text{g}\rightarrow\text{e}}$), the $\Delta\Delta E_{\text{stat}}^{\text{g}\rightarrow\text{e}}$ term is further separated into $\Delta E_{\text{Xs}}^{\text{g}\rightarrow\text{e}}$ and $\Delta\Delta E_{\text{Xs}}^{\text{g}\rightarrow\text{e}}$, two terms using eq 14. The $\Delta E_{\text{Xs}}^{\text{g}\rightarrow\text{e}}$ represents the energy change of the solute–solvent interaction due to different solute charge distributions in the ground state and excited state, and $\Delta\Delta E_{\text{Xs}}^{\text{g}\rightarrow\text{e}}$ is the change of the intrinsic excitation energy of the solute in solution. In the vapor, supercritical and near supercritical, dense-liquid and ambient states, the energy component of $\Delta E_{\text{Xs}}^{\text{g}\rightarrow\text{e}}$ is the dominant component of the total solvatochromic shift, and the computed values are 23–167 cm^{-1} , 256–394 cm^{-1} , 484–1110 cm^{-1} , and 1253 cm^{-1} , respectively (Figure 4). Together with the solvent polarization correction, $\Delta E_{\text{pol}}^{\text{g}\rightarrow\text{e}}$, we find that the observed spectral shifts nearly come entirely from the difference in solute–solvent interaction energy between the excited and the ground states, comprising 95% of the total $\Delta\nu$. Surprisingly, the intrinsic excitation energy of the solute does not change significantly relative to the gas-phase value, with the computed $\Delta\Delta E_{\text{Xs}}^{\text{g}\rightarrow\text{e}}$ less than 50 cm^{-1} in all density ranges (Figure 4). Evidently, the polarization of the solute wave function does not affect the energy gap between the ground state and the $n \rightarrow \pi^*$ excited state. The results of the energy decompositions for the solvatochromic shift of the acetone $n \rightarrow \pi^*$ excitation reveal that the electrostatic stabilization from the solute–solvent interaction in the ground state over that in the excited state ($\Delta E_{\text{Xs}}^{\text{g}\rightarrow\text{e}}$) primarily dominates the spectra blue-shift. Furthermore, $\Delta E_{\text{Xs}}^{\text{g}\rightarrow\text{e}}$ increases continuously without a plateau behavior in SCW region, although it shows a clear transition in that region, leading to a rapid increase as the solvent density further increases (Figure 4). It is interesting to comment that for systems involving $\pi \rightarrow \pi^*$ transitions where dispersion and inductive polarization effects might be more significant, the quantitative picture could have greater polarization and intrinsic energy contributions. It would be interesting to make similar analyses of these types of compounds.

Dipole Moment vs Spectra Shift. A further measure of the molecular polarization is provided by calculating the

**Figure 5.** Computed radial-distribution functions for the acetone oxygen and water hydrogen ($g_{\text{OH}}(R)$) in ambient water (dashed line) and in the supercritical water states of 500 °C (solid lines).**Table 2.** Computed Positions of the First Peaks (r_1) in Radial Distribution Functions (rdfs) and Coordination Numbers of Water Molecules in the First Solvation Shell of Acetone ($N_{\text{H}_2\text{O}}$) in the Supercritical, Near-Critical, Dense-Liquid, and Ambient Water Fluids

	r_1 (Å)	$N_{\text{H}_2\text{O}}$
supercritical region	2.1–2.3	0.08–0.72
near-critical region	2.1–2.3	0.61–1.44
dense-liquid region	2.0–2.2	1.22–2.51
ambient water	2.0	2.90

ground-state and excited-state induced dipole moments of acetone due to solvation (Table 1). The calculated average and induced dipole moments continuously increase as the fluid density increases, whereas the values of $\langle\mu^g\rangle - \langle\mu^e\rangle$ are relatively small. Thus, although the ground state is more strongly solvated than the excited state, the similarity in the computed dipoles in Table 1 show that specific solute–solvent interactions are critically important in molecular solvation, and the overall dipole moment of a molecule is not a direct indication of its strength of solvation.

Solvent Structure vs Spectral Shift. The structural interpretation of energy component analyses is confirmed

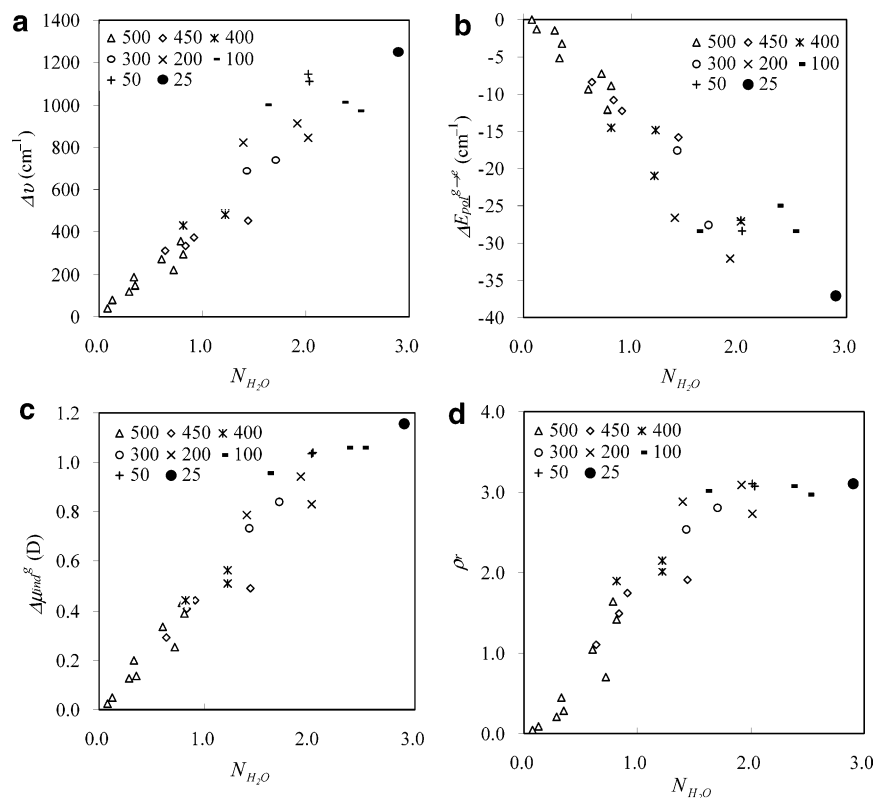


Figure 6. Correlations with the computed coordination number of water molecules in the first-solvation layer of acetone ($N_{\text{H}_2\text{O}}$) for (a) solvatochromic shift ($\Delta\nu$), (b) solvent polarization contribution ($\Delta E_{\text{pol}}^{\text{g}\rightarrow\text{e}}$), (c) ground-state induced dipole moment of acetone ($\Delta\mu_{\text{ind}}^{\text{g}}$), and (d) reduced density (ρ_r) of fluid. Temperatures in degrees Celsius used for different simulations are indicated.

by examining the radial distribution functions (rdfs) between the solute and solvent. In particular, we focus on the acetone oxygen (O) and the water hydrogen (Hw) rdf, $g_{\text{OH}}(R)$, which gives the probability of finding a water hydrogen atom (Hw) at a distance R from the acetone oxygen (O). Figure 5 shows the rdfs obtained in the ambient water ($T = 25$ °C, $\rho_r = 3.11$) and a state just above the supercritical conditions of 500 °C at a reduced density of 0.70. Table 2 presents the positions of the first peaks of the rdfs and the coordination number of water molecules in the first solvation layer for acetone in these four water conditions. In the ambient water, the position of the first solvation peak is well-defined appearing at 2.0 Å, but the second peak is less structured. The calculated solvent structure in the ambient water is similar to that observed by Thompson,⁴⁸ by Takebayashi et al.,^{51,52} and by Martin et al.⁴⁰ In contrast, there is no well-defined first peak of $g_{\text{OH}}(R)$ in the supercritical conditions (Figure 5), which is also in accord with the results obtained by Takebayashi et al.^{51,52}

The positions of the first peak of $g_{\text{OH}}(R)$ in the vapor, supercritical and near supercritical, and dense-liquid water fluids were shifted to longer distances in comparison with that in ambient water by 0.1–0.3 Å, 0.1–0.3 Å, and 0.0–0.2 Å, respectively (Table 2). Furthermore, the coordination numbers in the first solvation layer about the oxygen of acetone ($N_{\text{H}_2\text{O}}$) are 0.08–0.72, 0.61–1.44, and 1.22–2.51 in the corresponding conditions (Table 2 and Figure 6d). These results show that the average number of $N_{\text{H}_2\text{O}}$ is an increasing function of water density, which is consistent with

the finding by Takebayashi et al.^{51,52} It is worthy to note that the trend of the coordination numbers of water ($N_{\text{H}_2\text{O}}$) in the first layer around acetone reflects the total solvatochromic shifts ($\Delta\nu$ in the $n \rightarrow \pi^*$ excitation of acetone with a linear correlation of $r^2 = 0.91$ (Figure 6a). The solvent polarization correction ($\Delta E_{\text{pol}}^{\text{g}\rightarrow\text{e}}$) and the ground-state induced dipole ($\Delta\mu_{\text{ind}}^{\text{g}}$) also show linear correlations with $N_{\text{H}_2\text{O}}$ (Figure 6b,c). Overall, it implies that the density-dependent $N_{\text{H}_2\text{O}}$ of specific hydrogen-bond interactions between acetone and water molecules directly influences the magnitude of the solvatochromic shift, solvent polarization correction, and the induced dipole of acetone. In fact, the changes in coordination number shown in Figure 6d as a function of the fluid reduced density mirrors completely with the trends of the spectral shifts in Figure 1.

Conclusions

Hybrid QM-CI/MM Monte Carlo simulations have been carried out to investigate the solvatochromic shifts of the acetone $n \rightarrow \pi^*$ excitation in the supercritical ($\rho_r < 0.7$), near-critical ($0.7 < \rho_r < 1.9$), dense-liquid ($\rho_r > 1.9$), and ambient water conditions. In the present work, the solvent polarization correlation following the solute electronic excitation was included. The computed $n \rightarrow \pi^*$ blue-shift in ambient water (1245 cm^{-1}) is in reasonable agreement with the experimental value ($\Delta\nu_{\text{exp}} = 1560 \text{ cm}^{-1}$).^{28,68} The trend of the solvatochromic shift as a function of reduced fluid density with the range from 0.05 to 3.11 was in accord with the experiment probed by the UV–visible absorption spec-

troscopy.¹ The results of energy decomposition show that the solvatochromic shifts in the supercritical, near-critical, dense-liquid, and ambient water fluids are mainly determined by the electrostatic interactions between acetone and water molecules during the solute excitation. Furthermore, the energy required to orient solvent molecules following the acetone excitation is quite small and decreases linearly with water density. The solvent-density dependent blue-shift and the solvent polarization correction for the acetone $n \rightarrow \pi^*$ excitation in water fluid are governed by the induced dipole of acetone in the ground and excited states and the specific hydrogen-bond interactions between the oxygen of acetone and the hydrogen of water. In addition, both energy terms are more obvious in the ambient water than in the supercritical water because the solvent stabilization of the ground state over the excited state is more significant in the former condition.

Acknowledgment. This work has been supported in part by the National Institutes of Health (GM46736) and by the Army Research Laboratory through the Army High-Performance Computing Research Center (AHPARC) under the auspices of Army Research Laboratory DAAD 19-01-2-0014 and by the Office of Naval Research under grant number N00014-05-1-0538.

References

- (1) Bennett, G. E.; Johnston, K. P. *J. Phys. Chem.* **1994**, *98*, 441.
- (2) Bermejo, M. D.; Cocero, M. J. *AIChE J.* **2006**, *52*, 3933.
- (3) Marrone, P. A.; Hodes, M.; Smith, K. A.; Tester, J. W. *J. Supercrit. Fluids* **2004**, *29*, 289.
- (4) Marrone, P. A.; Cantwell, S. D.; Dalton, D. W. *Ind. Eng. Chem. Res.* **2005**, *44*, 9030.
- (5) Ryan, E. T.; Xiang, T.; Johnston, K. P.; Fox, M. A. *J. Phys. Chem. A* **1997**, *101*, 1827.
- (6) Williams, P. T.; Onwudili, J. A. *Environ. Technol.* **2006**, *27*, 823.
- (7) Shaw, R. W.; Brill, T. B.; Clifford, A. A.; Eckert, C. A.; Franck, E. U. *Chem. Eng. News* **1991**, *69* (51), 26.
- (8) Savage, P. E. *Chem. Rev.* **1999**, *99*, 603.
- (9) Minett, S.; Fenwick, K. *Eur. Water Manage.* **2001**, *4*, 54.
- (10) Haar, L.; Gallagher, J. S.; Kell, G. S. *NBSATRC Steam Tables*; Hemisphere: Washington, DC, 1984.
- (11) Archer, D. G.; Wang, P. *J. Phys. Chem. Reg. Data* **1990**, *19*, 371.
- (12) Paulaitis, M. E.; Krukonis, V. J.; Kurnik, R. T.; Reid, R. C. *Rev. Chem. Eng.* **1983**, *1*, 179.
- (13) Johnston, K. P.; Rossky, P. J. *NATO Sci. Ser. Ser. E* **2000**, *366*, 323.
- (14) Galkin, A. A.; Lunin, V. V. *Russ. Chem. Rev.* **2005**, *74*, 21.
- (15) Gao, J. *J. Am. Chem. Soc.* **1993**, *115*, 6893.
- (16) Reichardt, C. *Solvents and Solvent Effects in Organic Chemistry*, 2nd ed.; VCH: Weinheim, 1990.
- (17) Amos, A. T.; Hall, G. G. *Proc. R. Soc. London, Ser. A* **1961**, *263*, 482.
- (18) (a) Karelson, M. M.; Katritzky, A. R.; Zerner, M. C. *Int. J. Quantum Chem., Quantum Chem. Symp.* **1986**, *20*, 521. (b) Karelson, M. M.; Zerner, M. C. *J. Am. Chem. Soc.* **1990**, *112*, 9405. (c) Karelson, M. M.; Zerner, M. C. *J. Phys. Chem.* **1992**, *96*, 6949.
- (19) Lerf, C.; Suppan, P. *J. Chem. Soc. Faraday Trans.* **1992**, *88*, 963.
- (20) Li, J.; Cramer, C. J.; Truhlar, D. G. *Int. J. Quantum Chem.* **2000**, *77*, 264.
- (21) Nugent, S.; Ladanyi, B. M. *J. Chem. Phys.* **2004**, *120*, 874.
- (22) Gao, J. *J. Am. Chem. Soc.* **1994**, *116*, 9324.
- (23) Gao, J.; Byun, K. *Theor. Chem. Acc.* **1997**, *96*, 151.
- (24) Rajamani, R.; Gao, J. *J. Comput. Chem.* **2002**, *23*, 96.
- (25) Poulsen, T. D.; Ogilby, P. R.; Mikkelsen, K. V. *J. Chem. Phys.* **2002**, *116*, 3730.
- (26) Buncel, E.; Rajagopal, S. *Acc. Chem. Res.* **1990**, *23*, 226.
- (27) Reichardt, C. *Chem. Rev.* **1994**, *94*, 2319.
- (28) Bayliss, N. S.; McRae, E. G. *J. Phys. Chem.* **1954**, *58*, 1006.
- (29) Canuto, S.; Coutinho, K.; Zerner, M. C. *J. Chem. Phys.* **2000**, *112*, 7293.
- (30) Roesch, N.; Zerner, M. C. *J. Phys. Chem.* **1994**, *98*, 5817.
- (31) Cossi, M.; Barone, V. *J. Chem. Phys.* **2000**, *112*, 2427.
- (32) Pappalardo, R. R.; Reguero, M.; Robb, M. A.; Frish, M. *Chem. Phys. Lett.* **1993**, *212*, 12.
- (33) Minezawa, N.; Kato, S. *J. Chem. Phys.* **2007**, *126*, 054511/1.
- (34) Mennucci, B.; Cammi, R.; Tomasi, J. *J. Chem. Phys.* **1998**, *109*, 2798.
- (35) Martin, M. E.; Sanchez, M. L.; Olivares del Valle, F. J.; Aguilar, M. A. *J. Chem. Phys.* **2000**, *113*, 6308.
- (36) Bernasconi, L.; Sprik, M.; Hutter, J. *J. Chem. Phys.* **2003**, *119*, 12417.
- (37) Fonseca, T. L.; Coutinho, K.; Canuto, S. *J. Chem. Phys.* **2007**, *126*, 034508.
- (38) Neugebauer, J.; Louwarse, M. J.; Baerends, E. J.; Wesolowski, T. A. *J. Chem. Phys.* **2005**, *122*, 094115.
- (39) Coutinho, K.; Canuto, S. *THEOCHEM* **2003**, *632*, 235.
- (40) Martin, M. E.; Sanchez, M. L.; Olivares del Valle, F. J.; Aguilar, M. A. *J. Chem. Phys.* **2000**, *113*, 6308.
- (41) Luzhkov, V.; Warshel, A. *J. Am. Chem. Soc.* **1991**, *113*, 4491.
- (42) Blair, J. T.; Krogh-Jespersen, K.; Levy, R. M. *J. Am. Chem. Soc.* **1989**, *111*, 6948.
- (43) DeBolt, S. E.; Kollman, P. A. *J. Am. Chem. Soc.* **1990**, *112*, 7515.
- (44) Gao, J.; Li, N.; Freindorf, M. *J. Am. Chem. Soc.* **1996**, *118*, 4912.
- (45) Gao, J.; Alhambra, C. *J. Am. Chem. Soc.* **1997**, *119*, 2962.
- (46) Gao, J. *J. Comput. Chem.* **1997**, *18*, 1061.
- (47) Thompson, M. A.; Schenter, G. K. *J. Phys. Chem.* **1995**, *99*, 6374.
- (48) Thompson, M. A. *J. Phys. Chem.* **1996**, *100*, 14492.
- (49) Heitz, M. P.; Bright, F. V. *J. Phys. Chem.* **1996**, *100*, 6889.

- (50) Wyatt, V. T.; Bush, D.; Lu, J.; Hallett, J. P.; Liotta, C. L.; Eckert, C. A. *J. Supercrit. Fluids* **2005**, *36*, 16.
- (51) Takebayashi, Y.; Sugeta, S. Y., T.; Otake, K.; Nakahara, M. *J. Phys. Chem. B* **2003**, *107*, 9847.
- (52) Takebayashi, Y.; Yoda, S. S., T.; Otake, K.; Sako, T.; Nakahara, M. *J. Chem. Phys.* **2004**, *120*, 6100.
- (53) Balbuena, P. B.; Johnston, K. P.; Rossky, P. J. *J. Phys. Chem.* **1996**, *100*, 2716.
- (54) Gao, J. *J. Phys. Chem.* **1994**, *98*, 6049.
- (55) Kubo, M.; Levy, R. M.; Rossky, P. J.; Matubayasi, N.; Nakahara, M. *J. Phys. Chem. B* **2002**, *106*, 3979.
- (56) Dang, L. X. *J. Chem. Phys.* **1992**, *97*, 2659.
- (57) Gao, J. *J. Phys. Chem.* **1992**, *96*, 537.
- (58) Gao, J. *MCQUB; v3.0*; Department of Chemistry, SUNY: Buffalo, NY, 1998.
- (59) Gao, J. *MCQUM; v4.0*; Department of Chemistry, University of Minnesota: Minneapolis, MN, 2000.
- (60) Stewart, J. J. P. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1.
- (61) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- (62) Gao, J.; Xia, X. *Science* **1992**, *258*, 631.
- (63) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (64) Owicki, J. C.; Scheraga, H. A. *Chem. Phys. Lett.* **1977**, *47*, 600.
- (65) Westacott, R. E.; Johnston, K. P.; Rossky, P. J. *J. Am. Chem. Soc.* **2001**, *123*, 1006.
- (66) Westacott, R. E.; Johnston, K. P.; Rossky, P. J. *J. Phys. Chem. B* **2001**, *105*, 6611.
- (67) Balbuena, P. B.; Johnston, K. P.; Rossky, P. J.; Hyun, J.-K. *J. Phys. Chem. B* **1998**, *102*, 3806.
- (68) Suppan, P. J. *Photochem. Photobiol.* **1990**, *A50*, 293.
- (69) De Vries, A. H.; Van, Duijnen, P. T. *Int. J. Quantum Chem.* **1996**, *57*, 1067.
- (70) Fox, T.; Roesch, N. *Chem. Phys. Lett.* **1992**, *191*, 33.
- (71) Georg, H. C.; Coutinho, K.; Canuto, S. *Chem. Phys. Lett.* **2006**, *429*, 119.
- (72) Roehrig, U. F.; Frank, I.; Hutter, J.; Laio, A.; VandeVondele, J.; Rothlisberger, U. *ChemPhysChem* **2003**, *4*, 1177.

CT700058C

Solid–Liquid Interfacial Free Energy of Water: A Molecular Dynamics Simulation Study

Jun Wang, Yuk Wai Tang, and X. C. Zeng*

Department of Chemistry, University of Nebraska–Lincoln, Lincoln, Nebraska 68588

Received November 27, 2006

Abstract: The superheating-undercooling hysteresis method and molecular dynamics simulation [Luo et al. *Phys. Rev. B* 2003, 68, 134206] were applied to estimate solid–liquid interfacial free energy (γ) of model water at ambient pressure. Two models of water were selected, the TIP4P-Ew and TIP5P-Ew, which are the improved TIP4P and TIP5P model (for the use with Ewald technique), respectively. The calculated γ at 1 bar is 37 mJ/m² for TIP4P-Ew and 42 mJ/m² for TIP5P-Ew, consistent with a previous direct MD simulation (39 mJ/m²), as well as within the range of measured values (25–44 mJ/m²).

Introduction

The free energy of the interface (γ) at a given pressure is one of the fundamental thermodynamic properties of interfacial systems. For example, the liquid–vapor interfacial free energy (or the surface tension) is relevant to capillary rise, and the solid–liquid interfacial free energy plays an important role in understanding the mechanism of nucleation and crystal growth. Despite its key role in interfacial systems, γ is difficult to measure experimentally. In most cases γ can be measured either indirectly from measurements of crystal nucleation rates or directly by contact angle measurements.¹ The former method is limited by the fact that nucleation primarily occurs heterogeneously, while the latter method has been used to study only a few materials to date due to the difficulty of such experiments.

Theoretically, density-functional theory has been a primary choice to evaluate γ . However, previous studies have been primarily focused on simple model systems (hard-sphere and Lennard-Jones models), and calculations of solid–liquid interfacial free energies are not fully consistent in the literature.^{2–4} Accurate γ can also be obtained through atomistic simulations such as using molecular dynamics (MD). To calculate liquid–vapor surface tension, four types of MD simulation techniques can be selected, including the Kirkwood-Buff mechanical relation, thermodynamic free energy difference, finite-size scaling, and thermodynamic free-energy perturbation.⁵ In the case of solid–liquid interface, however, the mechanical relation method only gives

the excess surface stress, rather than the interfacial free energy γ . Two simulation methods have been developed to compute solid–liquid interfacial free energy γ , namely the fluctuation method and the cleaving potential method. The fluctuation method^{6–9} examines the fluctuations in the height of the interface followed by a Fourier transform to compute the interfacial stiffness which can be fitted to obtain γ . The fluctuation method is able to distinguish weak anisotropy of a system since the anisotropy of the stiffness is an order of magnitude larger than that of the free energy but is less accurate in determining γ due to the fitting process involved. The method cannot be used to resolve faceted interfaces because the fluctuation of interface height is too small. Broughton and Gilmer¹⁰ proposed the cleaving potential method which consists of four reversible steps: cleaving solid phase, cleaving liquid phase, merging solid and liquid interfaces, and removing the fictitious cleaving potential. The total work computed through thermodynamic integration in the four steps is directly related to γ . Davidchack and Laird^{11,12} later proposed to use cleaving walls instead of cleaving potential, which resulted in accuracy sufficient to resolve the anisotropy of interfacial free energy. More recently, Mu and Song¹³ further improved the efficiency of the cleaving potential technique with a multistep thermodynamic perturbation method.

Although both the fluctuation and cleaving potential methods can yield accurate values of solid–liquid interfacial free energy, the simulations are computationally expensive even for simple fluid systems such as hard sphere and Lennard-Jones. An efficient simulation approach to obtain

* Corresponding author e-mail: xczeng@phase2.unl.edu.

Table 1. Comparison of Temperature Hysteresis at Pressure $P = 36.32 \text{ } \epsilon/\sigma^3$ ^a

	T_+ (ϵ/k_B)	T_- (ϵ/k_B)	T_m (ϵ/k_B)	θ_c^+	θ_c^-	β	V_s (σ^3)	V_l	$\Delta H_{m,v}$ (ϵ/σ^3)	γ_{sl} (ϵ/σ^3)	Q (K/ps)
Luo et al. ¹⁵	3.314	1.852	2.688	1.233	0.689	1.954	0.833	0.881	3.380	1.530	8.33
this work	3.14(8)	1.85(6)	2.75(7)	1.24(6)	0.67(4)	2.0(9)	0.828(3)	0.877(4)	3.54(7)	1.6(3)	8.33

^a Extensive quantities are presented per atom. Numbers in parentheses indicate the estimated error on the last digit(s) shown.

orientation averaged value of solid–liquid interfacial free energy is the superheating–undercooling hysteresis method developed by Luo et al.¹⁴ These authors demonstrated that this simulation method can give fair estimation of the solid–liquid interfacial free energy for the Lennard-Jones (LJ) system.¹⁵ They also estimated the interfacial free energy of liquid water/ice system based on experimental undercooling data.¹⁶ Moreover, a direct comparison of solid–liquid interfacial free energy for the LJ system computed from the hysteresis method and the fluctuation method or the cleaving potential technique was also made.¹⁶ The excellent agreement demonstrated the accuracy of the hysteresis method. Here, we employed such a superheating–undercooling hysteresis method to estimate the orientation averaged solid–liquid interfacial free energy of liquid water/ice system, for which the required superheating–undercooling data were obtained from MD simulations with two models of water.

Computational Method

Details of the superheating–undercooling hysteresis method are described elsewhere.¹⁴ The method results in a formula which relates interfacial free energy γ_{sl} with melting temperature T_m , enthalpy change of melting per unit volume $\Delta H_{m,v}$, and a dimensionless nucleation barrier parameter β

$$\gamma_{sl} = \left(\frac{3}{16\pi} \beta k_B T_m \Delta H_{m,v}^2 \right)^{1/3} \quad (1)$$

where k_B is the Boltzmann constant. Based on the classical nucleation theory and undercooling experiments, the maximum superheating $\theta_c^+ = T_+/T_m$ and undercooling $\theta_c^- = T_-/T_m$ can be established as

$$\beta = (A_0 - b \log_{10} Q) \theta_c (1 - \theta_c)^2 \quad (2)$$

where A_0 and b were fitted to be 59.4 and 2.33, respectively, for a number of elements and compounds. The heating/cooling rate Q is normalized by 1 K/s.

We adopted a procedure similar to that reported in the original paper¹⁴ to determine the highest temperature T_+ achievable in a superheated solid and the lowest temperature T_- achievable in an undercooled liquid, before a phase transformation occurs. First, a proton-disordered hexagonal ice I_h is equilibrated at an initial temperature (153.6 K for TIP4P-Ew¹⁷ and 150.5 K for TIP5P-Ew¹⁸) in the MD simulation with the isobaric–isothermal (NPT) ensemble. The temperature and pressure (1 bar) were controlled by using Nose-Hoover¹⁹ technique. Standard periodic boundary conditions were applied in all directions of the orthorhombic box containing 768 water molecules. Both TIP4P-Ew and TIP5P-Ew water molecules were treated as rigid bodies in the MD simulations, and the corresponding rotation equations were solved by using the quaternion algorithm with a time step of 1.0 and 0.5 fs, respectively. Next, the solid (ice) phase

is subjected to incremental heating until it melts. Thereafter, the melt (liquid water) is subjected to incremental cooling. Thermodynamic properties were calculated in every 50 ps heating/cooling step, after a 50 ps run for system equilibration. At the end of each heating/cooling step the temperature was increased or decreased by 3.8 K, corresponding to a heating/cooling rate of 0.076 K/ps. All MD simulations were performed using the DL_POLY2 package.²⁰ The long-range charge–charge interactions were treated with the smooth-particle-mesh-Ewald (SPME) technique.

Results and Discussions

A. Benchmark Test: Melting and Freezing of the Lennard-Jones (LJ) System. We first carried out a benchmark simulation to calculate the solid–liquid interfacial free energy of the LJ system. This test allowed us to examine the feasibility of this approach and to compare results of the unshifted LJ potential obtained in this work with results of the modified LJ potential reported.¹⁵ Only one case ($P=36.32 \text{ } \epsilon/\sigma^3$, where ϵ and σ are the energy and length parameters of LJ potential) was considered for the purpose of comparison. All the simulation parameters were set the same as those reported,¹⁵ except that we used the unshifted LJ potential with a cutoff at 2.5σ . In general, our results are in good agreement with the previous ones,¹⁵ except that we obtained slightly higher values of T_+ and $\Delta H_{m,v}$. A more detailed comparison is shown in Table 1.

B. Solid–Liquid Interfacial Free Energy of TIP4P-Ew and TIP5P-Ew Water. Although homogeneous nucleation has been demonstrated in undercooling experiments, accurate superheating data for ice are rarely reported largely because heterogeneous melting renders measuring the correct superheating limit T_+ difficult. Conversely, homogeneous crystallization of liquid water is rarely reported in MD simulations except for one work.²¹ This is because ice nucleus formation is a rare event in the MD simulation of undercooled water. Similarly, Zheng et al.²² reported that recrystallization of complex molecules by cooling the liquid is very difficult to achieve in MD simulations. Although it is challenging to determine the limiting value of T_- from MD simulation, β and γ_{sl} can be deduced from either T_+ or T_- for given T_m and $\Delta H_{m,v}$. Note that without T_- the melting point T_m cannot be estimated using the formula given in the hysteresis method.¹⁴

$$T_m = T_+ - \sqrt{T_+ T_-} + T_- \quad (3)$$

However, the equilibrium melting temperature T_m for both TIP4P-Ew and TIP5P-Ew water models can be determined using other independent computational approach, for example, the two-phase coexistence approach reported previously.^{23–24} ΔH_m can be calculated from the enthalpy difference between the solid and liquid at T_m , while $\Delta H_{m,v}$

Table 2. Comparison of Calculated Interfacial Free Energy at $P = 1$ Bar for the Two Water Models^a

	T_+ (K)	T_m (K)	θ_c^+	β	V_s (\AA^3)	V_l (\AA^3)	$\Delta H_{m,v}$ ($\times 10^8$ J/m ³)	γ_{sl} (mJ/m ²)	Q (K/ps)
TIP4P-Ew	321(6)	244(1) ^b	1.32(3)	4.5(9)	32.0(1)	30.2(2)	2.40(5)	37(3)	0.0762
TIP4P-Ew	317(7)	244(1) ^b	1.30(4)	4.1(9)	32.0(1)	30.2(2)	2.40(5)	36(3)	0.0200
TIP5P-Ew	314(6)	254(1)	1.24(3)	2.4(7)	31.4(1)	29.9(3)	3.90(7)	42(4)	0.0762
TIP5P-Ew	314(6)	254(1)	1.24(3)	2.5(7)	31.4(1)	29.9(3)	3.90(7)	42(4)	0.0200

^a Extensive quantities are presented per molecule. Numbers in parentheses indicate the estimated error on the last digit(s) shown. ^b The melting temperature of TIP4P-Ew ice is updated from the previously reported value 257.0K²³ to 244 ± 1 K based on a much longer (1 ns) two-phase-coexistence (in NPT ensemble) simulation with 12 288 water molecules. This new T_m value is very close to the result of $T_m = 242$ K reported by Fernandez et al.²⁴ using the same simulation method and a smaller system. The previously reported melting point 254 ± 1 K of TIP5P-Ew ice remains the same on basis of the longer (1 ns) simulation in NPT ensemble (see Supporting Information, Figure S3), about 16 K lower than $T_m = 270$ K reported by Fernandez et al.²⁴ Assuming $T_m = 270$ K, we also estimated the corresponding interfacial free energy for TIP5P-Ew, which is 36 mJ/m² (see Supporting Information, Table S1).

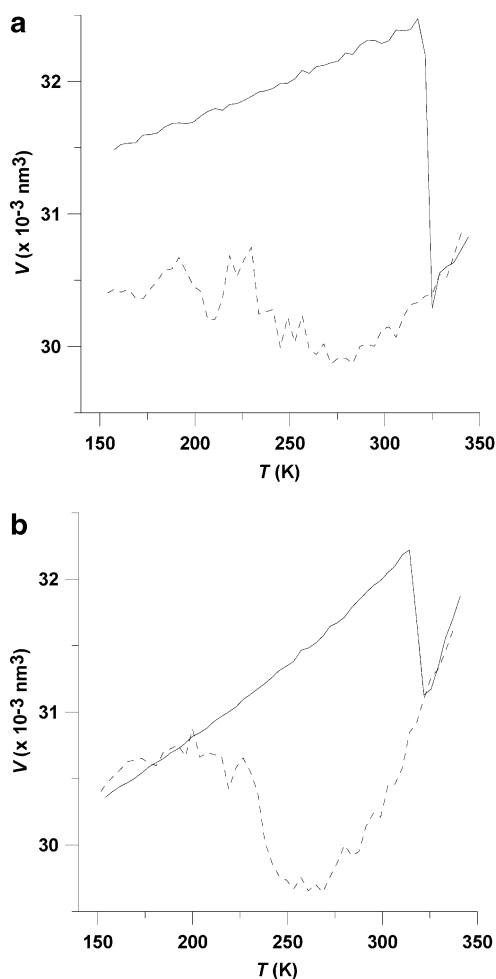


Figure 1. Temperature dependence of volume ($Q=0.0762$ K/ps) for (a) the TIP4P-Ew model and (b) the TIP5P-Ew model. Solid line represents superheating, and dashed line represents undercooling.

is normalized to the average volume of solid and liquid at the melting temperature.

As expected, upon superheating, the volume of solid ice gradually increases with increasing the temperature before a sudden reduction of the volume (due to the collapse of ice structure) (Figure 1). This behavior is unique in heating tetrahedral structure materials.²⁵ Near the superheating limit, there is an obvious potential energy jump (Figure 2) as well as one order-of-magnitude increase of diffusion coefficient (Figure 3). These observations confirmed that melting occurs at 321 K for TIP4P-Ew and 314 K for TIP5P-Ew. Moreover,

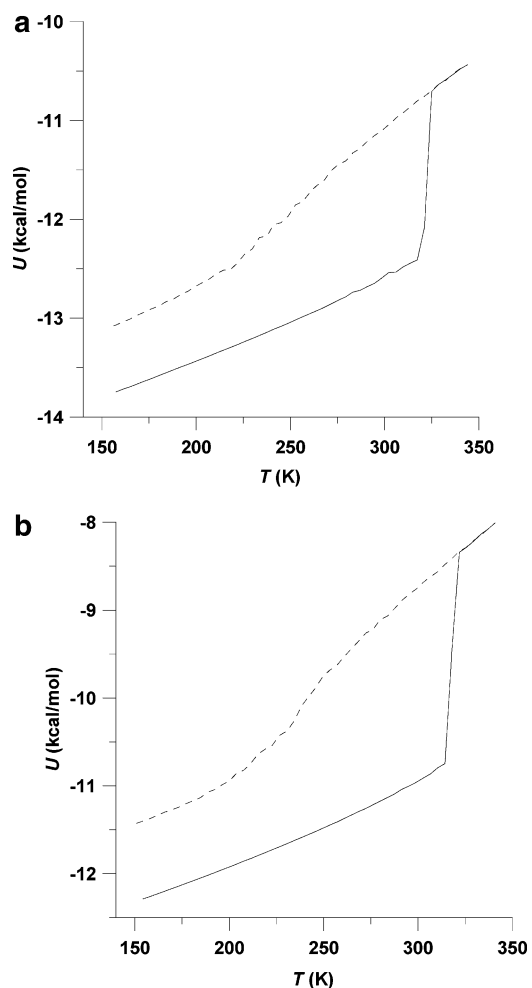


Figure 2. (a) Temperature dependence of potential energy ($Q=0.0762$ K/ps) for (a) the TIP4P-Ew model and (b) the TIP5P-Ew model. Solid line represents superheating and dashed line represents undercooling.

additional simulations using the constant stress-constant temperature ensemble and NPT ensemble with 2592 water molecules were performed to ensure that the superheating limit is not very sensitive to system size and box shape (Supporting Information, Figures S1 and S2). Although the diffusion coefficient of liquid water can decrease to the same magnitude as that of I_h ice below 210 K upon undercooling (Figure 3), no ordered structure was observed from the analysis of configuration snapshots at the low temperatures. A stiffer undercooling curve of volume change is obtained for TIP5P-Ew (Figure 1) but still not sufficient to locate T_-

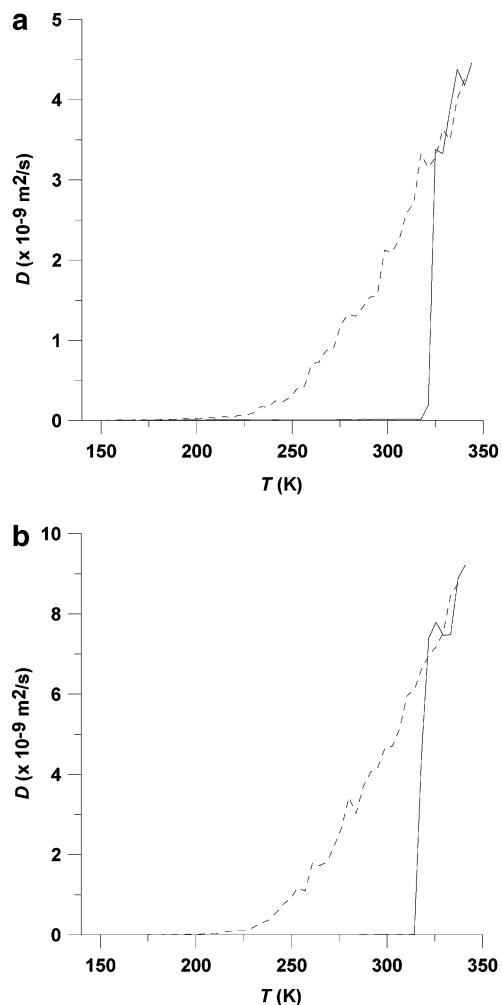


Figure 3. Temperature dependence of diffusion coefficient ($Q=0.0762\text{K/ps}$) for (a) the TIP4P-Ew model and (b) the TIP5P-Ew model. Solid line represents superheating and dashed line represents undercooling.

due to the continuous decrease of potential energy (Figure 2). The volume of liquid water eventually fluctuates near a constant after a slow increase from 280 to 230 K (Figure 1). Based on the temperature dependence of radial distribution function (Figure 4) the liquid water may undergo a continuous transformation toward an amorphous ice upon undercooling.

The calculated interfacial free energies γ_{sl} with two different heat/cooling rates for two water models are shown in Table 2. It appears that the heating/cooling rate has little effect on the calculated γ_{sl} . Overall, the calculated γ_{sl} are consistent with a previous MD simulation result²⁶ (39 mJ/m^2) as well as within the range of measured values¹⁶ ($25\sim 44 \text{ mJ/m}^2$). Conversely, both TIP4P-Ew and TIP5P-Ew models give rise to higher γ_{sl} compared to the result (28.0 mJ/m^2)¹⁶ and direct measurement of solid–liquid interfacial energy²⁷ (29.1 mJ/m^2). The discrepancy may be due in part to the empirical TIP4P-Ew and TIP5P-Ew models of water employed in this work. For example, both models underestimate the melting temperatures of water, which renders the material dependent parameter β larger by a factor of 4 (two for TIP5P-Ew) compared to the reported value¹⁶ (1.0).

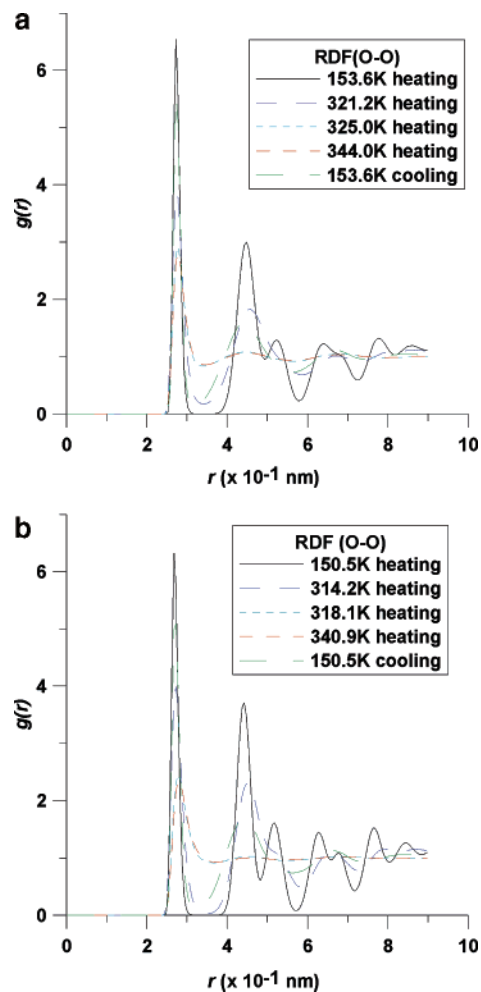


Figure 4. Temperature dependence of radial distribution function of oxygen atoms ($Q=0.0762\text{K/ps}$) for (a) the TIP4P-Ew model and (b) the TIP5P-Ew model.

Conclusion

In summary, we employed the Luo et al.'s method¹⁴ and superheating/undercooling data directly from MD simulations to estimate the solid–liquid interfacial free energy γ_{sl} of liquid water/ice interface with two water models. With the melting temperature T_{m} obtained from independent simulations,^{23,24} the calculated γ_{sl} are consistent with a previous direct MD simulation²⁵ but appreciably higher than the results obtained based on experimental undercooling data.¹⁶ More accurate values of the liquid water/ice interfacial free energy for the two model systems can be computed by using either the fluctuation or cleaving potential method. Research in this direction is under way.

Acknowledgment. We are grateful to Professor J. R. Morris and Professor X. Y. Song for valuable discussions. This research was supported by grants from DOE (DE-FG02-04ER46164), NSF (CHE-0427746 and CHE-0701540), John Simon Guggenheim Foundation, the Nebraska Research Initiative (X.C.Z.), and by the Research Computing Facility at University of Nebraska–Lincoln.

Supporting Information Available: Figures S1–S3 and Table S1. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Woodruff, D. P. The experimental determination of the solid-liquid interfacial free energy. In *The Solid-Liquid Interface*, 1st ed.; Cahn, R. W., Forty, A. J., Ward, I. M., Eds.; Cambridge University Press: London, U.K., 1973; Vol. 2, pp 12–31.
- (2) McMullen, W. E.; Oxtoby, D. W. *J. Chem. Phys.* **1988**, *88*, 1967.
- (3) Curtin, W. A. *Phys. Rev. Lett.* **1987**, *59*, 1228.
- (4) Marr, D. W.; Gast, A. P. *Phys. Rev. E* **1993**, *47*, 1212.
- (5) Gloor, G. J.; Jackson, G.; Blas, F. J.; de Miguel, E. *J. Chem. Phys.* **2005**, *123*, 134703.
- (6) Hoyt, J. J.; Asta, M.; Karma, A. *Phys. Rev. Lett.* **2001**, *86*, 5530.
- (7) Morris, J. R. *Phys. Rev. B* **2002**, *66*, 144104.
- (8) Asta, M.; Hoyt, J. J.; Karma, A. *Phys. Rev. B* **2002**, *66*, 100101.
- (9) Morris, J. R.; Song, X. *J. Chem. Phys.* **2003**, *119*, 3920.
- (10) Broughton, J. Q.; Gilmer, G. H. *J. Chem. Phys.* **1986**, *84*, 5759.
- (11) Davidchack, R. L.; Laird, B. B. *Phys. Rev. Lett.* **2000**, *85*, 4751.
- (12) Davidchack, R. L.; Laird, B. B. *J. Chem. Phys.* **2003**, *118*, 7651.
- (13) Mu, Y.; Song, X. *J. Chem. Phys.* **2006**, *124*, 034712.
- (14) Luo, S. N.; Ahrens, T. J.; Cagin, T.; Strachan, A.; Goddard, W. A., III; Swift, D. C. *Phys. Rev. B* **2003**, *68*, 134206.
- (15) Luo, S. N.; Strachan, A.; Swift, D. C. *J. Chem. Phys.* **2004**, *120*, 11640.
- (16) Luo, S. L.; Strachan, A.; Swift, D. C. *Modell. Simulat. Mater. Sci. Eng.* **2005**, *13*, 321.
- (17) Horn, W.; Swope, W. C.; Pitera, J. W.; Madura, J. C.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665.
- (18) Rick, S. W. *J. Chem. Phys.* **2004**, *120*, 6085.
- (19) Melchionna, S.; Ciccotti, G.; Holian, B. L. *Mol. Phys.* **1993**, *78*, 533.
- (20) Smith, W.; Forester, T. R. *DL_POLY_2.15*; Daresbury Laboratory: Cheshire, U.K. 2003.
- (21) Matsumoto, M.; Saito, S.; Ohmine, I. *Nature* **2002**, *416*, 409.
- (22) Zheng, L. Q.; Luo, S. N.; Thompson, D. L. *J. Chem. Phys.* **2006**, *124*, 154504.
- (23) Wang, J.; Yoo, S.; Bai, J.; Morris, J. R.; Zeng, X. C. *J. Chem. Phys.* **2005**, *123*, 036101.
- (24) Fernandez, R. G.; Abascal, J. L. F.; Vega, C. *J. Chem. Phys.* **2006**, *124*, 144506.
- (25) Tang, Y. W.; Wang, J.; Zeng, X. C. *J. Chem. Phys.* **2006**, *124*, 236103.
- (26) Haymet, A. D. J.; Bryk, T.; Smith, E. J. Solute Ions at Ice/Water Interface. In *Ionic Soft Matter: Modern Trends in Theory and Applications*; Proceedings of the NATO Advanced Research Workshop on Ionic Soft Matter, Lviv, Ukraine, April 14–17, 2004; Henderson, D., Holovko, M., Trokhymchuk, A., Eds.; Springer: London, 2005; pp 333–359.
- (27) Hardy, S. C. *Philos. Mag.* **1977**, *35*, 471.

CT600345S

Combined QM/MM Molecular Dynamics Study on a Condensed-Phase S_N2 Reaction at Nitrogen: The Effect of Explicitly Including Solvent Polarization

Daan P. Geerke,[†] Stephan Thiel,[‡] Walter Thiel,[‡] and Wilfred F. van Gunsteren^{*†}

Laboratory of Physical Chemistry, Swiss Federal Institute of Technology Zürich, ETH, CH-8093 Zürich, Switzerland, and Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, D-45470 Mülheim, Germany

Received January 11, 2007

Abstract: In a previous combined QM/MM molecular dynamics (MD) study from our laboratory on the identity S_N2 reaction between a chloride anion and an amino chloride in liquid dimethyl ether (DME), an increase in the free energy activation barrier was observed in the condensed phase when compared to the gas-phase activation energy. Here we reproduce these findings, but when comparing the condensed-phase potential of mean force (PMF) with the *free* energy profile in the gas phase (obtained from Monte Carlo simulations), we observe a smaller solvent effect on the activation barrier of the reaction. In a next step, we introduce an explicit description of electronic polarization in the MM (solvent) part of the system. A polarizable force field for liquid DME was developed based on the charge-on-spring (COS) model, which was calibrated to reproduce thermodynamic properties of the nonpolarizable model in classical MD simulations. The COS model was implemented into the MNDO/GROMOS interface in a special version of the QM/MM software ChemShell, which was used to investigate the effect of solvent polarization on the free energy profile of the reaction under study. A higher activation barrier was obtained using the polarizable solvent model than with the nonpolarizable force field, due to a better solvation of and a stronger polarization of solvent molecules around the separate reactants. The obtained PMFs were subjected to an energy-entropy decomposition of the relative solvation free energies of the reactant complex along the reaction coordinate, to investigate in a quantitative manner whether the solvent (polarization) effects are mainly due to favorable QM-MM (energetic) interactions.

I. Introduction

In the present study, we use a combined QM/MM Hamiltonian^{1–4} to extend a previous molecular dynamics (MD) study⁵ of solvent effects on a S_N2 reaction at nitrogen in liquid dimethyl ether (DME). In combined QM/MM simulations, a small part of the system (the reactive subsystem) is treated quantum mechanically, whereas the rest

of the system (the MM part) is described by a classical force field. The total combined Hamiltonian H consists of three terms^{1–4}

$$H = H_{\text{QM}} + H_{\text{QM/MM}} + H_{\text{MM}} \quad (1)$$

H_{QM} and H_{MM} describe interactions within the QM and MM parts of the system, respectively. $H_{\text{QM/MM}}$ accounts for interactions between the QM and MM subsystems and is often described by an electrostatic coupling scheme^{3,4,6–8} in which the field generated by the MM atoms is included in the QM Hamiltonian as a set of additional point-charges (first two terms on the right of eq 2). Other nonbonded interactions

* Corresponding author fax: (+41)-44-6321039; e-mail: wfvgn@igc.phys.chem.ethz.ch.

[†] Swiss Federal Institute of Technology Zürich.

[‡] Max-Planck-Institut für Kohlenforschung.

between the QM and MM atoms are mimicked by adding the QM nuclei as van der Waals centers to the force field (last term on the right in eq 2). $H_{\text{QM/MM}}$ reads then as

$$H_{\text{QM/MM}} = - \sum_{i,m} \frac{q_m}{r_{im}} + \sum_{a,m} \frac{Z_a q_m}{r_{am}} + \sum_{a,m} \left(\frac{C_{12,a}^{1/2} C_{12,m}^{1/2}}{r_{am}^{12}} - \frac{C_{6,a}^{1/2} C_{6,m}^{1/2}}{r_{am}^6} \right) \quad (2)$$

in which the indices a and i run over the QM nuclei and electrons, respectively, and m runs over the MM atoms. Z and q are the (partial) charges of the QM nuclei and MM (united) atoms, respectively, and $C_{12}^{1/2}$ and $C_6^{1/2}$ are their repulsive and attractive Lennard-Jones parameters.

The reaction under investigation is the identity $S_{\text{N}}2$ reaction of a chloride anion with amino chloride



The gas-phase potential energy surface (PES) for this reaction is characterized by a double-well potential, with a reactant or product ion-dipole complex (RC) located at the minima and a classical $S_{\text{N}}2$ transition state (TS) at the central barrier.^{5,9} Liu et al.⁵ showed that upon solvation in DME, the potential of mean force (PMF) for reaction 3 still has a double-well shape, but the free energy of formation of RC is less negative and the activation free energy barrier is more positive than the corresponding gas-phase energies. This can be explained from the distribution of the net negative charge over the reactive system, which is maximally concentrated on the nucleophile in the separated reactants. As a result, interactions with the polar DME liquid are stronger in the reactant state than for the RC and TS. Indeed, radial distribution functions for the solvent atoms around the nucleophile showed⁵ a larger first solvation peak in the reactant state, indicating better solvation than in case of the complexes.

In the current study, we first repeated the QM/MM study on reaction 3 in DME. The reactive subsystem was described at the PM3 level of theory, which was shown by Liu⁵ to be a valid choice. Like Liu, we use an electrostatic coupling scheme for the QM-MM interactions and a nonpolarizable DME force field. Thus, polarization of the electrons in the QM subsystem by the MM environment is automatically accounted for, but the solvent molecules cannot adapt their charge distribution in the course of the reaction. In addition we investigate the effect of taking solvent electron polarization into account on the PMF of reaction 3 from simulations using a combined QM/MM Hamiltonian with an explicit description of electron polarization effects in the MM part (designated as QM/MM-pol Hamiltonian). Because the net charge of the reactive subsystem varies from being concentrated on the nucleophile in the reactant state to being more spread out over the reactant complex and transition state, changes in the solvent molecular dipole moments along the reaction coordinate might well affect the reaction profile. Several MD studies^{1,10–14} employed a QM/MM-pol Hamiltonian before, using either the induced point-dipole^{1,15,16} or

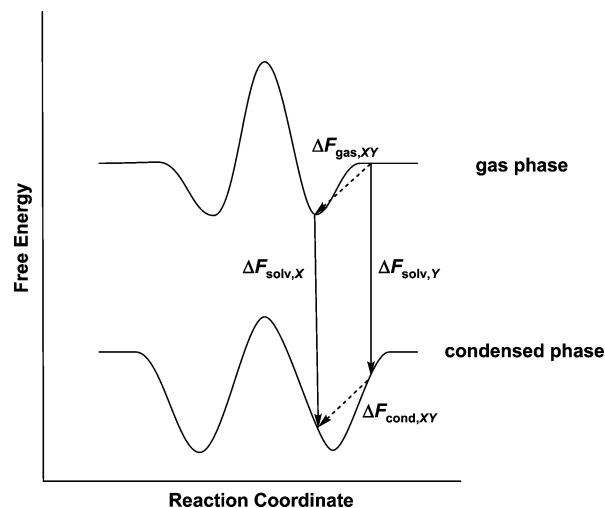


Figure 1. Schematic representation of the free energy profiles of a $S_{\text{N}}2$ reaction in the gas phase and in solution. The vertical arrows indicate the free energy of solvation ΔF_{solv} of the reactive complex at values for the reaction coordinate of X and Y , respectively. The dashed arrows are the relative free energy differences ΔF_{gas} and ΔF_{cond} along the reaction coordinate in the gas and condensed phase, respectively.

the fluctuating charge method^{17,18} to account for electron polarization in the MM subsystem. Here we use the charge-on-spring (COS) model (or Drude-oscillator or shell model)^{19–21} to describe the MM inducible dipoles. An attractive feature of the COS method for use in QM/MM simulations is that the induced dipoles directly enter $H_{\text{QM/MM}}$ via additional point charges in eq 2. For our QM/MM-pol simulations, a COS-based model for DME was parametrized based on the nonpolarizable force field, and the Charge-on-Spring model was implemented into a special version²² of the GROMOS96 software^{23,24} interfaced to the ChemShell QM/MM software package.⁸

Liu studied solvent effects on reaction 3 from a comparison between the condensed-phase free energy profile (PMF) and the gas-phase potential energy surface (PES). To obtain a more complete picture of the solvent effects, we also consider differences with the gas-phase free energy profile along the reaction coordinate. Once the free energy profiles in vacuum and the solvent are known, a direct quantitative measure for the solvent effect on the reaction profile can be obtained by defining the relative difference in the “free energy of solvation” of the reactive subsystem ($\Delta\Delta F_{\text{solv}}$) along the reaction coordinate rc

$$\Delta\Delta F_{\text{solv},XY} = \Delta F_{\text{solv},X} - \Delta F_{\text{solv},Y} \quad (4)$$

with $\Delta F_{\text{solv},X}$ and $\Delta F_{\text{solv},Y}$ the free energy change upon transferring the reactive subsystem from vacuum into the liquid at values X and Y for rc , respectively. Values for $\Delta\Delta F_{\text{solv}}$ can also be calculated from applying a thermodynamic cycle (see Figure 1)

$$\Delta\Delta F_{\text{solv},XY} = \Delta F_{\text{cond},XY} - \Delta F_{\text{gas},XY} \quad (5)$$

with $\Delta F_{\text{cond},XY}$ and $\Delta F_{\text{gas},XY}$ being the condensed-phase and gas-phase free energy difference along the reaction coordinate when going from $rc = Y$ to $rc = X$, respectively.

Liu's finding that better solvation of the charge-localized reactants is responsible for the increase in the barrier height when compared to the gas-phase PES hints at energetic interactions determining the solvent effect. In an attempt to quantify whether this is the case, we apply an energy-entropy decomposition scheme for solvation free energies ΔF_{solv} , which has recently successfully been applied to study driving forces behind hydrophobic solvation and cosolvent interactions in aqueous mixtures.^{25–29} Starting point of this decomposition scheme is to separate ΔF_{solv} (corresponding with the free energy change of bringing a solute molecule (indicated by the index “u”) from vacuum into a solvent (indicated by the index “v”) into energy and entropy contributions from solute–solute (uu), solute–solvent (uv), and solvent–solvent (vv) interactions

$$\begin{aligned}\Delta F_{\text{solv}} &= \Delta U - T\Delta S \\ &= \Delta U_{\text{uu}} + \Delta U_{\text{uv}} + \Delta U_{\text{vv}} \\ &\quad - (T\Delta S_{\text{uu}} + T\Delta S_{\text{uv}} + T\Delta S_{\text{vv}})\end{aligned}\quad (6)$$

From statistical mechanics, it can be shown that the energy and entropy terms arising from solvent–solvent interactions exactly cancel out,³⁰ that is

$$\Delta U_{\text{vv}} = T\Delta S_{\text{vv}}\quad (7)$$

Here, we include solute–solute energies and entropies of solvation into the corresponding solute–solvent terms, which reduces eq 6 to

$$\Delta F_{\text{solv}} = \Delta U_{\text{uv}} - T\Delta S_{\text{uv}}\quad (8)$$

Thus, ΔU_{uv} (accounting for energetic interactions between the solute and solvent) and ΔS_{uv} (a measure for the probability for the solvent to open up solute-sized cavities and to undergo favorable interactions with the solute) are the terms to analyze to understand trends in ΔF_{solv} at a microscopic level.^{25–29}

Here, we consider the reactive subsystem as the solute and monitor relative differences in its free energy of solvation along a reaction coordinate *rc* (see Figure 1). Using eq 4, $\Delta\Delta F_{\text{solv},XY}$ (the relative difference in ΔF_{solv} in going from an *rc* value of *Y* to *X*) reads then

$$\begin{aligned}\Delta\Delta F_{\text{solv},XY} &= \Delta F_{\text{solv},X} - \Delta F_{\text{solv},Y} \\ &= \Delta U_{\text{uv},X} - T\Delta S_{\text{uv},X} - (\Delta U_{\text{uv},Y} - T\Delta S_{\text{uv},Y}) \\ &= \Delta\Delta U_{\text{uv},XY} - T\Delta\Delta S_{\text{uv},XY}\end{aligned}\quad (9)$$

where $\Delta\Delta U_{\text{uv},XY}$ and $T\Delta\Delta S_{\text{uv},XY}$ are the relative differences in ΔU_{uv} and $T\Delta S_{\text{uv}}$ between *rc* values of *X* and *Y*, respectively. Here, $\Delta\Delta U_{\text{uv},XY}$ is simply the corresponding difference in solute–solvent (QM-MM) interaction energy, corrected by the solute–solute (QM) reorganization energy

$$\begin{aligned}\Delta\Delta U_{\text{uv},XY} &= \Delta U_{\text{uv},X} - \Delta U_{\text{uv},Y} \\ &= \hat{H}_{\text{QM/MM},X} + \hat{H}_{\text{QM,liq},X} - \hat{H}_{\text{QM,gas},X} \\ &\quad - (\hat{H}_{\text{QM/MM},Y} + \hat{H}_{\text{QM,liq},Y} - \hat{H}_{\text{QM,gas},Y})\end{aligned}\quad (10)$$

and

$$T\Delta\Delta S_{\text{uv},XY} = \Delta\Delta U_{\text{uv},XY} - \Delta\Delta F_{\text{solv},XY}\quad (11)$$

In the QM/MM-pol simulations, changes in the self-polarization energy of the solvent molecules along the reaction coordinate (ΔU_{self}) are included in $\Delta\Delta U_{\text{uv}}$ (because U_{self} accounts for the cost of the change in polarization of the solvent molecules due to a change in the value of *rc*), and eq 10 reads

$$\begin{aligned}\Delta\Delta U_{\text{uv},XY} &= \hat{H}_{\text{QM/MM},X} + \hat{H}_{\text{QM,liq},X} - \hat{H}_{\text{QM,gas},X} \\ &\quad - (\hat{H}_{\text{QM/MM},Y} + \hat{H}_{\text{QM,liq},Y} - \hat{H}_{\text{QM,gas},Y}) \\ &\quad + \Delta U_{\text{self},XY}\end{aligned}\quad (12)$$

where

$$\Delta U_{\text{self},XY} = U_{\text{self},X} - U_{\text{self},Y}\quad (13)$$

From trends in $\Delta\Delta U_{\text{uv}}$ and $T\Delta\Delta S_{\text{uv}}$ we analyze the origin of free energy differences along the reaction coordinate and, hence, the solvent (polarization) effect on reaction 3 in DME.

II. The Charge-on-Spring Model To Account for Electron Polarization in the MM Subsystem

Following the charge-on-spring (COS) model,^{21,22,31} the MM inducible dipoles are represented by an additional massless site with a point charge $q_{\text{pol},i}$ attached to the polarizable MM centers *i* (having a charge of $q_i - q_{\text{pol},i}$), via a spring with a force constant depending on its polarizability α_i . The induced dipoles $\vec{\mu}_{\text{ind},i}$ are then represented by a displacement $\Delta\vec{r}_{\text{pol},i}$ of the spring

$$\vec{\mu}_{\text{ind},i} = q_{\text{pol},i}\Delta\vec{r}_{\text{pol},i}\quad (14)$$

For a given configuration of QM and MM atoms, the QM wavefunction is optimized in an SCF calculation in which the wavefunction ‘feels’ the MM polarization by including the charges-on-spring and their (fixed) positions in the first two terms in eq 2. Subsequently, the displacements $\Delta\vec{r}_{\text{pol},i}$ of the charges-on-spring attached to the MM polarizable centers are calculated as

$$\Delta\vec{r}_{\text{pol},i} = \frac{\alpha_i(4\pi\epsilon_0)\vec{E}_i}{q_{\text{pol},i}}\quad (15)$$

where \vec{E}_i is the electric field at center *i* and is calculated as the sum of the contributions from the MM nuclei and charges-on-spring (see eq 59 in ref 31) and from the (fixed) QM wavefunction (determined from the QM gradients at *i*). Because the \vec{E}_i 's depend on the positions of all other charges-on-spring, the $\vec{\mu}_{\text{ind},i}$'s ($\Delta\vec{r}_{\text{pol},i}$'s) in the MM subsystem are determined in an iterative way, until the induced dipole interaction energies are converged. Thus, a doubly iterative scheme has to be employed: since the polarization of the QM electrons and MM polarizable centers affect each other, iterations over the QM and MM SCF procedures are performed until the total QM/MM-pol energy is converged. To ensure full convergence of the QM electron wavefunction,

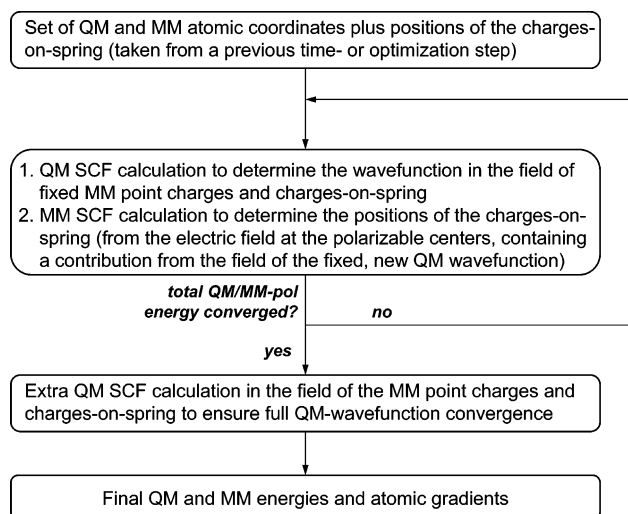


Figure 2. Schematic representation of the SCF procedure to determine the total QM/MM-pol energy and the gradients on the QM and MM nuclei for a set of atomic positions, as implemented in a special version of the GROMOS interface to ChemShell that is adapted to the charge-on-spring model.

an additional QM SCF calculation is performed after the last QM/MM iteration step. This procedure is summarized in Figure 2. The total forces on the MM polarizable centers for the use in the next MD or geometry optimization step are calculated by adding the contributions from the electrostatic gradients on the massless charges-on-spring to the forces acting on the atom they are attached to.^{21,22} Additionally to the extra point charges (the charges-on-spring) that enter H_{MM} and $H_{\text{QM/MM}}$, a self-polarization energy term (U_{self}) is to be added to the QM/MM-pol Hamiltonian to account for the energy cost of inducing the dipoles³²

$$U_{\text{self}} = \frac{1}{2} \sum_i \frac{\vec{\mu}_{\text{ind},i} \cdot \vec{\mu}_{\text{ind},i}}{\alpha_i (4\pi\epsilon_0)} = \frac{1}{2} \sum_i \frac{q_{\text{pol},i}^2 |\Delta\vec{r}_{\text{pol},i}|^2}{\alpha_i (4\pi\epsilon_0)} \quad (16)$$

U_{self} is a direct measure of the contribution from the induced dipoles to the total energy of the system, since it is the negative of this contribution.³²

III. Simulation Setup

Potential energy surfaces (PES) and potentials of mean force (PMF) for reaction 3 were obtained along a reaction coordinate rc defined in terms of the bond distances of the incoming and the leaving chloride anion to the central nitrogen, $r_{\text{N-Cl1}}$ and $r_{\text{N-Cl2}}$, respectively

$$rc = r_{\text{N-Cl1}} - r_{\text{N-Cl2}} \quad (17)$$

with $rc = 0.0$ nm corresponding to the transition state and rc adopting an infinite value for the separate reactants. The gas-phase PES was obtained from energy minimizations (using the hdlcopt geometry optimizer³³ implemented in ChemShell⁸) in which values for rc were kept fixed and gradually increased in steps of 0.01 nm starting from $rc = 0.0$ nm (the transition state) until an energy plateau was reached. The PMF in vacuum was obtained from a series of Monte Carlo (MC) simulations at 248.15 K, in which a

harmonic potential-energy function was employed to keep rc close to a target value rc_{target} (varied from 0.0 to 1.1 nm with increments of 0.01 nm)

$$V_{\text{umb}} = \frac{1}{2} k_{\text{umb}} (rc - rc_{\text{target}})^2 \quad (18)$$

The force constant k_{umb} was set to a large value (6×10^6 kJ mol⁻¹ nm⁻²) to ensure narrow distributions of rc values around rc_{target} . In every MC simulation, 10^6 trial steps were performed (after 1000 steps of equilibration) in which the Cartesian coordinates of one of the atoms were varied with a maximum random displacement of 10^{-3} nm. Gas-phase energies and gradients were calculated at the PM3 level of theory using the MNDO99 code³⁴ interfaced to ChemShell.

In the combined QM/MM simulations of reaction 3 in liquid dimethyl ether (DME), the reactive subsystem was described by the PM3 Hamiltonian and the solvent by either a nonpolarizable (DME_{nonpol}) or a polarizable force field (DME_{pol}). Parameters for the (rigid, united-atom) DME_{nonpol} model were taken from Liu et al.,⁵ see Table 1. The DME_{pol} model was developed based on DME_{nonpol} using the charge-on-spring (COS) method.^{19–21} Massless sites with a charge of $q_{\text{pol}} = -8.0 e$ were attached to DME's atomic centers. The polarizabilities α_i for O and CH₃ were taken from Miller,³⁵ where α_i of the CH₃ group was calculated as the sum of the polarizabilities of the atoms it is composed of. The nonbonded parameters were adapted to obtain a model that reproduces the density and heat of vaporization of the DME_{nonpol} model in MD simulations of the pure liquid under NpT boundary conditions. The rationale was to lower the Lennard-Jones well depth of one of the atom types (the oxygen) in the DME_{pol} model, since this parameter effectively accounts for polarization effects in nonpolarizable force fields. For the simulations of the DME_{nonpol} liquid, the GROMOS96 simulation package^{23,24} was used, and for the DME_{pol} simulations an adapted version²² of this package was used. Temperature and pressure were kept constant at 248.15 K and 1 atm using the weak-coupling approach³⁶ with coupling times of 0.1 and 0.5 ps, respectively, and an isothermal compressibility of 4.575×10^{-4} (kJ mol⁻¹ nm⁻³)⁻¹ for the pressure bath. Bond lengths were kept fixed using the SHAKE procedure³⁷ with a relative geometric accuracy of 10^{-4} . The time step was 2 fs. A triple-range cutoff scheme was applied: nonbonded interactions within 0.8 nm were calculated every step from a charge-group based pairlist that was updated every five steps. At these time points, interactions between 0.8 and 1.4 nm were calculated and kept constant between updates. A reaction-field contribution³⁸ was added to the electrostatic interactions and forces, in which the dielectric permittivity was set to 4. The system consisted of 512 molecules which were initially placed in a random orientation in a cubic box at the experimental density³⁹ of 737 kg m⁻³. After an initial energy minimization, the system was equilibrated for 20 ps and simulated for an additional 200 ps in which data were saved every 100 steps for analysis. The density and heat of vaporization of the liquid were calculated, as well as its static dielectric permittivity which was calculated from fluctuations in the total dipole moment of

Table 1. Nonpolarizable (DME_{nonpol})⁵ and Polarizable (DME_{pol}) Force-Field Parameter Sets for the Rigid United-Atom Models for Liquid Dimethyl Ether and van der Waals Parameters⁵ for the QM Subsystem (NH₂Cl₂)

parameter ^a	DME _{nonpol}	DME _{pol}	parameter ^a	NH ₂ Cl ₂
r _{O-CH₃} [nm]	0.141	0.141		
r _{CH₃-CH₃} [nm]	0.2324	0.2324		
q _O [e]	-0.36	-0.36		
q _{CH₃} [e]	0.18	0.18		
C ₆ ^{1/2} (O) [kJ mol ⁻¹ nm ⁶] ^{1/2}	0.04756	0.042663	C ₆ ^{1/2} (N) [kJ mol ⁻¹ nm ⁶] ^{1/2}	0.04936
C ₁₂ ^{1/2} (O) [10 ⁻³ (kJ mol ⁻¹ nm ¹²) ^{1/2}]	0.86115	0.7761	C ₁₂ ^{1/2} (N) [10 ⁻³ (kJ mol ⁻¹ nm ¹²) ^{1/2}]	1.301
C ₆ ^{1/2} (CH ₃) [kJ mol ⁻¹ nm ⁶] ^{1/2}	0.09421	0.09421	C ₆ ^{1/2} (Cl) [kJ mol ⁻¹ nm ⁶] ^{1/2}	0.1175
C ₁₂ ^{1/2} (CH ₃) [10 ⁻³ (kJ mol ⁻¹ nm ¹²) ^{1/2}]	5.1137	5.1137	C ₁₂ ^{1/2} (Cl) [10 ⁻³ (kJ mol ⁻¹ nm ¹²) ^{1/2}]	10.34
α _O [10 ⁻³ nm ³]		0.637	C ₆ ^{1/2} (H) [kJ mol ⁻¹ nm ⁶] ^{1/2}	0.0
α _{CH₃} [10 ⁻³ nm ³]		2.222	C ₁₂ ^{1/2} (H) [10 ⁻³ (kJ mol ⁻¹ nm ¹²) ^{1/2}]	0.0

^a r_{O-CH₃} and r_{CH₃-CH₃}: O-CH₃ and CH₃-CH₃ bond lengths corresponding with a CH₃-O-CH₃ angle of 111.0 degrees for both models. *q*: partial charges, C₆^{1/2} and C₁₂^{1/2}: attractive and repulsive van der Waals parameters, and α: atomic polarizability.

the simulation box, according to a Kirkwood-Fröhlich type equation derived by Neumann.^{22,40}

In the QM/MM simulations, interactions between the QM and MM subsystems were treated using an electrostatic coupling scheme (eq 2), and van der Waals parameters for the QM atoms are given in Table 1. Energy and gradient evaluations for the QM and MM regions were performed by the MNDO99³⁴ and GROMOS96^{23,24} codes, respectively, interfaced to ChemShell.⁸ For the energy and gradient evaluations in the QM/MM-pol simulations, a doubly iterative scheme (Figure 2) was implemented in the ChemShell interface, using a special version of the GROMOS96 code adapted to the COS model.²² For every configuration of atomic positions, a fixed number of QM/MM SCF iterations was performed, starting with the COS displacements Δr̄_{pol,*i*} as determined in the previous time or optimization step. The number of iterations was set to 4, which was found to be large enough to achieve an energy convergence of the total QM/MM-pol Hamiltonian within 0.025 kJ mol⁻¹. Energy convergence criteria for the separate QM and MM SCF calculations were set to 10⁻⁸ eV and 10⁻³ kJ mol⁻¹, respectively.

Molecular dynamics simulations of reaction 3 in liquid DME under NVT conditions were performed using the dynamics routine of ChemShell,⁸ with a time step of 0.5 fs. The temperature was kept constant at 248.15K using a Berendsen thermostat³⁶ with a coupling time of 0.1 ps. Geometries of the solvent molecules were kept rigid by constraining bond lengths using the SHAKE algorithm³⁷ with a relative geometric accuracy of 10⁻⁸. For the QM-MM interactions, a straight cutoff truncation scheme was applied. Interactions between the QM subsystem and DME solvent molecules were taken into account if the position of the DME molecule's center of geometry is within 1.1 nm of any of the QM atoms. For the MM-MM interactions, a twin-range cutoff was applied using a charge-group based pairlist which was updated every 5 steps: interactions between molecules with the distance between the center of geometries of less than 0.8 nm were calculated explicitly every step and between 0.8 and 1.1 nm every fifth step.

The free energy profile in DME was obtained from a series of QM/MM MD simulations at different values for *rc* in which its value was constrained to a target value *rc*_{target} using the SHAKE procedure.^{37,41} Simulations were performed at

63 different values for *rc*_{target}: between values of 0–0.04 nm, *rc*_{target} was incremented in steps of 0.004 nm; between 0.04 and 0.2 nm, *rc*_{target} was incremented in steps of 0.005 nm; and between 0.2 and 0.4 nm, *rc*_{target} was incremented in steps of 0.01 nm. Initial coordinates for the simulations were generated as follows. First, the gas-phase optimized geometry of the transition state was placed in the center of a cubic box which was filled by adding 350 DME molecules in a random orientation. The box volume (35.8929 nm³) was chosen such that the density of the solvent was equal to the density of the liquid model. After an energy minimization, initial atomic velocities were assigned from a random Maxwell-Boltzmann distribution corresponding to a temperature of 248.15 K, and a series of equilibration runs was performed with increasing values of *rc*_{target} (starting from 0.0 nm). Every equilibration run of 5 ps started with the final set of atomic positions and velocities from the previous simulation. In the initial geometry optimization and equilibration runs, the DME_{nonpol} model was used. The final configurations of the equilibration runs were used as starting configurations for the production runs using the nonpolarizable force field and for additional equilibration runs of 4 ps using the DME_{pol} model to generate starting structures for the production runs with the QM/MM-pol Hamiltonian. All production runs had a length of 20 ps. Constrained forces, energies, QM charge distributions, and positions of the MM atoms and charges-on-spring were saved every 100 step for analysis. To check for inaccuracies in the condensed-phase PMF due to hysteresis, we additionally performed the same sets of QM/MM simulations but starting from *rc* = 0.4 nm in the equilibration procedure: the separated reactants were solved in DME, and *rc* was gradually decreased (in steps of 0.01 nm) to *rc* = 0.0 nm in the equilibration simulations using the DME_{nonpol} model. In this way, starting structures for the production runs and for the QM/MM-pol equilibration runs were obtained.

Free energy differences along the reaction coordinate were calculated using the thermodynamic integration formalism:⁴² identifying ξ with the reaction coordinate, the free energy difference Δ*F*_{*a*-*b*} between two values *a* and *b* for ξ is evaluated as the potential of mean force of constraint⁴³⁻⁴⁵

$$\Delta F_{a-b} = \int_a^b d\xi \left\langle \frac{\partial V}{\partial \xi} \right\rangle_{\xi} \quad (19)$$

where $\langle \partial V / \partial \xi \rangle_{\xi}$ is the force of constraint along the reaction coordinate, defined as the derivative of the potential energy V of the system with respect to ξ . In the vacuum simulations, this force was directly calculated from the average value of the umbrella force⁴⁶ at every MC step. In the condensed-phase simulations, it was evaluated as the average of the difference between unconstrained and constrained forces.⁴¹ The use of a constraint on a reaction coordinate results in the introduction of a metric tensor effect,^{43,47–51} which in the case of a reaction coordinate as defined in eq 17 can be corrected for by adding a term to eq 19 after which the expression for $\Delta F_{a \rightarrow b}$ reads as

$$\Delta F_{a \rightarrow b} = \int_a^b d\xi \left\langle \frac{\partial V}{\partial \xi} \right\rangle_{\xi} - k_B T \ln \frac{\langle z^{-1/2} \rangle_{\xi=b}}{\langle z^{-1/2} \rangle_{\xi=a}} \quad (20)$$

with

$$z = m_{\text{Cl1}}^{-1} + 2m_{\text{N}}^{-1}(1 + \cos(\vec{r}_{\text{N-Cl1}}, \vec{r}_{\text{N-Cl2}})) + m_{\text{Cl2}}^{-1} \quad (21)$$

where $\cos(\vec{r}_{\text{N-Cl1}}, \vec{r}_{\text{N-Cl2}})$ is the cosine of the angle between the N–Cl₁ and N–Cl₂ (bond) vectors, and the m_i 's are the masses of the atoms. The integral in eq 19 was evaluated via trapezoidal integration. Errors in energy values and the constrained and restrained forces were estimated at every rc value using the block-averaging procedure described by Allen and Tildesley.⁵² Individual errors in the forces were integrated to yield the total error in the free energy differences along the reaction coordinate.

Free energies of solvation (ΔF_{solv}) in nonpolarizable DME were calculated for the transition state and the separated reactants. Three sets of simulations were performed: one for the complete QM subsystem in which rc was kept constrained at 0.0 nm (transition state). For the reactants, two separate sets of simulations of NH₂Cl and Cl[−] in DME_{nonpol} were performed. In all three cases, two subsets of simulations were performed: first, electrostatic interactions between the QM and MM subsystems were gradually turned off using a coupling parameter λ that linearly scales down QM-MM electrostatic interactions (with $\lambda = 1$ corresponding to full and $\lambda = 0$ corresponding to no interactions). Subsequently a set of simulations was performed in which QM-MM van der Waals interactions were linearly scaled down from $\lambda = 1$ to $\lambda = 0$. Simulations were performed at 21 evenly distributed λ -points, with 5 ps equilibration and 20 ps of production per λ -point. At every λ -point, the free energy change in going from the actual value λ_a to $\lambda_b = \lambda_a - 0.05$, was calculated using perturbation theory⁵³

$$\Delta F_{\lambda_a \rightarrow \lambda_b} = -k_B T \ln \langle e^{-\frac{H(\lambda_b) - H(\lambda_a)}{k_B T}} \rangle_{\lambda_a} \quad (22)$$

with $H(\lambda_x)$ being the total potential energy of the system calculated with the Hamiltonian corresponding to λ_x from configurations saved every 100 steps during the production runs. ΔF_{solv} was obtained as minus the total sum of the values for $\Delta F_{\lambda_a \rightarrow \lambda_b}$ (corresponding to the free energy of turning off electrostatic and van der Waals QM-MM interactions, respectively) calculated at $\lambda_a = 1, 0.95, \dots, 0.05$.

IV. Results and Discussion

IV.1. Parametrization of a Polarizable Force Field for Liquid Dimethyl Ether. The optimized polarizable param-

Table 2. Thermodynamic and Dielectric Properties for Liquid Dimethyl Ether at 248.15 K and 1 atm from Experiment and from MD Simulations Using the Nonpolarizable Force Field (DME_{nonpol})⁵ and the Polarizable Model (DME_{pol}) Developed in the Current Work

property ^a	experiment	DME _{nonpol}	DME _{pol}
T [K]	248.15	248.0	248.0
ρ [kg m ^{−3}]	737 ^b	751	749
ΔH_{vap} [kJ mol ^{−1}]	21.7 ^c	21.3	21.1
$-U_{\text{pot}}$ [kJ mol ^{−1}]		19.2	20.1
U_{self} [kJ mol ^{−1}]			1.1
ϵ		4.7	6.8

^a T : temperature, ρ : density, ΔH_{vap} : heat of vaporization, U_{pot} : potential energy, U_{self} : self-polarization energy, ϵ : static relative dielectric permittivity. ^b Reference 39. ^c Reference 54.

eter set for liquid dimethyl ether (DME) is given in Table 1 and reproduces the density and heat of vaporization of the nonpolarizable DME_{nonpol} model within 1%, see Table 2. Note that DME_{nonpol} slightly overestimates the experimental density (737 kg m^{−3})³⁹ and underestimates the experimental heat of vaporization (21.7 kJ mol^{−1}),⁵⁴ see Table 2. However, the goal is to parametrize a polarizable model which is as close as possible to DME_{nonpol}, in order to get a clear picture of the effect of treating solvent polarization effects explicitly in the study of reaction 3 in DME. The self-polarization energy (U_{self}) contribution to the total potential energy of the polarizable model (DME_{pol}) is only 5%, see Table 2. However, polarization effects will play a more important role in heterogeneous, more polar (ionic) media, as shown below for the case of a changing charge distribution over the solute system along the reaction coordinate. The static relative dielectric permittivities of the DME_{nonpol} and DME_{pol} models are also given in Table 2, with the permittivity being significantly higher for the latter one (6.8 versus 4.7). We could not find experimental data to compare these values with, and for this system of relatively low dielectric permittivity, 200 ps of simulation was enough to obtain convergence.

IV.2. The Use of the Charge-On-Spring Model in a QM/MM-pol Hamiltonian. After implementing the charge-on-spring model in the GROMOS96 code interfaced to ChemShell, we tested the variational character of the combined QM/MM-pol Hamiltonian by monitoring the convergence behavior of the ‘electronic’ energy of the system (energy of the QM electrons plus induced MM dipoles) for configurations taken every 100 steps from simulations for the separate reactants (rc = 0.4 nm) and the transition state (rc = 0.0 nm) in DME_{pol}. We explicitly looked at the convergence of the total electronic energy, since the separate SCF procedures to solve for the QM electronic wavefunction or the positions of the MM charges-on-spring are strictly not variational due to their mutual influence. The QM, MM, and QM/MM energies were followed every QM-MM iteration step. It was found that both the QM electronic and MM self-polarization energy converged within 0.001 kJ mol^{−1} after 2 or 3 steps. No large fluctuations of the separate QM and MM energy terms were observed during the iterative process, indicating that the applied doubly iterative procedure is variational for the investigated system. The convergence

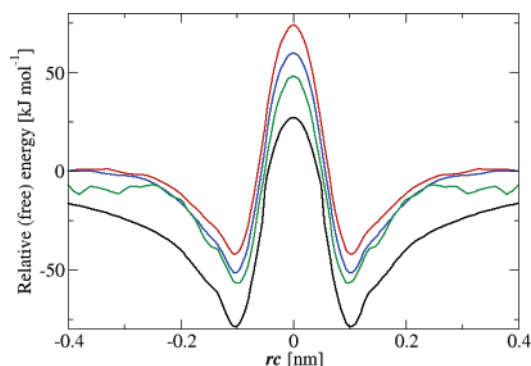


Figure 3. Gas-phase potential energy surface for the S_N2 reaction 3 along the reaction coordinate rc as defined in eq 17 at the PM3 level of theory (black line) and the corresponding free energy profile (green line) and the free energy profile for the same reaction in liquid dimethyl ether from QM/MM MD simulations using a nonpolarizable (blue line) and polarizable (red line) force field for the solvent. Values for the (free) energies are relative to the values for the separated reactants (corresponding with $rc = \pm 1.1$ nm in the gas phase and $rc = \pm 0.4$ nm in the condensed phase).

of the total ‘electronic’ energy of the system to within 0.002 kJ mol⁻¹ is considered to be sufficient: the introduced inaccuracy is negligible in size when compared to the kinetic energy of the atomic degrees of freedom ($1/2 k_B T = 1$ kJ mol⁻¹ at 248.15 K).

In our QM/MM-pol simulations, we did not observe artifacts due to the ‘polarization catastrophe’.³¹ That is, we did not see any induced MM dipole adopting infinitely large values due to close QM-MM van der Waals contacts and accordingly unrealistic large QM-MM contributions to the electric field. Apparently, the size of the van der Waals radii of the QM and MM nuclei was chosen large enough. Additionally, the large value for q_{pol} assures that displacements of the charges-on-spring from their adjacent polarizable center are small compared to the van der Waals radius of the atom, preventing the charges-on-spring to collapse onto the QM or other MM nuclei.

IV.3. The (Free) Energy Profiles for Reaction 3 in the Gas Phase and in Liquid DME. The gas-phase potential energy surface (PES) and the potentials of mean force (PMF) for reaction 3 in vacuum and in liquid dimethyl ether (DME) are presented in Figure 3. The PES and PMFs are given in the range of values for the reaction coordinate rc from -0.4 to 0.4 nm, whereas simulations were only performed at $rc \geq 0.0$ nm. The symmetric profiles were obtained by mirroring the explicitly obtained part of the curves. Plotted values for the potential energy are relative to the value for the infinitely separated reactants, and free energies are relative to the value for which the PMF reaches a plateau.

The gas-phase formation energy of the reactant complex (RC) and transition state (TS) out of the reactants was found to be -79.0 kJ mol⁻¹ and 26.9 kJ mol⁻¹, respectively. These values are in agreement with the values reported by Liu et al. (-18.9 kcal mol⁻¹ and 6.4 kcal mol⁻¹, respectively)⁵ and correspond with a clear picture of a double-well PES. Figure 3 shows that the calculation of the gas-phase PMF suffers from poor sampling in the dissociative regime ($|rc| > 0.2$ nm).

Values for the restraining force are not converged in this regime, since the part of the phase space to be sampled increases due to a weakening of dipole-ion interactions between the separating reactants. Performing tenfold longer MC simulations at selected values for rc between 0.2 and 1.1 nm did not significantly improve convergence (results not shown). Thus, from the gas-phase MC simulations we could not estimate the plateau value for the separate reactants relative to the free energy values for smaller values of rc . Here it is obtained in an alternative way, via relative differences between the free energies of solvation ΔF_{solv} of the reactants Cl⁻, NH₂Cl, and the TS in nonpolarizable DME. ΔF_{solv} was estimated at -109.6 , -12.3 and -110.2 kJ mol⁻¹, respectively. Taking for X and Y in eq 4 the transition and reactant state, respectively, $\Delta\Delta F_{\text{solv},XY}$ was estimated at 11.7 kJ mol⁻¹. Thus, the separated reactants are better solvated in DME, in accordance with Liu’s findings discussed in section I. Using eqs 4 and 5 and a value of $\Delta F_{\text{cond},XY} = 59.8$ kJ mol⁻¹ (see Figure 3), the plateau value for the reactants in the gas phase is found to be 48.1 kJ mol⁻¹ lower than the free energy of the transition state. An estimate of the errors in ΔF_{solv} is difficult to obtain when using the perturbation formula (eq 22). However, these errors do not contribute to $\Delta\Delta F_{\text{solv}}$ profiles along rc when calculated from eq 5, as is done in the next sections.

Figure 3 also presents the free energy profiles of reaction 3 in liquid DME obtained from our QM/MM simulations. These profiles have been corrected for the metric tensor effect (second term on the right of eq 20), which was found to contribute less than 1 kJ mol⁻¹ to free energy differences along rc . Accumulated errors in the reactant complexation free energy and activation barrier were 6.0 and 7.3 kJ mol⁻¹ for the simulations with the QM/MM-nonpol Hamiltonian (which employs the DME_{nonpol} force field) and 8.0 and 9.7 kJ mol⁻¹ for the QM/MM-pol simulations, respectively, which are smaller than the corresponding free energy differences along rc and the relative differences with respect to the gas-phase potential energy values. To check for hysteresis, we redid the series of simulations in opposite direction (changing rc gradually from 0.4 to 0 nm). No noticeable differences between the PMFs for the different pathways were observed with maximum deviations in free energy differences along the reaction profile of 4 kJ mol⁻¹, which is within the estimated errors. Besides, we did not observe a solvent memory effect along the reaction coordinate, as indicated by the observed overlap of the radial distribution functions (rdf) of solvent molecules around the chlorides in the transition state and from a comparison of rdfs for the solvent molecules around the reactants (products) from simulations of the forward and backward reaction (results not shown). From these findings, we conclude that the simulation time and number of data points along the reaction coordinate are chosen sufficiently large to exclude hysteresis effects.

IV.4. The Effect of the Inclusion of Solvent Polarization Effects. Figure 3 shows that the inclusion of solvent electron polarization effects does not change the qualitative picture of a double-well free energy profile for reaction 3 in liquid DME, but the plateau value for the reactants is reached at a

smaller value for rc in the QM/MM-pol simulations than when using the DME_{nonpol} model. This indicates a stronger solvent screening by DME_{pol} of ion-dipole interactions between the separating reactants, in line with its larger dielectric permittivity (Table 2). Quantitative differences in the PMFs obtained from simulations using nonpolarizable or polarizable DME model are relatively small when compared to the change with respect to the PES in vacuum. However, when comparing the condensed-phase PMFs with the gas-phase *free* energy profile, solvent electron polarization effects are found to play a significant role in the change of the reaction profile upon solvation. Changes in the formation free energy of the TS and RC out of the separate reactants are much smaller upon solvation than when comparing the PMF in DME with the PES in vacuum. The increase in activation barrier when going from the gas-phase PES to the PMF in vacuum can be explained from the lower density of rotational and vibrational states of the classical S_N2 transition states, resulting in a loss of entropy upon complexation.⁵⁵ The loss in entropy upon RC and TS formation has a large effect on the gas-phase activation barrier and reduces the difference in activation barrier when going from the gas-phase simulations to the simulations using the DME_{nonpol} model to 11.7 kJ mol⁻¹. The increase in activation barrier when comparing the PMFs in vacuum and in DME_{pol} was estimated at 25.9 kJ mol⁻¹. Thus, we found a doubling of the solvent effect on the free energy barrier upon inclusion of solvent electron polarization effects.

Including solvent polarization effects does not substantially affect the properties of the QM subsystem along the reaction coordinate, such as the geometry of the substrates or its polarization by the solvent. When using either the nonpolarizable or polarizable solvent models, maximum differences of 0.002 nm for average bond distances in the reactive subsystem were observed (results not shown). Mulliken partial charges of the QM atoms along the reaction are shown in Figure 4a. Only in the range of $rc = 0.02$ nm to $rc = 0.2$ nm a small difference in polarization of the complex is found. In this regime, the attacking Cl⁻ has a slightly more negative charge (and the Cl more tightly bound to the nitrogen an accordingly more positive partial charge) in the QM/MM-pol simulations, probably due to solvent back polarization effects. This difference vanishes when going to the separated reactants (in which charge transfer from the NH₂Cl subunit to the chloride anion is not possible) or to the symmetric transition state.

In contrast, significant changes in the polarization of the solvent molecules are observed along rc . The localization of the QM subsystem's negative charge on the separate chloride anion leads to a stronger polarization of the solvent molecules in the reactant state. This is clearly indicated by trends in the total self-polarization energy U_{self} of the solvent. Figure 4b shows an increase of about 20 kJ mol⁻¹ in U_{self} values when going from $rc = 0.0$ to 0.4 nm. Indeed, when comparing the transition and the reactant state we see a significant induction in the radial component of the average molecular dipole moment of the DME solvent molecules around the leaving chloride anion, see Figure 4c. In contrast,

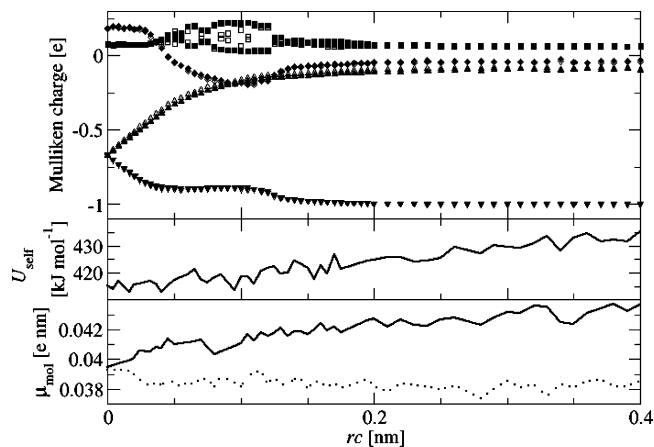


Figure 4. (a) Mulliken charges on atoms of the QM subsystem in liquid dimethyl ether along the reaction coordinate rc of reaction 3 as defined in eq 17, using the QM/MM-nonpol (closed symbols) and QM/MM-pol Hamiltonians (open symbols) (diamonds: nitrogen, squares: hydrogens, triangles up: chloride in NH₂Cl, triangles down: approaching or leaving chloride anion), (b) total self-polarization energy (in kJ mol⁻¹) of the solvent, and (c) radial component of the molecular dipole in the first solvation shell (radius = 0.6 nm) around the chlorides in the QM subsystem (approaching or leaving Cl⁻: solid line, and Cl residing in NH₂Cl product/reactant: dotted line) along rc for reaction 3 in liquid dimethyl ether, in simulations using the QM/MM-pol Hamiltonian.

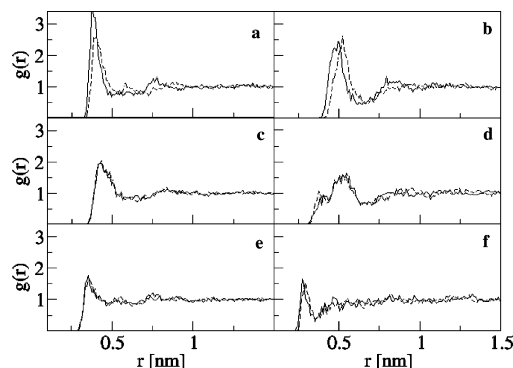


Figure 5. Radial distribution functions of solvent atoms (C, O) around QM solute atoms (Cl⁻, Cl, N) for the separated reactants of reaction 3 (corresponding with a value for the reaction coordinate rc as defined in eq 17 of 0.4 nm) in liquid dimethyl ether from simulations using the QM/MM-nonpol (dashed lines) and the QM/MM-pol (solid lines) Hamiltonians: (a) Cl⁻-C, (b) Cl⁻-O (c) Cl-C, (d) Cl-O, (e) N-C, and (f) N-O atom pairs.

this value only moderately decreases for the DMEs around the chloride atom that remains attached to N.

The stronger induction of the MM molecular dipoles in the reaction state leads to a better solvation of the Cl⁻ anion by the polarizable solvent, when compared to the simulations using the nonpolarizable force field. Figure 5a shows an increase in the first peak of the radial distribution function (rdf) of the carbon of DME around the Cl⁻ upon inclusion of solvent polarization effects in the Hamiltonian. Additionally, the peak slightly shifts to the left, hinting at a closer approach of the DME molecules. This is even more strongly indicated by the shift in the first peak in the Cl⁻-O rdf

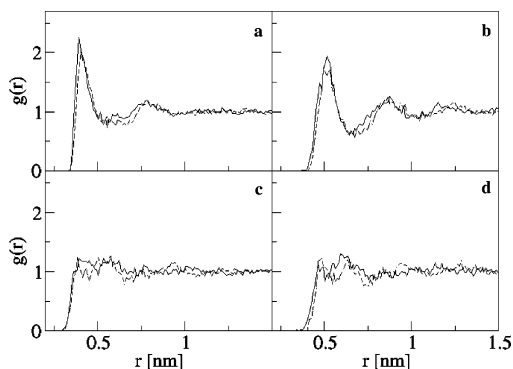


Figure 6. Radial distribution functions of solvent atoms (C, O) around QM solute atoms (Cl, N) for the transition state of reaction 3 (corresponding with a value for the reaction coordinate rc as defined in eq 17 of 0.0 nm) in liquid dimethyl ether from simulations using the QM/MM-nonpol (dashed line) and the QM/MM-pol (solid line) Hamiltonians: (a) Cl–C, (b) Cl–O, (c) N–C, and (d) N–O atom pairs.

(Figure 5b). The difference in shift of the first peaks in the Cl[−]–C and Cl[−]–O rdfs hints at a difference in orientation of the polarizable DME molecules around the chloride anion compared to the DME_{nonpol} molecules. However, this could not be unambiguously confirmed from a comparison of the Cl[−]–C–O angle involving carbons in the first solvation shell around the ion: only a small difference was found in the simulations using DME_{nonpol} and DME_{pol} with values of 123.6 and 125.6 degrees, respectively. The solvent structure around the neutral NH₂Cl substrate is hardly affected upon inclusion of solvent polarization, as shown by the similar shape of the radial distribution functions shown in Figure 5c–f. From the rdfs of the solute–solvent atom pairs for the TS (see Figure 6) we see a slight improvement of TS solvation upon using a polarizable DME force field, as reflected by the small shifts to the left in the first solvation peaks for the Cl–C, Cl–O, and N–O rdfs (Figure 6a,b,d). However, these changes are much smaller than the increase in and shift of the first peak in the Cl[−]–C and Cl[−]–O rdfs for the separate anion (Figure 5a,b). Thus, inclusion of solvent polarization effects leads to a further improvement of the solvation of the charge-separated reactants when compared to the transition state. Together with the enhancement of the dipole of the solvent molecules around the chloride ion making energetic interactions stronger than in the TS, this causes an increase in the activation barrier of reaction 3 in DME when going from a nonpolarizable to a polarizable description of the solvent.

IV.5. An Energy-Entropy Decomposition of the Relative “Free Energy of Solvation” of the QM Subsystem.

Figure 7a shows the differences in the free energy of solvation of the QM subsystem along the reaction coordinate relative to the reactants at infinite separation ($\Delta\Delta F_{\text{solv}}$) in polarizable and nonpolarizable DME, as calculated from Figure 3 using eq 5. This profile quantitatively shows that the separated reactants are better solvated than the reactant complex and transition state (values for $\Delta\Delta F_{\text{solv}}$ are positive over the whole range of rc values) and that this effect is more pronounced in DME_{pol} (values for $\Delta\Delta F_{\text{solv}}$ being more positive). Values for $\Delta\Delta F_{\text{solv}}$ along rc were decomposed into

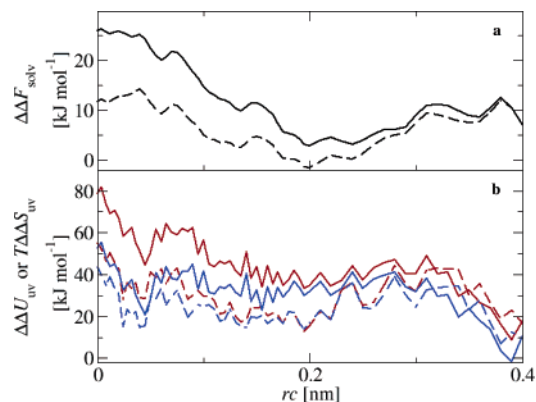


Figure 7. (a) Free energy of solvation relative to the reactants at infinite separation ($\Delta\Delta F_{\text{solv}}$) and (b) the corresponding differences in the solute–solvent interaction energy ($\Delta\Delta U_{\text{uv}}$, red lines) and solute–solvent entropy ($T\Delta\Delta S_{\text{uv}}$, blue lines) of the reactive subsystem along the reaction coordinate rc as defined in eq 17 for reaction 3 in liquid dimethyl ether from simulations using a QM/MM-nonpol (dashed lines) and QM/MM-pol (solid lines) Hamiltonian. $\Delta\Delta F_{\text{solv}}$, $\Delta\Delta U_{\text{uv}}$, and $T\Delta\Delta S_{\text{uv}}$ were calculated using eqs 5 and 10–13.

differences in the energetic ($\Delta\Delta U_{\text{uv}}$) and entropic ($T\Delta\Delta S_{\text{uv}}$) interactions between the solute (QM) and solvent (MM) subsystems relative to the reactants at infinite separation, see Figure 7b. Errors in $\Delta\Delta U_{\text{uv}}$ were estimated as the sum in the errors of the separate terms in eqs 10 or 12 and sum up to 2–3 kJ mol^{−1}.

From Figure 7b, it can be seen that values for the solute–solvent interaction energy with respect to the separate reactants increase in going to the reactant complex and transition state. The trend in $\Delta\Delta U_{\text{uv}}$ being positive over the whole range of rc is counteracted by a gain in entropy upon going to $rc = 0.0$ nm which (from Figures 5 and 6) can be understood in terms of a loss in solvent structure compared to the reactant state. However, the trend in solute–solvent entropy only partly counteracts the one in $\Delta\Delta U_{\text{uv}}$, resulting in $\Delta\Delta F_{\text{solv}}$ being positive over the whole range of rc . This is a quantitative measure of the solute–solvent interaction energy determining the increase in activation free energy upon solvation.

From Figure 7b, the solute–solvent interaction energy also determines the further increase in activation barrier and RC formation free energy when going from the nonpolarizable to the polarizable DME solvent. In the QM/MM-pol simulations, $\Delta\Delta U_{\text{uv}}$ is more positive when rc approaches zero (in line with the simultaneous decrease in U_{self} , see Figure 4b), again only partly counteracted by a further increase in entropy at rc going to zero (which can be explained from the change in the first peak in the Cl[−]–C and Cl[−]–O rdfs when going from the QM/MM-nonpol to the QM/MM-pol simulations, see Figure 5).

Finally, we shortly comment on the use of an atomistic representation of the solvent based on our results. First of all, the use of a continuum-electrostatic model in the gas-phase MC simulations to account for solvent effects in a mean-field manner might yield a condensed-phase PMF similar in shape to the ones obtained in our more expensive QM/MM-(non)pol simulations. However, these simulations

can of course not capture the change of the solvent structure around the QM subsystem (as reflected by Figures 5 and 6) or the change in $\text{Cl}^- - \text{C}_{\text{DME}} - \text{O}_{\text{DME}}$ angle involving DME's carbon in the first solvation shell of the anion (from about 120 to 160 degrees) when going from the reactant to the transition state. Similarly, one may try to correct for the missing solvent electronic polarization effects in the QM/MM-nonpol simulations in an average way, by including the energetic contribution of the polarization of the MM atoms via linear-response theory. This would be less expensive than using the iterative QM/MM-pol approach. However, only in the latter case one can observe changes in the microscopic structure around the reactive subsystem induced by the polarization of the solvent, such as the improved solvation of the Cl^- anion (see Figure 5a,b).

V. Conclusions

In the present work, we repeated a combined QM/MM MD simulation study⁵ on the free energy profile of reaction 3 in liquid dimethyl ether (DME). Additionally, we performed the same set of simulations in which electronic polarization of the solvent was explicitly taken into account using the charge-on-spring (COS) model with the aim of analyzing explicit solvent polarization effects upon the reaction. For this purpose, a COS-based force field for DME was parametrized based on the nonpolarizable parameter set used, and the COS model was implemented into the GROMOS interface to ChemShell using a doubly iterative scheme. The combined Hamiltonian for reaction 3 in polarizable DME was found to behave variationally: under the chosen settings, total energies converge within a few iteration steps. No occurrence of the polarization catastrophe was observed.

Including solvent electronic polarization effects does not change the qualitative picture of the double-well free energy profile of reaction 3 in DME. However, the higher dielectric permittivity of the polarizable solvent results in stronger solvent screening of the interactions between the separated reactants, resulting in a plateau value for the free energy corresponding to the separated reactants at a smaller value of r_c than in the simulations using the nonpolarizable $\text{DME}_{\text{nonpol}}$ force field. Moreover, when compared to the gas-phase potential of mean force (PMF), we find a doubling of the change in activation free energy upon solvation when comparing its value from the simulations using the $\text{DME}_{\text{nonpol}}$ model with those from the QM/MM-pol simulations. This could be explained from a stronger polarization of the polarizable solvent molecules around the reactants than those surrounding the transition state in which the net-charge of the reactive subsystem is more smeared out. This leads not only to stronger QM-MM electrostatic interactions for large r_c values but also to a better solvation of the Cl^- anion when compared to the simulations in $\text{DME}_{\text{nonpol}}$. The origin of the increase in activation barrier upon solvation and upon explicit inclusion of solvent polarization can be understood from a quantitative energy-entropy decomposition of the solute-solvent interactions. According to this analysis, solvent (polarization) effects on the PMF of reaction 3 in DME are driven by changes in the solute-solvent interaction energy along r_c , which are only partly counteracted by the solute-

solvent entropy increase upon loss in solvent structure when going from the reactant to the transition state.

Acknowledgment. The authors want to thank Dirk Bakowies, Salomon Billeter, and Haiyan Liu for stimulating discussions. Financial support from the Max-Planck Gesellschaft, from the National Center of Competence in Research (NCCR) in Structural Biology, and from grant number 200021-109227 of the Swiss National Science Foundation is gratefully acknowledged.

References

- (1) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.
- (2) Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes and Solutions*; John Wiley: New York, 1991.
- (3) Senn, H. M.; Thiel, W. *Top. Curr. Chem.* **2007**, *268*, 173.
- (4) Lin, H.; Truhlar, D. G. *Theor. Chim. Acc.* **2007**, *117*, 185.
- (5) Liu, H. Y.; Müller-Plathe, F.; van Gunsteren, W. F. *Chem. Eur. J.* **1996**, *2*, 191.
- (6) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986**, *7*, 718.
- (7) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700.
- (8) Sherwood, P.; de Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; Billeter, S.; Terstegen, F.; Thiel, S.; Kendrick, J.; Rogers, S. C.; Casci, J.; Watson, M.; King, F.; Karlsen, E.; Sjøvoll, M.; Fahmi, A.; Schafer, A.; Lennartz, C. *J. Mol. Struct. (Theochem)* **2003**, *632*, 1.
- (9) Glukhovtsev, M. N.; Pross, A.; Radom, L. *J. Am. Chem. Soc.* **1995**, *117*, 9012.
- (10) Thompson, M. A.; Schenter, G. K. *J. Phys. Chem.* **1995**, *99*, 6374.
- (11) Thompson, M. A. *J. Phys. Chem.* **1996**, *100*, 14492.
- (12) Bryce, R. A.; Buesnel, R.; Hillier, I. H.; Burton, N. A. *Chem. Phys. Lett.* **1997**, *279*, 367.
- (13) Field, M. J. *Mol. Phys.* **1997**, *91*, 835.
- (14) Dupuis, M.; Aida, M.; Kawashima, Y.; Hirao, K. *J. Chem. Phys.* **2002**, *117*, 1242.
- (15) Vesely, F. J. *J. Comput. Phys.* **1977**, *24*, 361.
- (16) van Belle, D.; Couplet, I.; Prevost, M.; Wodak, S. J. *J. Mol. Biol.* **1987**, *198*, 721.
- (17) Rappe, A. K.; Goddard, W. A. *J. Phys. Chem.* **1991**, *95*, 3358.
- (18) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141.
- (19) Drude, P. *The Theory of Optics*; Longmans, Green, and Co.: New York, 1902.
- (20) Born, M.; Huang, K. *Dynamic Theory of Crystal Lattices*; Oxford University Press: Oxford, UK, 1954.
- (21) Straatsma, T. P.; McCammon, J. A. *Mol. Simul.* **1990**, *5*, 181.
- (22) Yu, H. B.; Hansson, T.; van Gunsteren, W. F. *J. Chem. Phys.* **2003**, *118*, 221.

- (23) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; vdf Hochschulverlag: ETH Zürich, Switzerland, 1996.
- (24) Scott, W. R. P.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596.
- (25) van der Vegt, N. F. A.; van Gunsteren, W. F. *J. Phys. Chem. B* **2004**, *108*, 1056.
- (26) van der Vegt, N. F. A.; Trzesniak, D.; Kasumaj, B.; van Gunsteren, W. F. *Chem. Phys. Chem.* **2004**, *5*, 144.
- (27) Trzesniak, D.; van der Vegt, N. F. A.; van Gunsteren, W. F. *Phys. Chem. Chem. Phys.* **2004**, *6*, 697.
- (28) Lee, M. E.; van der Vegt, N. F. A. *J. Am. Chem. Soc.* **2006**, *128*, 4948.
- (29) van der Vegt, N. F. A.; Lee, M. E.; Trzesniak, D.; van Gunsteren, W. F. *J. Phys. Chem. B* **2006**, *110*, 12852.
- (30) Yu, H. A.; Karplus, M. *J. Chem. Phys.* **1988**, *89*, 2366.
- (31) Yu, H. B.; van Gunsteren, W. F. *Comput. Phys. Commun.* **2005**, *172*, 69.
- (32) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269.
- (33) Billeter, S. R.; Turner, A. J.; Thiel, W. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2177.
- (34) Thiel, W. *MNDO99*, V 6.1 ed.; Max-Planck-Institut für Kohlenforschung: Mülheim an der Ruhr, Germany, 2003.
- (35) Miller, K. J. *J. Am. Chem. Soc.* **1990**, *112*, 8533.
- (36) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
- (37) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.
- (38) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451.
- (39) Maass, O.; Booner, E. H. *J. Am. Chem. Soc.* **1922**, *44*, 1709.
- (40) Neumann, M. *Mol. Phys.* **1983**, *50*, 841.
- (41) Senn, H. M.; Thiel, S.; Thiel, W. *J. Chem. Theory Comput.* **2005**, *1*, 494.
- (42) Beveridge, D. L.; DiCapua, F. M. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431.
- (43) Carter, E. A.; Ciccotti, G.; Hynes, J. T.; Kapral, R. *Chem. Phys. Lett.* **1989**, *156*, 472.
- (44) Ciccotti, G.; Ferrario, M.; Hynes, J. T.; Kapral, R. *Chem. Phys.* **1989**, *129*, 241.
- (45) Paci, E.; Ciccotti, G.; Ferrario, M.; Kapral, R. *Chem. Phys. Lett.* **1991**, *176*, 581.
- (46) Billeter, S. R.; van Gunsteren, W. F. *J. Phys. Chem. A* **2000**, *104*, 3276.
- (47) Sprik, M.; Ciccotti, G. *J. Chem. Phys.* **1998**, *109*, 7737.
- (48) den Otter, W. K.; Briels, W. J. *J. Chem. Phys.* **1998**, *109*, 4139.
- (49) Schlitter, J.; Klahn, M. *J. Chem. Phys.* **2003**, *118*, 2057.
- (50) Schlitter, J.; Klahn, M. *Mol. Phys.* **2003**, *101*, 3439.
- (51) Trzesniak, D.; Kunz, A. P. E.; van Gunsteren, W. F. *Chem. Phys. Chem.* **2007**, *8*, 162.
- (52) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: New York, 1987.
- (53) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420.
- (54) Kennedy, R. M.; Sagenkahn, M.; Aston, J. G. *J. Am. Chem. Soc.* **1941**, *63*, 2267.
- (55) Olmstead, W. N.; Brauman, J. I. *J. Am. Chem. Soc.* **1977**, *99*, 4219.

CT7000123

Basis Set Superposition Error along the Free-Energy Surface of the Water Dimer

Jens Thar,[†] Rainer Hovorka,[‡] and Barbara Kirchner^{*,†}

Lehrstuhl für Theoretische Chemie, Wilhelm-Ostwald Institut für Physikalische und Theoretische Chemie, Universität Leipzig, Linnestrasse 2, D-04103 Leipzig, Germany, and Organische Chemie, Kekulé Institut für Organische Chemie und Biochemie, Universität Bonn, Gerhard-Domagk-Strasse 1, D-53121 Bonn, Germany

Received August 9, 2006

Abstract: In this article we review the behavior of static plane wave basis set calculations in comparison to Gaussian basis set calculations. This was done in the framework of density functional theory for description of hydrogen bonds with the water dimer as an example. Furthermore we carried out molecular dynamics simulations enforcing the self-dissociation reaction of the water dimer to study the influence of the basis set onto the reaction. Not surprisingly, we find strongly varying results of the calculated forces for a chosen cutoff along the reaction coordinates. The basis set superposition errors of the dimer interaction energy are analyzed along the free-energy surface, i.e., along the trajectories. Based on the analysis along the trajectories a qualitative and quantitative estimate depending on the particular point of the free-energy surface can be provided. Namely, at the intermolecular O...H distance close to the equilibrium geometry the errors are smaller than at shorter O...H distances. However, the distribution at the equilibrium distance is more unsymmetrical than the distribution at short distances. It is wider, and the standard deviation is larger than at shorter distances where the basis set superposition error is larger.

1. Introduction

The plane wave basis set (PWBS) combined with pseudo-potentials and density functional theory (DFT) is the standard method used in many first-principles simulations (FPMD).^{1–6} A reason for the heavy use of this combination lies in the nice property of low computational costs together with the advantage of easy technical applicability.⁷ The importance of low computational costs is understandable when keeping in mind that along a trajectory in each step a quantum chemical calculation has to be carried out. For a total simulation time of 10 ps with a time step of 0.1 fs about 100 000 of such calculations are needed.

Basis sets used in standard quantum chemical calculations are usually built up by atom-centered functions.^{8,9}

When comparing interaction energies (E_I) at different geometries the atom-centered basis set (e.g., the Gaussian Basis Set (GBS)) introduces the well studied basis set superposition error (BSSE).¹⁰ The quality of the basis set is not the same at all geometries, owing to the fact that the electron density around one nucleus may be described by functions centered at another nucleus, see Figure 1. The counterpoise correction procedure introduced by Boys and Bernardi is the standard method to correct for the BSSE.¹¹ The BSSE represents a strong disadvantage of the Gaussian basis set and a strong argument in favor of the plane wave basis set especially for simulations of large systems and many electronic structure calculations. Since plane waves are not atom centered functions, such an effect does not appear in calculations employing them. On the other hand, plane wave basis sets have an extended dimension and cannot describe the compact charge densities as accurately as the Gaussian basis set.

* Corresponding author e-mail: bkirchner@uni-leipzig.de.

[†] Universität Leipzig.

[‡] Universität Bonn.

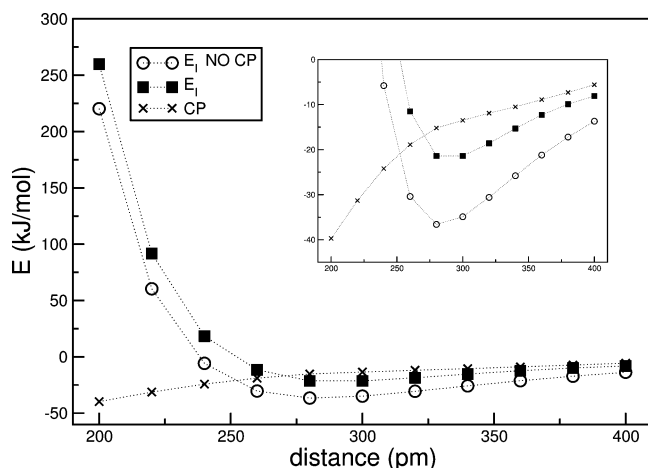


Figure 1. Interaction energy E_1 in kJ/mol not counterpoise corrected (NO CP) (empty circles), counterpoise corrected (filled squares), and counterpoise correction (CP) (crosses) of the water dimer at different $r_{O^*H^*}$ intermolecular distances in pm but otherwise in the global minimum conformation. All calculations were performed with the BP86/SV(P) combination of density functional and basis set.

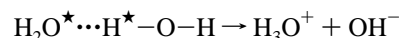
Attempts to combine the advantages of both types of basis sets—GBS as well as PWBS—were successfully carried out for example in the Hutter group.^{12–14} One needs to add here that all plane-wave and mixed basis set calculations can be used in combination with pseudopotentials, for a detailed description see refs 15–17. An approach that allows O(N)-scaling which employs pure Gaussians was introduced by Schlegel and co-workers.¹⁸ A different route is taken by the Tuckerman group where a complete basis set limit for simulations is achieved with a discrete variable representation basis set.^{19–21} While ref 19 is dedicated to the development and implementation of this real-space approach for the electronic structure calculations in FPMD, refs 20 and 21 deal with the simulation of liquid water. It was found by Lee and Tuckerman that less overstructuring in the radial pair distribution functions occurs with this kind of basis set ansatz. In ref 19 the authors also provide an informative overview over alternatives to the PWBS approach in condensed phase. BSSE-free methods which, however, are not described in the context of condensed phase are discussed in the literature, see ref 22 for an overview and some selected examples in refs 23–27. There are more technical issues than the question of the BSSE to investigate when electronic structure is calculated on the fly. A recent article by Kuo is highlighted here as an example.²⁸ In a very important and reliable comparison of the Car–Parrinello method¹ versus the Born–Oppenheimer molecular dynamics simulation technique the authors could show that, despite some beliefs, the Car–Parrinello method and the Born–Oppenheimer simulation technique produce equivalent results.²⁸

Due to the problems associated with FPMD it is worth mentioning that other approaches to describe the liquid-phase such as water^{29–32} exists. Of course there is the wealth of simulations applying empirical water potentials^{33,34} and polarizable force fields.^{35–38} Next to these important methods, Weinhold and Ludwig developed and refined an alternative to treat the liquid state by considering the so-called quantum

cluster equilibrium (QCE) theory.^{39,40} So far many problems using the QCE theory^{39–42} were studied, for example, it was possible to determine the triple point of water and describe an icelike phase of water.⁴¹ Investigating isotopically substituted water⁴² and studying the cooperative versus dispersion effects⁴³ in liquid water was also the scope of the QCE studies. The importance of cooperative effects to be correctly described was also demonstrated by quantum calculations.⁴⁴ Employing pair potentials from ab initio calculations can be considered to be another area of liquid-phase simulations.⁴⁵ For water pioneering work was carried out by the Clementi group.^{46,47} An ab initio constructed force field in order to provide “chemical accuracy” was constructed by Liu and co-workers.⁴⁸ Furthermore the very important and BSSE-free symmetry-adapted perturbation method⁴⁹ was also used to provide ab initio pair potentials for simulations of water.⁵⁰

It can be recognized from Figure 1 that the size of the BSSE strongly depends on the geometry of the water dimer. It would therefore be difficult to work with Gaussian basis sets in simulations. So far no detailed study was published concerning the basis set superposition error in simulations. The reliability of (Gaussian-type) basis sets is usually only tested for static quantum chemical calculations and in the majority of cases for intramolecular properties like bond lengths; see, e.g., refs 51–53. However, the very recent article of Haynes and co-workers is interesting in our context, because the authors showed that the BSSE is eliminated by optimizing the local orbitals in situ using a basis set that is similar to a PWBS.²³

The objective of this article is to try to assess the reliability of simulations as obtained from different PWBS calculations with varying basis set size and to assess the BSSE along the trajectories. We do that by simulating the barrier for the autoprotolysis within a water dimer employing the technique of thermodynamic integration. This means that we force one proton to move from one water molecule to the other water molecule within a distance constraint to form the $\text{OH}^- + \text{H}_3\text{O}^+$ ion-pair:



The reaction coordinate is herewith determined by the stepwise increased hydrogen bond distance $r_{O^*H^*}$. The technique of thermodynamic integration is a standard technique so we recall here only the essential equation. For a detailed discussion see refs 34 and 54. The difference in free energy ΔA is obtained by integrating the negative averaged force $-\langle f \rangle_{r_{O^*H^*}}$

$$-\langle f \rangle_{r_{O^*H^*}} = \frac{\partial A}{\partial r_{O^*H^*}} \quad (1)$$

The self-dissociation of water is not a reaction that takes place under usual condition—standard pressure, temperature, and absent solvent—in the gas phase. Sobolewski and Domcke observed in their study about the hydrated hydronium that the ground state of the dissociation reaction correlates adiabatically with the formation of the $\text{OH}^- + \text{H}_3\text{O}^+$ ion-pair, while the excited state of the water dimer correlates with the biradical $\text{OH}-\text{H}_3\text{O}$ complex.⁵⁵ Their

studies are based on CASPT2 calculations using a modified ANO-L basis set upon B3LYP/6-311++G** structures. They found that the ion-pair configuration does not exist as a minimum on the ground-state potential energy surface. Nevertheless, the observed shallow plateau around 180 pm develops into a local minimum for a larger water cluster. Thus the $\text{OH}^- + \text{H}_3\text{O}^+$ ion-pair is stabilized by solvation. An overview over the present knowledge about the auto-protolysis can be found in refs 56 and 66.

Despite the fact that this reaction does not take place in the gas phase, it serves us to obtain many chemically different situations. During its course, each of this situation is associated with a unique BSSE when a Gaussian type orbital is applied, thus this reaction allows for the study of the BSSE during the course of the simulations. We are able to model the influence of the basis set quality onto the simulations by increasing or decreasing the number of plane waves in a particular calculation. In subsequent quantum chemical calculations applying the “cluster ansatz”^{57,58} we then analyze the BSSE with two particular GBSs along the trajectories. It can be expected that by changing the chemical situation in the system we create situations which are sometimes more and sometimes less affected by the quality of the basis set.

2. Technical Details

The general setup for the simulations was chosen to be the same as for the static PWBS calculations, see the Supporting Information. Molecular dynamics simulations were performed in the NVT ensemble at 300 K using a Nosé–Hoover chain thermostating scheme.^{59–61} All simulations were performed with the CPMD code.⁶² It was shown previously that BLYP provides the best results for liquid water; therefore, our dynamical calculations are mainly done for this functional. Other functionals are only tested at one cutoff.⁶³ The quality of the plane wave basis set was determined by the energy cutoff E_{cut} which we selected to be 20, 50, 70, and 90 Ryd. We simulated each run with a time step of 5 au ($=0.12094$ fs) and with a fictitious mass of 600 au. The total number of time steps per simulation was 50 000, i.e., 6 ps. To obtain the free energy difference for each point such a trajectory was carried out, i.e., we ran in total 45 trajectories. Subsequently to this we calculated along one set (BLYP/70 Ryd with ranging from $r_{\text{O}^*\text{H}^*}=100$ –180 pm) of the obtained trajectories the BSSE, i.e., we carried out 45000 single point quantum chemical calculation for each of the TZVPP and the SVP basis sets, see the Supporting Information.

3. Results

3.1. Static Considerations Revisited. We recall here well-known results for static calculations, because they provide the basis for the dynamical calculations. For the interested reader we provide a full discussion in the Supporting Information.

We start by summarizing the total energy results. For the BLYP functional the basis set limit as shown previously by Lee and Tuckerman²⁰ of 300 Ryd deviates from the value at 150 Ryd only by approximately 5 kJ/mol. We see that independent of the functional and of the water dimer

structures, the calculations are converged within the first digit before the decimal point for $E_{\text{cut}} = 50$ Ryd, with respect to the reference (150 Ryd) calculation. The difference between 50 Ryd and 70 Ryd is thus in the range of 800 kJ/mol, except for the BP86 functional which only shows a difference of 250 kJ/mol. Using the standard cutoff of 70 Ryd, the energies converge within the first digit after the decimal point. This means that the error lowers to about 180 kJ/mol. Choosing the cutoff of 90 Ryd improves the convergence behavior to 0.01 hartree, i.e., an error of about 25 kJ/mol. Therefore we also recommend using a cutoff of 90 Ryd in calculations with systems containing water molecules if computer time is available as it is in general recommended but currently seldomly used. For a qualitative discussion a cutoff of 70 Ryd might be sufficient to capture all important chemical effects. The behavior for all functionals is similar, except that the BP86 values start at lower energies and converge faster.

The difference in interaction energies between the two structures converges to approximately 3 kJ/mol independently of the particular functional. We also see the usual functional dependencies, for example both PBE functionals give stronger binding energies than all other functionals and BP86 yields a higher binding energy than BLYP. Obviously more than one structure should be investigated, because for all values of E_{cut} the global minimum structure yields a reasonable interaction energy. Inspecting the results for a local minimum structure we realize that a cutoff energy of 20 Ryd is leading to absurd results, while cutoff 20 Ryd yields good results for global minimum, see the Supporting Information. From cutoff 70 Ryd and even 50 Ryd on we obtain comparable interaction energies for all chosen cutoffs of a particular functional. This is the reason why we expect cutoff 70 Ryd to capture important “chemical effects” despite the fact that for all chosen cutoffs and some functionals the results are not within “chemical accuracy” which is simply due to density functional theory and has nothing to do with basis set convergence. Similar trends were found by the Hutter group in their study of hybrid functionals applied to water simulations.⁶⁴ The authors calculated interaction energies of 18.16 kJ/mol for BLYP, 19.37 kJ/mol for B3LYP, 19.08 kJ/mol for PBE, and 19.96 kJ/mol for PBE0. The difference of the latter two values to our values are about 3 kJ/mol and might be attributed to the choice of the pseudo-potentials. From our considerations (see the Supporting Information) we advise to compare the convergence behavior of the SVP basis set to a cutoff E_{cut} of 50 Ryd. The TZVP basis set convergence may be compared to $E_{\text{cut}} = 70$ Ryd and the TZVPP to $E_{\text{cut}} = 90$ Ryd. The interaction energies for the global minimum structure calculated with MP2/TZVPP and CCSD(T) in the basis set limit are -19.2 kJ/mol and -20.7 kJ/mol, respectively.⁴³ A recent and high-level correlated (R_{12} method) thus trustworthy value for the water dimer was provided by Klopper et al. with -21.00 kJ/mol.⁶⁵ PBE and PBE0 energies compare best with the CCSD(T) basis set limit values, see the Supporting Information. BLYP provides values that least correspond to the CCSD(T) data. If we compare the stability of the water dimer given by the different functionals, all the PWBS calculations

Table 1. Constraint Force $\langle f \rangle$ in au at Hydrogen Bridge Distance $r_{\text{O}^*\text{H}^*}$ in pm for Different Functionals and Energy Cutoffs E_{cut} in Ryd

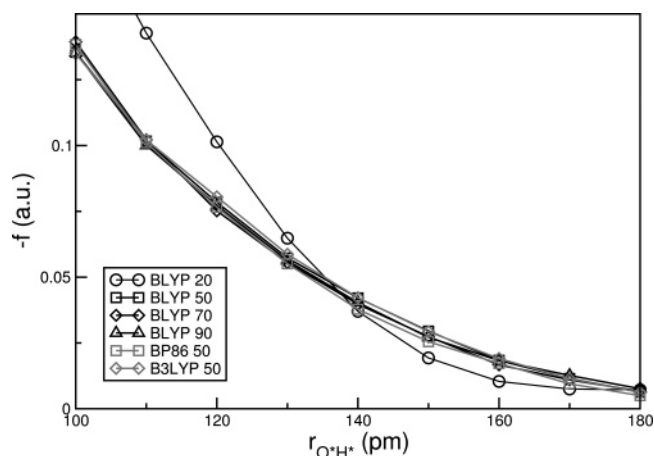
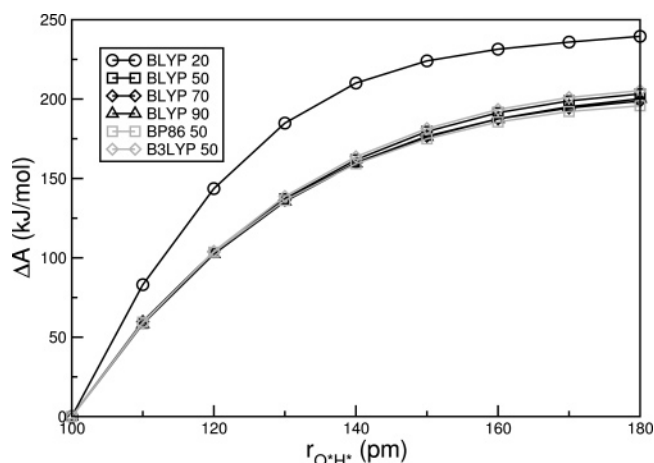
$r_{\text{O}^*\text{H}^*}$	BLYP				BP86 50	B3LYP 50
	20	50	70	90		
100	0.1922	0.1354	0.1394	0.1384	0.1353	0.1349
110	0.1427	0.1017	0.1022	0.1011	0.1001	0.1021
120	0.1014	0.0784	0.0752	0.0766	0.0772	0.0805
130	0.0648	0.0573	0.0555	0.0551	0.0563	0.0587
140	0.0370	0.0421	0.0397	0.0378	0.0404	0.0422
150	0.0193	0.0295	0.0274	0.0255	0.0271	0.0293
160	0.0104	0.0184	0.0167	0.0172	0.0183	0.0191
170	0.0075	0.0115	0.0110	0.0095	0.0127	0.0114
180	0.0072	0.0063	0.0066	0.0050	0.0076	0.0062

as well as GBS results show the same order. The intermolecular bond of the water dimer is strongest using the PBE functional, followed by PBE0, B3LYP, BP86, and BLYP. It is remarkable that the differences in binding energy between $E_{\text{cut}} = 150$ Ryd PWBS calculations and the largest GBS calculations are less than 0.5 kJ/mol for all functionals. Again we want to recall that for all functional the difference between 70 Ryd and 150 Ryd is less (< 1 kJ/mol) and so is the difference between the plane wave basis set results at 150 Ryd and the TZVPPP Gaussian basis set. Importantly, the deviations provided by different functionals (5 kJ/mol) and thus the difference to the benchmark value of Klopper and co-workers⁶⁵ is much larger.

Turning now to the geometry, we find in general that the PWBS geometries do not show trends like distances becoming shorter with a larger basis set. We note that the BP86 geometries independent of the basis set are relatively constant. This is the reason that usually BP86 is preferred over other GGA functionals if reliable structures are sought for. Comparing now TZVPPP and cc-pV5Z with PWBS/150 we find that for all functionals the geometry values agree within 3 pm. The angle varies for the GBS between 2 and 9 degrees. There is one very large angle of 38 degrees for the PBE/SVP combination. The plane wave basis set angles are a bit smaller ranging from 2 to 4 degrees. Again there is no convergence behavior. The difference between PWBSs and GBSs is mostly within 2° .

3.2. Thermodynamic Integration. We now investigate the influence of the basis set quality onto the outcome of the simulations. Table 1 and Figure 3 show the mean absolute constraint values $\langle f \rangle$ for a given distance and at a given E_{cut} obtained from a 6 ps long trajectory.

Please note that the first value for the cutoff of 20 Ryd was obtained at a distance of 103 pm instead of 100 pm. Simulations at 100 pm led to the enforced proton transfer and the subsequent back-transfer of another proton. We recall that at $E_{\text{cut}} = 20$ Ryd the dimer geometries show larger distances than at larger E_{cut} . This means that at 20 Ryd the chemically stable ion H_3O^+ can only be formed at larger distances. Obviously, from this it can be deduced that the chemistry of a system is altered by the basis set. At short enforced distances we see the largest deviations between the calculations with different cutoffs for the PWBS. These are situations where bond cleavage of O^*-H^* occurs, and the

**Figure 2.** The negative constraint force $-\langle f \rangle$ in au for different energy cutoffs with functional BLYP (black) and different functionals with energy cutoff of 50 Ryd (grey) obtained from PWBS trajectories.**Figure 3.** The free energy difference ΔA in kJ/mol of different functionals and different energy cutoffs in Ryd from PWBS molecular dynamics simulations.**Table 2.** Free Energy Difference ΔA in kJ/mol for Different Functionals and Different Energy Cutoffs E_{cut} in Ryd

E_{cut}	BLYP				BP86 50	B3LYP 50
	20	50	70	90		
ΔA	239.5	203.3	198.8	200.2	195.8	205.4

proton is transferred to the second water molecule in order to form the Eigen ion.⁶⁶

Surprisingly, BP86 values at a cutoff of 50 Ryd resemble more the BLYP data at 70 Ryd than the BLYP values at 50 Ryd. The integrated free energy difference ΔA is given in Table 2 and in Figure 3. All chosen cutoff and functional results give free energies around 200 kJ/mol, except for the value obtained with a cutoff of 20 which is 40 kJ/mol above the other, see the first entry in Table 2. The data compare well with the single-minimum-path difference of 220 kJ/mol–230 kJ/mol obtained by Sobolewski and Domcke within these distances. GGA functionals are said to underestimate reaction barriers as compared to exact-exchange functionals.⁶⁷ Comparing the GGA functional at $E_{\text{cut}} = 50$

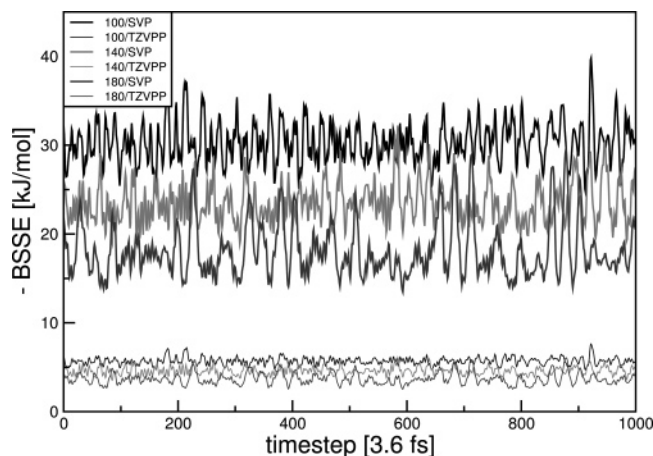


Figure 4. Development of the BSSE with the SVP and the TZVPP basis set along PWBS trajectories of different $r_{O^*H^*}$ distance (100 pm, 140 pm, and 180 pm) constraints.

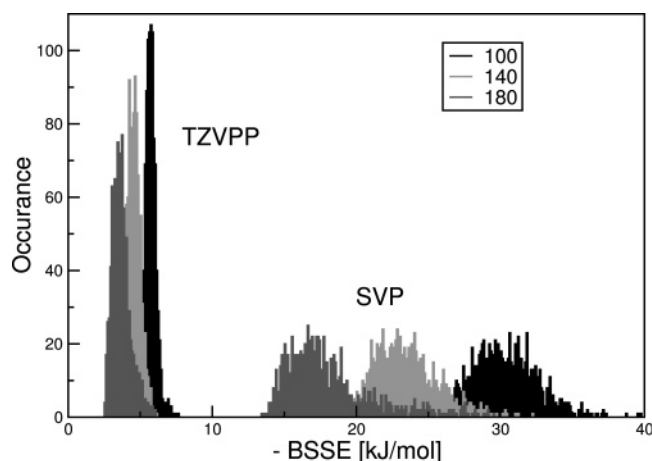


Figure 5. Distribution of the BSSE obtained with the SVP and the TZVPP basis set along PWBS trajectories of different $r_{O^*H^*}$ distance (100 pm, 140 pm, and 180 pm) constraints.

Ryd with B3LYP, we find that the B3LYP functional gives the largest value being approximately 10 kJ/mol above the BP86 data and only 2 kJ/mol above the BLYP free energy difference ΔA .

3.3. BSSE along Trajectories. For every trajectory of the 70 Ryd constraint dynamics 1000 snapshots were taken and the BSSE of each snapshot was calculated. The results are presented in Figures 4 and 5.

The BSSE-development along some trajectories of constraint distance ($r_{O^*H^*}$ =100, 140, and 180 pm) is depicted in Figure 4 for the two basis sets SVP and TZVPP. Obviously the BSSE varies during the course of the simulation. For the SVP basis set along the $r_{O^*H^*}$ = 180 pm trajectory there are several regions where the BSSE changes rapidly from about 14 kJ/mol to over 29 kJ/mol. It is also interesting that its behavior does not appear like fluctuations around an average of about 21 kJ/mol but resembles more a series of fluctuations around an average considerably smaller than 20 kJ/mol with large outliers to higher values. The same pattern can be observed for the TZVPP basis set but with a more confined range of maximum and minimum values due to the overall smaller average BSSE. Because the BSSE calculations for both basis sets are based on the same

Table 3. Statistical Analysis of the BSSE of the TZVPP and SVP Basis Set for Different $r_{O^*H^*}$ (in pm) Distances^a

$r_{O^*H^*}$	TZVPP				SVP			
	$\langle \rangle$	σ	min	max	$\langle \rangle$	σ	min	max
100	5.7	0.41	4.7	7.6	30.3	2.25	24.6	39.8
110	5.5	0.29	4.6	6.7	28.9	1.66	24.5	35.4
120	5.2	0.38	4.3	6.6	27.3	2.11	22.7	34.1
130	4.8	0.43	3.8	6.2	24.9	2.11	20.5	32.5
140	4.6	0.44	3.5	6.1	23.5	2.20	18.9	32.2
150	4.3	0.51	3.1	6.1	21.9	2.66	16.8	30.5
160	4.1	0.46	3.0	6.2	20.4	2.47	16.1	32.0
170	3.8	0.56	2.7	5.7	19.1	2.73	14.5	29.3
180	3.7	0.64	2.5	6.1	18.1	3.02	13.5	29.6

^a $\langle \rangle$: mean value of BSSE; σ : standard deviation; min: minimum value of BSSE; max: maximum value of BSSE. All values are in kJ/mol.

trajectory for each distance, the local maxima and minima of the BSSE are found at the same time step. For the BSSE along the $r_{O^*H^*}$ = 100 pm trajectory large fluctuations are also present; however, these values vary in average around the mean value for the $r_{O^*H^*}$ = 100 pm trajectory. The pattern for $r_{O^*H^*}$ = 140 pm trajectory lies between that of the $r_{O^*H^*}$ = 180 pm and the $r_{O^*H^*}$ = 100 pm trajectories.

In order to investigate the BSSE behavior further, we provide histograms with 0.1 kJ/mol bins of the BSSE data, see Figure 5. Here it is shown more clearly what has been observed before in Figure 4. Obviously the BSSE occurs at larger values for the SVP basis set than for the TZVPP basis set. This is also true for the smaller distances, i.e., for shorter constraint distances the BSSE is larger than for longer constraint values as can be expected also from Figure 1. The fluctuations of the BSSE for the TZVPP basis set is smaller compared to these at the SVP basis set. All distributions are not symmetric with respect to the average value but instead fade out toward higher BSSE errors. This observation is most present for $r_{O^*H^*}$ = 180 pm and least obvious for $r_{O^*H^*}$ = 100 pm. It is also apparent that the distribution of different constraint trajectories is closer for the TZVPP series of BSSE than for the SVP basis set values.

In Table 3 we list the maximum, average, and minimum BSSE and its standard deviation σ obtained from the trajectories at all different values of $r_{O^*H^*}$. We observe a decrease of the BSSE with increasing $r_{O^*H^*}$ from 30 kJ/mol to 18 kJ/mol for the SVP basis set and from 5.7 kJ/mol to 3.7 kJ/mol for the TZVPP basis set. The standard deviation σ being in the 10% range of the mean BSSE value shows the opposite trend, i.e., it rises while the distance enlarges. This nicely corresponds with the behavior reflected in the histogram, see Figure 5, that the distribution fades out for larger distances of $r_{O^*H^*}$. The observations from Figure 4 regarding the deviations of local maxima and minima from the assumed average value can also be quantified: For the SVP basis set the difference between the average BSSE value and the minimal BSSE value decreases from 5.7 kJ/mol to 4.6 kJ/mol while changing from $r_{O^*H^*}$ = 100 pm to $r_{O^*H^*}$ = 180 pm, whereas the difference between the maximal BSSE value and the mean BSSE value increases from 9.5 kJ/mol to 11.5 kJ/mol. The same trend although with smaller values is observed for the TZVPP basis set. Comparing both basis

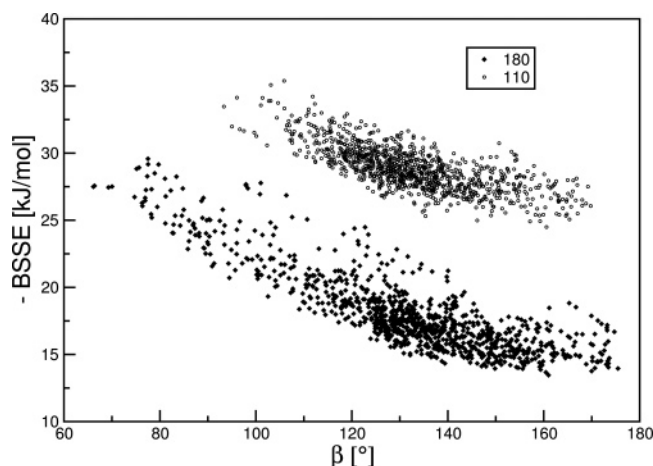


Figure 6. BSSE of the SVP basis set plotted against the angle β depicted in Figure 1 of the Supporting Information obtained from PWBS trajectories at $r_{O^*H^*} = 180$ pm and 110 pm.

sets at a given distance, the average BSSE as well as the standard deviation is considerably smaller for the TZVPP basis set for obvious reasons.

We now select the trajectories with the lowest and highest standard deviation σ , i.e., the trajectories at constraint $r_{O^*H^*} = 110$ pm and $r_{O^*H^*} = 180$ pm to gain further insight into the unsymmetric distribution of the BSSE. Figure 6 depicts the BSSE plotted against the angle β as defined in Figure 1 of the Supporting Information. β is the angle between the hydrogen bond vector and the bisector of the accepting water molecule. We observe two regions of the BSSE for the $r_{O^*H^*} = 180$ pm trajectory, see the filled black diamonds in Figure 6: left and right of $\beta = 145^\circ$. Right of 145° the BSSE behaves almost constant, whereas left of 145° the BSSE increases with smaller values of β . The BSSE of the $r_{O^*H^*} = 110$ pm trajectory shows a similar behavior. However, it completely lacks geometries with $\beta < 95^\circ$, which are responsible for most of the high BSSE values at the $r_{O^*H^*} = 180$ pm trajectory. This behavior provides a qualitative answer for the question why σ rises with increasing constraint distance: In water dimers with larger values of the distance $r_{O^*H^*}$ the hydrogen atoms of the water molecule that accepts the hydrogen bond can still approach the water molecule that donates the hydrogen bond as closely as in dimers of shorter $r_{O^*H^*}$. The average BSSE becomes smaller at $r_{O^*H^*} = 180$ pm, because the centers of the basis sets of the two molecules are further away from each other than for example in geometries of the $r_{O^*H^*} = 110$ pm trajectory. This means that conformations which are sterically unfavorable with regard to β can rather be populated at the $r_{O^*H^*} = 180$ pm trajectory, i.e., conformations where the hydrogen atoms of the accepting water are rather close to the donating water. These conformations then show a comparatively high BSSE, because the centers of the basis functions at the hydrogen atoms are closer to the first molecule than in any other conformation of a given $r_{O^*H^*}$ distance.

The remaining question is why the distribution of the BSSE only fades out into the direction of large BSSE values. This can be understood with the aid of Figure 6. Minimal values of the BSSE are found for geometries with $\beta > 155^\circ$.

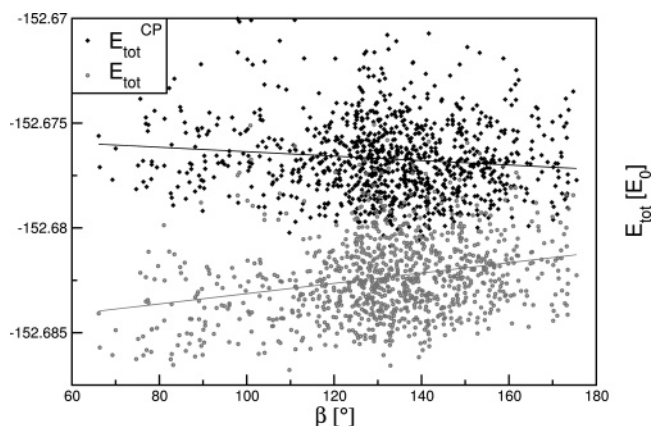


Figure 7. Total energies along the constraint trajectory $r_{O^*H^*} = 180$ pm plotted against the angle β depicted: E_{tot} , uncorrected total SVP energy; E_{tot}^{CP} , counterpoise corrected total SVP energy.

For these angles structures are populated independent of the size of distance $r_{O^*H^*}$. On the other hand angles of $\beta < 90^\circ$ are only populated for the $r_{O^*H^*} = 180$ pm but not for the shorter $r_{O^*H^*} = 110$ pm distance geometries. This shows why also at favorable $r_{O^*H^*}$ distances close to the equilibrium large BSSE values can occur and do occur.

In Figure 7 we plotted the total energies as well as the total energies with subtracted counterpoise corrections against the angle β for the constraint simulation with $r_{O^*H^*} = 180$ pm. The corrected values (black diamonds) are higher in the graph than the uncorrected values (gray circles). The straight lines in Figure 7 show linear regressions in order to clarify the tendencies. It is apparent from Figure 7 that configurations with small distances between atoms of different molecules (represented here by small angles β) would be overpopulated since their energy is more favorable than if correct energies would be calculated, i.e., the BSSE is not a simple shift of the potential energy surface.

4. Discussion and Conclusion

We recalled static calculation results of the energies and geometries for the water dimer and investigated the dynamical behavior of the water self-dissociation dependent on the basis set size and BSSE. We can confirm that the plane wave basis set is able to describe the structure as reasonable as a Gaussian basis set provided that a reliable value for the cutoff is chosen. This is also true for the interaction energies. Pulay, who investigated three small peptide molecules, found similar results with larger deviations between the plane wave basis set and Gaussian basis sets only in the dihedral angle.⁶⁸ In an assessment of the bulk properties of lithium tetraborate, Islam and co-workers found good results for geometries using the PWBS as compared to GBS.⁶⁹ The differences in energetics appeared to be a little more pronounced. More interestingly, Geissler et al. observed in the transition state region of a proton-transfer reaction in $(H_2O)_3H^+$ good agreement between plane wave DFT results and MP2 data using a GBS.⁷⁰ The authors compared these combination of methods and basis sets for geometries as well as frequencies for two transition states and the global minimum structure.

Judging from the behavior of the total energies we recommend cutoff 90 Ryd in general for calculations involving hydrogen bonds. The smaller cutoffs of 50 Ryd and 70 Ryd, however, lead to reasonable structures and interaction energies which is important for a discussion of chemical effects. Comparing PWBSs with GBSs the cutoff 50 Ryd for plane waves provides results similar to the SVP basis set, 70 Ryd similar to the TZVP basis set, and 90 Ryd similar to the TZVPP basis set.

Controlling the accuracy of liquid water simulations is still a delicate issue. There are several parameters changeable to improve the simulations which are the electronic structure method, the quality of the basis set, and the usual molecular dynamics approximations, such as the pairwise additivity.⁴³ Improving just one of these can worsen the previously found results because changing only the method or the basis set can undo the error compensation of fitted approximations for the given system and will then lead away from a realistic behavior of the system. Of course changing all of the variables to amount to more precise results always goes along with additional costs in computer time. This is the reason why compromises with respect to simulation time, system size, and accuracy have to be made. Thereby other errors can be introduced.

Using different cutoffs to obtain the free energy difference by thermodynamic integration, i.e., along several trajectories, shows that a change in the basis set can indeed lead to “different chemical behavior”. It was for example found that the proton transfer at cutoff 20 Ryd occurs already at larger distances than if larger cutoff values are chosen. The analysis of the BSSE along several trajectories revealed that the basis set superposition errors might introduce new problems when a proper trajectory should be calculated, because the errors are not of the same size for different chemical (hydrogen-bonding) situations. For the investigated water dimer the distribution of the BSSE is more unsymmetrical, and most importantly the fluctuations are bigger for larger intermolecular distances than for shorter intermolecular distances. Conformations in which atoms of the two different water molecules approach each other closer than they approach each other in the equilibrium geometry or in structures from the attractive region of the potential energy surface show a higher BSSE. In simulations these structures will be overpopulated when the BSSE plays a role, i.e., geometries with smaller average distances or less probable angles will be sampled more often. This would then lead to overstructured liquids which corresponds indirectly to the observations of Lee and Tuckerman.^{20,21} Based on similar forces of a GAPW and a PWBS simulation for liquid water it was assumed that the BSSE is mostly a shift of the potential energy surface.¹⁷ For our autoprotolysis reaction of the water dimer the opposite was observed. The question whether the basis set superposition error cancels out in a fully solvated reactions remains open. The amount of the BSSE cannot be estimated a priori, and the obtained trajectory might deviate more from the “true” trajectory when using atom-centered basis functions instead of plane waves. This may result in a stronger violation of the ergodic principle.^{34,71}

Acknowledgment. The authors gratefully acknowledge the financial support of the DFG priority program SPP 1191 “Ionic Liquids” and the ERA Chemistry program that allows fruitful collaboration under the project “A Modular Approach to Multi-responsive Surfactant/Peptide (SP) and Surfactant/Peptide/Nanoparticle (SPN) Hybrid Materials”. We furthermore like to acknowledge the financial support under the collaborative research center SFB 624 “Templates” at the University of Bonn. Computational time is gratefully acknowledged from the NIC supercomputers in Jülich. J.T. thanks funding by a “Chemiefonds-Stipendium” of the Fonds der Chemischen Industrie.

Supporting Information Available: Plane waves, computational details, static calculations, Tables 1–8, and Figures 1 and 2. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.
- (2) Marx, D.; Hutter, J. *Ab Initio Molecular Dynamics: Theory and Implementation*. In *Modern Methods and Algorithms of Quantum Chemistry*; Grotendorst, J., Ed.; John von Neumann Institute for Computing, Forschungszentrum Jülich: Jülich, 2000. See <http://www.theochem.rub.de/go/cprev.html> (accessed May 1, 2007).
- (3) Thar, J.; Reckien, W.; Kirchner, B. *Top. Curr. Chem.* **2007**, *268*, 133.
- (4) Carloni, P.; Röthlisberger, U.; Parrinello, M. *Acc. Chem. Res.* **2002**, *35*, 455.
- (5) Colombo, M. C.; Guidoni, L.; Laio, A.; Magistrato, A.; Maurer, P.; Piana, S.; Röhrig, U.; Spiegel, K.; Sulpizi, M.; Vondele, J. V.; Zumstein, M.; Röthlisberger, U. *Chimia* **2002**, *56*, 13.
- (6) Röhrig, U. F.; Guidoni, L.; Laio, A.; Frank, I.; Röthlisberger, U. *J. Am. Chem. Soc.* **2004**, *126*, 15328.
- (7) Hutter, J.; Curioni, A. *Chem. Phys. Chem.* **2005**, *6*, 1788.
- (8) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry*; Dover: New York, U.S.A., 1996; p 480.
- (9) Jensen, F. *Introduction to Computational Chemistry*; Wiley-VCH: Chichester, 2002; p 624.
- (10) van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Lenthe, J. H. *Chem. Rev.* **1994**, *94*, 1873.
- (11) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.
- (12) Lippert, G.; Hutter, J.; Parrinello, M. *Mol. Phys.* **1997**, *92*, 477.
- (13) Lippert, G.; Hutter, J.; Parrinello, M. *Theor. Chem. Acc.* **1999**, *103*, 124.
- (14) Blöchl, P. E. *Phys. Rev. B* **1994**, *50*, 17953.
- (15) Krack, M.; Parrinello, M. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2105.
- (16) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Comput. Phys. Commun.* **2005**, *167*, 103.
- (17) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Comp. Phys. Comm.* **2005**, *167*, 103.

- (18) Schlegel, H. B.; Millam, J. M.; Iyengar, S. S.; Voth, G. A.; A. D. D.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **2001**, *114*, 9758.
- (19) Liu, Y.; Yarne, D. A.; Tuckerman, M. E. *Phys. Rev. B* **2003**, *68*, 125110.
- (20) Lee, H. S.; Tuckerman, M. E. *J. Phys. Chem. A* **2006**, *110*, 5549.
- (21) Lee, H. S.; Tuckerman, M. E. *J. Chem. Phys.* **2006**, *125*, 154507.
- (22) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic-Structure Theory*; Wiley-VCH: Chichester, 2002; p 938.
- (23) Haynes, P. D.; Skylaris, C.-K.; Mostofi, A. A.; Payne, M. C. *Chem. Phys. Lett.* **2006**, *422*, 345.
- (24) Bende, A.; Knapp-Mohammady, M.; Suhai, S. *Int. J. Quantum Chem.* **2003**, *92*, 152.
- (25) Salvador, P.; Asturiol, D.; Mayer, I. *J. Comput. Chem.* **2006**, *27*, 1505.
- (26) Famulari, A.; Gianinetti, E.; Raimondi, M.; Sironi, M. *Int. J. Quantum Chem.* **1998**, *69*, 151.
- (27) Famulari, A.; Raimondi, M.; Sironi, M.; Gianinetti, E. *Chem. Phys.* **1998**, *232*, 275.
- (28) Kuo, I.-F. W.; Mundy, C. J.; McGrath, M. J.; Siepmann, J. I. *J. Chem. Theory Comput.* **2006**, *2*, 1274.
- (29) Ludwig, R. *Angew. Chem., Int. Ed. Engl.* **2006**, *45*, 3402.
- (30) Ludwig, R. *Chem. Phys. Chem.* **2007**, *8*, 44.
- (31) Ludwig, R.; Paschek, D. *Chem. Unserer Zeit* **2005**, *39*, 164.
- (32) Ludwig, R.; Paschek, D. *Angew. Chem., Int. Ed. Engl.* **2001**, *40*, 1809.
- (33) Rahman, A.; Stillinger, F. *J. Chem. Phys.* **1971**, *55*, 3336.
- (34) Frenkel, D.; Smit, B. *Understanding Molecular Simulation – From Algorithms to Applications*; Academic Press: San Diego, 2002; p 638.
- (35) Halgren, T. A.; Damm, W. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236.
- (36) Yu, H. B.; Hansson, T.; van Gunsteren, W. *J. Chem. Phys.* **2003**, *118*, 221.
- (37) Yu, H. B.; van Gunsteren, W. *J. Chem. Phys.* **2004**, *121*, 9549.
- (38) Yu, H. B.; van Gunsteren, W. *Comput. Phys. Commun.* **2005**, *172*, 69.
- (39) Weinhold, F. *J. Chem. Phys.* **1998**, *109*, 367.
- (40) Weinhold, F. *J. Chem. Phys.* **1998**, *109*, 373.
- (41) Ludwig, R.; Weinhold, F. *J. Chem. Phys.* **1999**, *110*, 508.
- (42) Ludwig, R.; Weinhold, F. *Z. Phys. Chem.* **2002**, *216*, 659.
- (43) Kirchner, B. *J. Chem. Phys.* **2005**, *123*, 204116.
- (44) Znamenskiy, V. S.; Green, M. E. *J. Chem. Theory Comput.* **2007**, *3*, 103.
- (45) Huber, H.; Dyson, A.; Kirchner, B. *Chem. Soc. Rev.* **1999**, *28*, 121.
- (46) Matsuoka, O.; Clementi, E.; Yoshimine, M. *J. Chem. Phys.* **1976**, *64*, 1351.
- (47) Lie, G. C.; Clementi, E.; Yoshimine, M. *J. Chem. Phys.* **1976**, *64*, 2314.
- (48) Liu, Y. P.; Kim, K.; Berne, B. J.; Friesner, R. A.; Rick, S. W. *J. Chem. Phys.* **1998**, *108*, 4739.
- (49) Mas, E. M.; Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; Wormer, P. E. S.; van der Avoird, A. *J. Chem. Phys.* **2000**, *113*, 6687.
- (50) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; van der Avoird, A. *J. Chem. Phys.* **2006**, *125*, 044301.
- (51) Helgaker, T.; Gauss, J.; Jørgensen, P.; Olsen, J. *J. Chem. Phys.* **1997**, *106*, 6430.
- (52) Bak, K. L.; Gauss, J.; Jørgensen, P.; Olsen, J.; Helgaker, T.; Stanton, J. F. *J. Chem. Phys.* **2001**, *114*, 6548.
- (53) Pawłowski, F.; Halkier, A.; Jørgensen, P.; Bak, K. L.; Helgaker, T.; Klopper, W. *J. Chem. Phys.* **2003**, *118*, 2539.
- (54) Sprik, M.; Ciccotti, G. *J. Chem. Phys.* **1998**, *109*, 7737.
- (55) Sobolewski, A. L.; Domcke, W. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4.
- (56) Ludwig, R. *Angew. Chem., Int. Ed. Engl.* **2003**, *42*, 258.
- (57) Hermansson, K.; Knuts, S.; Lindgren, J. *J. Chem. Phys.* **1991**, *95*, 7486.
- (58) Eggenberger, R.; Gerber, S.; Huber, H.; Searles, D.; Welker, M. *J. Chem. Phys.* **1992**, *97*, 5898.
- (59) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511.
- (60) Martyna, G. J.; Klein, M. L.; Tuckerman, M. E. *J. Chem. Phys.* **1992**, *97*, 2635.
- (61) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.
- (62) CPMD V3.8; Copyright IBM Corp. 1990–2003, Copyright MPI für Festkörperforschung Stuttgart 1997–2001. See also www.cmpd.org (accessed May 1, 2007).
- (63) Sprik, M.; Hutter, J.; Parrinello, M. *J. Chem. Phys.* **1996**, *105*, 1142.
- (64) Todorova, T.; Seitsonen, A. P.; Hutter, J.; Kuo, I.-F. W.; Mundy, C. J. *J. Phys. Chem. B* **2006**, *110*, 3685.
- (65) Klopper, W.; van Duijneveldt-va de Rijdt, J. G. C. M.; van Duijneveldt, F. B. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2227.
- (66) Kirchner, B. *ChemPhysChem* **2007**, *8*, 41.
- (67) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: Weinheim, 2000; p 528.
- (68) Pulay, P.; Saebo, S.; Malagoli, M.; Baker, J. *J. Comput. Chem.* **2005**, *26*, 599.
- (69) Islam, M. M.; Maslyuk, V. V.; Bredow, T.; Minot, C. *J. Phys. Chem. B* **2005**, *109*, 13597.
- (70) Geissler, P. L.; Van Voorhis, T.; Dellago, C. *Chem. Phys. Lett.* **2000**, *324*, 149.
- (71) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, 1987; reprinted 1990; p 408.

Theoretical Studies on Optical and Electronic Properties of Propionic-Acid-Terminated Silicon Quantum Dots

Q. S. Li,[†] R. Q. Zhang,^{*,†} T. A. Niehaus,^{‡,§} Th. Frauenheim,[‡] and S. T. Lee[†]

Centre of Super-Diamond and Advanced Films (COSDAF) and Department of Physics and Materials Science, City University of Hong Kong, Hong Kong SAR, China, Bremen Center for Computational Material Science, University Bremen, 28334 Bremen, Germany, and Department of Molecular Biophysics, German Cancer Research Center, D-69120 Heidelberg, Germany

Received February 19, 2007

Abstract: The origin and stability of photoluminescence (PL) are critical issues for silicon nanoparticles to be used as biological probes. Optical and electronic properties of propionic-acid (PA)-terminated silicon quantum dots (SiQDs) were studied using the density-functional tight-binding method. We find that the adsorbed PA molecules slightly affect the structure of silicon core. The PA adsorption does not change the optical properties of SiQDs, while it substantially decreases the ionization potentials in the excited state and results in some new active orbitals with adjacent energies around the Fermi energy level. Accordingly, the modified surface of SiQDs can serve as a reaction substrate to oxygen and solvent molecules, which is responsible for the increase in both PL stability and water solubility.

1. Introduction

Silicon is the leading semiconductor material in microelectronic industry.¹ At nanometer sizes, a very important feature of the material is the enormous surface area to volume ratio.² Modification or functionalization of the nanoscale silicon surface might open the possibility of integrating solid-state electronics with optical sensing techniques.^{3–6}

Due to the well-known quantum confinement effect, silicon quantum dots (SiQDs) possess novel optical properties that can potentially be exploited to make optoelectronic devices^{7–11} and new biological chromophores.^{12–15} Unfortunately, this promise is hampered by the easiness of surface oxidation of free-standing SiQDs. It is found that hydrogen-terminated SiQDs exhibit strong PL in the blue region of visible spectrum, but their surface oxidized easily at room temperature,¹⁶ and this oxide surface passivation leads to a dipole-forbidden yellow-red emission.¹⁷

Besides good photoluminescence (PL) stability, good water solubility is also essentially required for silicon nanocrystals

to be employed for bioimaging.¹⁸ Actually, the hydrogenated SiQDs have poor dispersibility in water and some other common solvents. For this reason, much experimental work^{9,19–28} has been carried out on surface functionalization of SiQDs.

For instance, Warner and co-workers²⁰ attached allylamine to silicon particles using a Pt catalyst. The allylamine-capped SiQDs are water-soluble and exhibit strong blue PL with a rapid rate of recombination. In contrast, Li and Ruckenstein^{21,22} added acrylic acid on the SiQDs surface by a UV-induced graft method and alleged the potential of produced propionic-acid (PA)-terminated SiQDs as biological staining agents with tremendous photostability. The high density of carboxylic acid moieties can be used to covalently immobilize molecules containing amines groups, such as proteins.^{29,30} In a subsequent contribution, water-dispersible PA-terminated SiQDs were prepared by photoinitiated hydrosilylation.²³ Sato and Swihart²³ demonstrated that the strongpoint of their experiment is that the SiQD size and corresponding PL emission color changing continuously from yellow to green could be controlled by varying the etching time, while other water-dispersible particles in previous reports^{20–22} only exhibit a single emission color.

* Corresponding author e-mail: aprqz@cityu.edu.hk.

[†] City University of Hong Kong.

[‡] University Bremen.

[§] German Cancer Research Center.

As stated above, the surface modification of SiQDs has been the subject of numerous experimental investigations. Yet, as far as we know, there are only a few theoretical studies^{31–35} on electronic structure and optical properties of functional group-terminated SiQDs up to date. Early in 1996, one of the present authors and his co-workers have pointed out the importance of surface saturation on the stability of silicon nanostructures.³¹ In 2005, Reboredo and Galli³⁴ reported that steric repulsion is dominant in determining the stability of alkyl-passivated clusters. It is also concluded that alkyl passivation weakly affects optical gaps of SiQDs, while it substantially decreases ionization potentials and electron affinities.³⁴ Recently, the excited-state properties of allylamine-capped SiQDs have been theoretically studied in our group.³⁵ The calculation results verified that allylamine is a good protecting molecule, as it reduces the surface oxidation possibility and maintains optical properties of SiQDs in the visible region.³⁵ However, the physical mechanism and chemical nature responsible for the optical properties of water-soluble luminescent SiQDs are far from fully understood and still need further investigation.

In this work, a systematic theoretical study on the electronic structures and optical properties of PA-terminated SiQDs is presented as a function of adsorbed PA amounts. The obtained results will be compared with related experimental^{19,28} and theoretical^{31–35} results. We expect that the present work could reveal the main changes in structure and properties induced by surface modifications with organic molecules and thus provide new insights and guidance to the experimentalists. Particular efforts will be made to characterize the physical mechanism responsible for the optical properties, especially the luminescence, because the application of nanoparticles depends on their luminescence upon insertion into biological cells.

2. Computational Details

In this study, the self-consistent charge density-functional-based tight-binding approach, SCC-DFTB, and its time-dependent linear response extension TD-DFTB were employed to study the electronic and optical properties of selected hydrogenated and PA-terminated SiQDs. The DFTB method has been described in detail elsewhere^{36–40} and will be outlined here only briefly.

The SCC-DFTB model was derived from a second-order expansion of the density functional theory (DFT) total energy functional with respect to the charge-density fluctuations, and the Hamiltonian matrix elements are calculated with a two-center approximation, which are tabulated together with the overlap matrix elements with respect to the interatomic distance.^{36,37}

The TD-DFTB method³⁸ following the TD-DFT route of Casida^{39,40} is capable of efficiently handling excited-state calculations of large systems. In excited-state energy calculations, a self-consistent field (SCF) calculation is conducted first to obtain the single-particle Kohn–Sham (KS) orbitals and the corresponding KS energies ϵ_i . Then, a coupling matrix which gives the response of the SCF potential with respect to a change in the electronic density is obtained as follows

$$K_{ij\sigma,kl\tau} = \sum_{\alpha\beta} q_{\alpha}^{ij} q_{\beta}^{kl} [\gamma_{\alpha\beta} + (2\delta_{\sigma\tau} - 1)m_{\alpha\beta}]$$

where δ and m respectively represent the charge-density fluctuations and the magnetization; σ and τ are spin indices; q represents the Mulliken charge; i and k are indices of the occupied KS orbitals, whereas j and l are unoccupied ones. The exchange-correlation energy has been included in the γ and m . The excitation energy (ω_i) is obtained by solving the following eigenvalue problem

$$\sum_{ij\sigma} [\omega_{ij}^2 \delta_{ik} \delta_{jl} \delta_{\sigma\tau} + 2\sqrt{\omega_{ij} K_{ij\sigma,kl\tau}} \sqrt{\omega_{kl}} F_{ij\sigma}^j] = \omega_l^2 F_{kl\tau}^l \quad (\omega_{ij} = \epsilon_j - \epsilon_i)$$

where F denotes a normalized spherical density fluctuation. The total energy of the excited state is given as a sum of ground-state energy E_{GS} and the excitation energy ω_l :

$$E_{\Sigma} = E_{GS} + \omega_l$$

According to Kasha's rule, optical emissions always occur from the lowest state. For all the SiQDs studied here, the lowest singlet–singlet transition is optically allowed. Therefore, attention will be paid to the structure changes and properties related to the first singlet excited state (S_1 state). In the following section, the excited state refers to the S_1 state except when otherwise stated.

Additionally, in the present work, we used a basis of numerically described s, p, and d atomic orbitals for Si atoms, s and p atomic orbitals for C, N, and O atoms, and an s atomic orbital for H atoms.

To validate the reliability of the DFTB method, test calculations were performed for Si_5H_{12} and $Si_{35}H_{36}$. Our calculated absorption gap of Si_5H_{12} (6.40 eV) is close to the experimental value (6.5 eV)⁴¹ as well as other high-level ab initio results.⁴² For $Si_{35}H_{36}$, our optical gap (4.37 eV) compares well with MR-MP2 result (4.33 eV).⁴² The above tests indicate that the accuracy of TD-DFTB is comparable with the high-level ab initio calculations to study the silicon nanostructures.

3. Results and Discussion

3.1. Geometrical Structures. The optimized ground-state geometries of hydrogenated and PA-terminated SiQDs are shown in Figure 1. We chose $Si_{35}H_{36}$ as the initial model because its diameter 1.1 nm is close to the experimental value,^{20–23} and numerous related theoretical studies^{17,32,35,43–50} have used the same model. Thus, it is convenient to compare the obtained results, such as structural parameters and optical gaps, with the corresponding values in previous experimental or theoretical studies. Our calculations confirmed that ground-state $Si_{35}H_{36}$ is in T_d symmetry. The Si–Si bond lengths are about 2.33–2.37 Å, and the bond lengths of the inner atoms are a little longer than those on the surface. The Si–H bond length is about 1.50 Å, which is in good agreement with the experimental value 1.48 Å.⁵¹

Like allylamine-capped SiQDs,³⁵ PA-terminated SiQDs favor adopting high symmetries, such as D_{2d} , S_4 , and C_2 . We chose to maximize the distance between passivants at

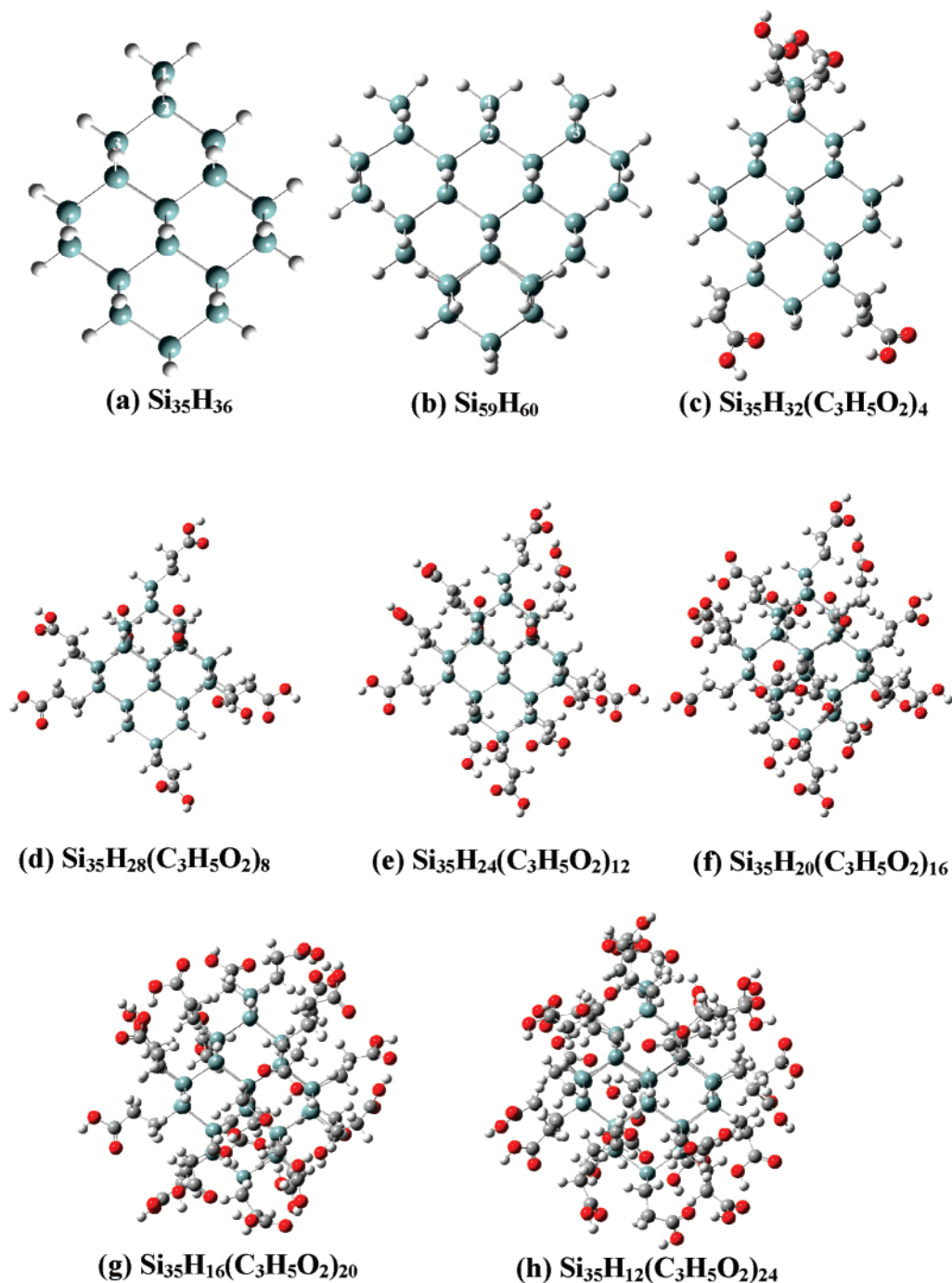


Figure 1. The optimized ground-state structures of SiQDs. The silicon, carbon, hydrogen, and oxygen atoms are cyan, gray, white, and red, respectively. The three different adsorption positions are labeled by number 1, 2, or 3 in (a) and (b).

the surface, to minimize steric repulsions. The optimized symmetrical structures were confirmed to be minimum points by vibrational frequency calculations that give all real frequencies. In the optimized structures, the amount of adsorbed PA molecules varies from 0 to 24. Note that there are likely many isomers for partly PA-terminated SiQDs, and the structures we discussed are the most minimum points in energy among all the isomers. We attempted to optimize the structures of fully PA-terminated SiQDs, that is, $\text{Si}_{35}(\text{C}_3\text{H}_5\text{O}_2)_{36}$, but the minimum-point search does not converge at all. This indicates that it is very difficult to make the surface of $\text{Si}_{35}\text{H}_{36}$ fully PA-terminated due to steric hindrance. This supposition was confirmed by FTIR spectroscopy

copy²³ of PA-terminated SiQDs which showed that surface oxidation occurred during ultrasonication. In addition, our result is consistent with the previous theoretical finding that steric repulsion prevents full alkyl passivation of SiQDs with unreconstructed surfaces.³⁴ Note that the Si–Si or Si–H bond length of PA-terminated SiQDs is slightly longer than that in $\text{Si}_{35}\text{H}_{36}$, which indicates that the PA adsorption causes only small changes on the geometrical structure of the silicon core.

The most striking change in structure is usually associated with electronic excitation. Upon excitation, the T_d symmetry of $\text{Si}_{35}\text{H}_{36}$ cluster is broken down, resulting in a distortion in structure. The distortion leads to some Si–Si bonds

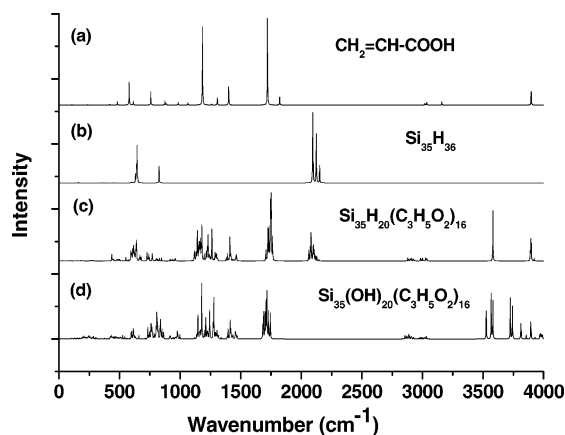


Figure 2. Calculated IR spectrum of acrylic acid, $\text{Si}_{35}\text{H}_{36}$, $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$, and $\text{Si}_{35}(\text{OH})_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$.

increasing, while most of the Si–H bonds remain unchanged. The largest increase of Si–Si distance is up to 0.43 Å. Our calculations alleged that both hydrogenated and PA-terminated SiQDs are in C_1 symmetry in the S_1 state. Similarly, Luppi et al.⁴³ had verified that electronic and geometrical properties of silicon nanoclusters obtained by keeping the symmetry constraint for excited-state calculations are far from the actual energy minimum and lead to a wrong geometry and charge density of the excited state.

In order to gain more information about PA-terminated SiQDs, we simulated the IR spectrum of acrylic acid, $\text{Si}_{35}\text{H}_{36}$ and $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$ (see Figure 2). For $\text{Si}_{35}\text{H}_{36}$, the peaks that appeared ranging from 2097 cm^{-1} to 2153 cm^{-1} are attributed to the strong Si–H stretching vibration. The Si–H bending mode leads to a very obvious peak at 644 cm^{-1} . Experimentally, these two vibrational modes were found at 2085 cm^{-1} and 631 cm^{-1} , respectively.⁵¹ The peak at 827 cm^{-1} is attributed to the scissor vibration of the H–Si–H group. The absorption positions of Si–Si bonds are at the side lower than 500 cm^{-1} , with very weak peak intensity. The bonding of PA on the surface of the SiQDs is reflected by the peak at about 1250 cm^{-1} for the Si–CH₂ stretching vibration. The symmetric and asymmetric vibrations of C–CH₂ and C–COOH lead to the absorption between 2700 cm^{-1} and 3900 cm^{-1} . As shown in Figure 2(c), the vibrational spectrum of $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$ is similar to the experimental spectrum,²³ except for the Si–H characteristic absorption peaks around 2100 cm^{-1} and 650 cm^{-1} . Considering the space steric effects, it is impossible to substitute all surface H atoms with PA molecules, so the absence of vibrational absorption of Si–H bonds in experiments possibly results from the oxidation of a small amount of unsubstituted H atoms. Then we simulated and showed the vibrational spectra of $\text{Si}_{35}(\text{OH})_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$ in Figure 2(d), which is in reasonable agreement with the experimental result.²³ This supports the conclusion that the PA adsorption decreases the oxidation rate of the SiQDs surface.

3.2. Optical Properties. According to the topological structure of $\text{Si}_{35}\text{H}_{36}$, there are three different positions on its surface (denoted by 1, 2, and 3 in Figure 1(a)). For comparison, the binding energies, HOMO–LUMO energy gaps, absorption energies of the SiQDs, and the net Mulliken charges of the PA molecule adsorbed on different positions

Table 1. Binding Energies, HOMO–LUMO Energy Gaps, and Absorption Energies of Single PA-Terminated SiQDs on Different Adsorption Positions of $\text{Si}_{35}\text{H}_{36}$ or $\text{Si}_{59}\text{H}_{60}$ ^a

adsorption position	$E(\text{binding})$ (eV)	$E(\text{HOMO-LUMO})$ (eV)	$E(\text{absorption})$ (eV)	Mulliken charge of PA
$\text{Si}_{35}\text{H}_{36}$				
1	2.194	4.306	4.345	−0.07
2	2.132	4.269	4.330	−0.06
3	2.155	4.287	4.299	−0.06
$\text{Si}_{59}\text{H}_{60}$				
1	4.428	3.549	3.599	−0.07
2	4.396	3.550	3.594	−0.06
3	4.394	3.551	3.591	−0.06

^a The different adsorption positions are labeled in Figure 1. The net Mulliken charges of the PA branch are also presented.

are presented in Table 1. In detail, there is a little amount of negative charge on the PA branch chain due to its receiving electrons from the silicon cluster. This can be ascribed to the moderately higher electronegativity of carbon (2.55) compared to silicon (1.90).⁵³ With the expansion of the silicon core from Si_{35} to Si_{59} , the HOMO–LUMO gap and absorption gap decrease by about 0.7–0.8 eV due to the quantum confinement effect, while the data are still independent of the adsorption position. In general, there is no obvious difference found in the binding energy, HOMO–LUMO energy gap, absorption energy, and net Mulliken charge of the branch. This indicates that the SiQDs are nearly isotropic, that is, the adsorption positions affect the optical properties to such a small degree that they could be ignored in the following study.

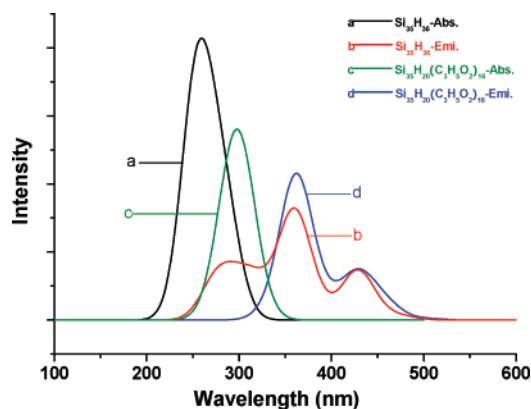
Identifying the first allowed optical transition in the case of large clusters is an important but difficult task, because in this case the absorption and emission spectra became quasicontinuous. In many cases, theoretical calculations do not evaluate oscillator strength and cannot explicitly identify optically allowed and dark transition. In the references where the oscillator strength is considered, the optical gap is usually defined as the point at which the integrated oscillator strength is nonzero or exceeds a threshold value. In the present work, we set the threshold value at 10^{-4} of the total oscillator strength. The chosen value of 10^{-4} stands above the level of numerical “noise” but is sufficiently small as to not suppress the experimentally detectable dipole-allowed transitions. The same definition for the optical gap has been used in previous experimental⁵⁴ and theoretical⁴⁵ work. Besides the optical gap, we pay special attention to the maximal peak position on the absorption or emission spectrum, since in many cases the maximal peak position is not corresponding to the optical gap but evident on the spectrum.

Table 2 presents our calculated HOMO–LUMO energy gaps, the maximal peak positions of absorption and emission, and corresponding oscillator strengths for hydrogenated and PA-terminated SiQDs. In Table 2, we can see that our calculated absorption energy for $\text{Si}_{35}\text{H}_{36}$ is 4.37 eV (283.8 nm), which is close to the recently reported TDDFT/B3LYP value 4.42 eV⁴⁸ and MR-MP2 result 4.33 eV.⁴² After surface modification, the HOMO–LUMO energy gap decreases remarkably with an increase in the adsorbed molecules in ground-state or excited-state configuration, while the emis-

Table 2. HOMO–LUMO Energy Gaps,^a the Maximal Absorption and Emission Wavelengths, and the Oscillator Strengths^b

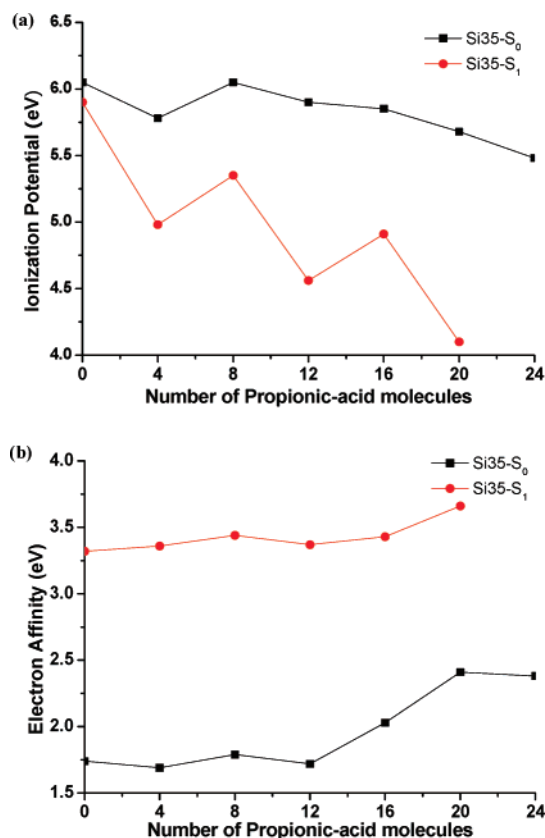
	ΔE_1 (eV)	absorption (nm)	f_1	ΔE_2 (eV)	emission (nm)	f_2
Si ₃₅ H ₃₆	4.316	283.8	0.08	2.588	358.5	0.06
Si ₃₅ H ₃₂ (C ₃ H ₅ O ₂) ₄	4.096	289.0	0.02	1.628	359.2	0.05
Si ₃₅ H ₂₈ (C ₃ H ₅ O ₂) ₈	4.262	289.5	0.01	1.909	360.2	0.03
Si ₃₅ H ₂₄ (C ₃ H ₅ O ₂) ₁₂	4.181	291.9	0.03	1.189	363.6	0.03
Si ₃₅ H ₂₀ (C ₃ H ₅ O ₂) ₁₆	3.822	306.4	0.03	1.481	360.6	0.08
Si ₃₅ H ₁₆ (C ₃ H ₅ O ₂) ₂₀	3.275	316.6	0.04	0.440	364.9	0.04
Si ₅₉ H ₆₀	3.648	334.4	0.25	2.894	390.4	0.13
Si ₅₉ H ₅₆ (C ₃ H ₅ O ₂) ₄	3.539	343.8	0.12	1.064	412.0	0.08
Si ₅₉ H ₅₂ (C ₃ H ₅ O ₂) ₈	3.515	346.5	0.14	1.051	413.6	0.07

^a ΔE_1 for ground-state configurations, ΔE_2 for excited-state configurations. ^b f_1 for absorption, f_2 for emission.

**Figure 3.** Calculated absorption and emission spectra (see text) for Si₃₅H₃₆ and Si₃₅H₂₀(C₃H₅O₂)₁₆. The units on the y-axis are arbitrary. Fifty roots were calculated, and the peaks were broadened using a Gaussian line shape with a line HWHM (half-width at half-maximum) of 20.

sion peak positions only slightly red-shift. This is consistent with the experimental finding that the PL maximum peak only slightly red-shifts after carboxyl functionalization.^{21,24} Moreover, we can see that the HOMO–LUMO energy gap in ground-state configuration is much larger than that in excited-state configuration, mainly because the LUMO energy level moves down significantly and the HOMO energy level moves up remarkably after structure relaxation in excited state. In addition, from the data in Table 2, we can observe that the maximum absorption gap of Si₅₉H₅₆–(C₃H₅O₂)₄ is about 0.7 eV lower than that of Si₃₅H₃₂–(C₃H₅O₂)₄, which is consistent with the corresponding gap difference between Si₅₉H₆₀ and Si₃₅H₃₆ due to the well-known quantum confinement effect. As a result, we can conclude that it is the size of SiQDs, especially the size of the silicon core, that determines the optical properties, while the amounts of adsorbed PA molecules have little effect on the optical spectra.

We further analyzed the simulated absorption and emission spectra of Si₃₅H₃₆ and Si₃₅H₂₀(C₃H₅O₂)₁₆ (see Figure 3). It could be seen that the PA adsorption leads to a red-shift of about 50 nm on the absorption spectrum, while the emission peak only slightly red-shifts. The absorption peak of Si₃₅H₂₀–(C₃H₅O₂)₁₆ appears at about 300 nm, which is in good agreement with the experimental values 320 nm²⁰ and 290

**Figure 4.** (a) –HOMO energy and (b) –LUMO energy for SiQDs in ground-state (S₀) or excited-state (S₁) configuration (see text) as a function of the adsorbed PA number.

nm.²¹ On the other hand, the emission peak of Si₃₅H₂₀–(C₃H₅O₂)₁₆ appears at around 435 nm, which is close to the experimental value 480 nm in ref 20 but a little far from the experimental value 600 nm in ref 21. This could be explained in terms of different sizes, that is, the diameter of Si₃₅H₂₀–(C₃H₅O₂)₁₆, 1.1 nm, is close to the size distribution of 1.4 ± 0.3 nm in ref 20 but a little smaller than that of 1.9–2.4 nm in ref 21. Moreover, we find that the optical absorption of Si₃₅H₂₀(C₃H₅O₂)₁₆ ranges from 230 nm to 360 nm, while the light emission ranges from 300 nm to 520 nm. Since these spectra show a substantial PL quantum yield in the visible region, it is possible to use PA-terminated SiQDs as candidates of biological chromophores.⁵⁵

3.3. Electronic Properties. In order to gain insight into the physical mechanisms responsible for the optical properties, it is necessary to examine the nature of the electronic states responsible for absorption and emission, that is, the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). In most occasions, the change in charge density introduced by the excitation of an electron from the HOMO to the LUMO induces forces on each atom due to the changes in the corresponding orbital densities.

In principle, electron affinity (A) corresponds to the energy given by the system when an additional electron is added, while the ionization potential (I) is referred to as the energy provided to the system to remove an electron. Melnikov and Chelikowsky⁵⁶ pointed out that for SiQDs LUMO and HOMO energies behave qualitatively as –A and –I, and

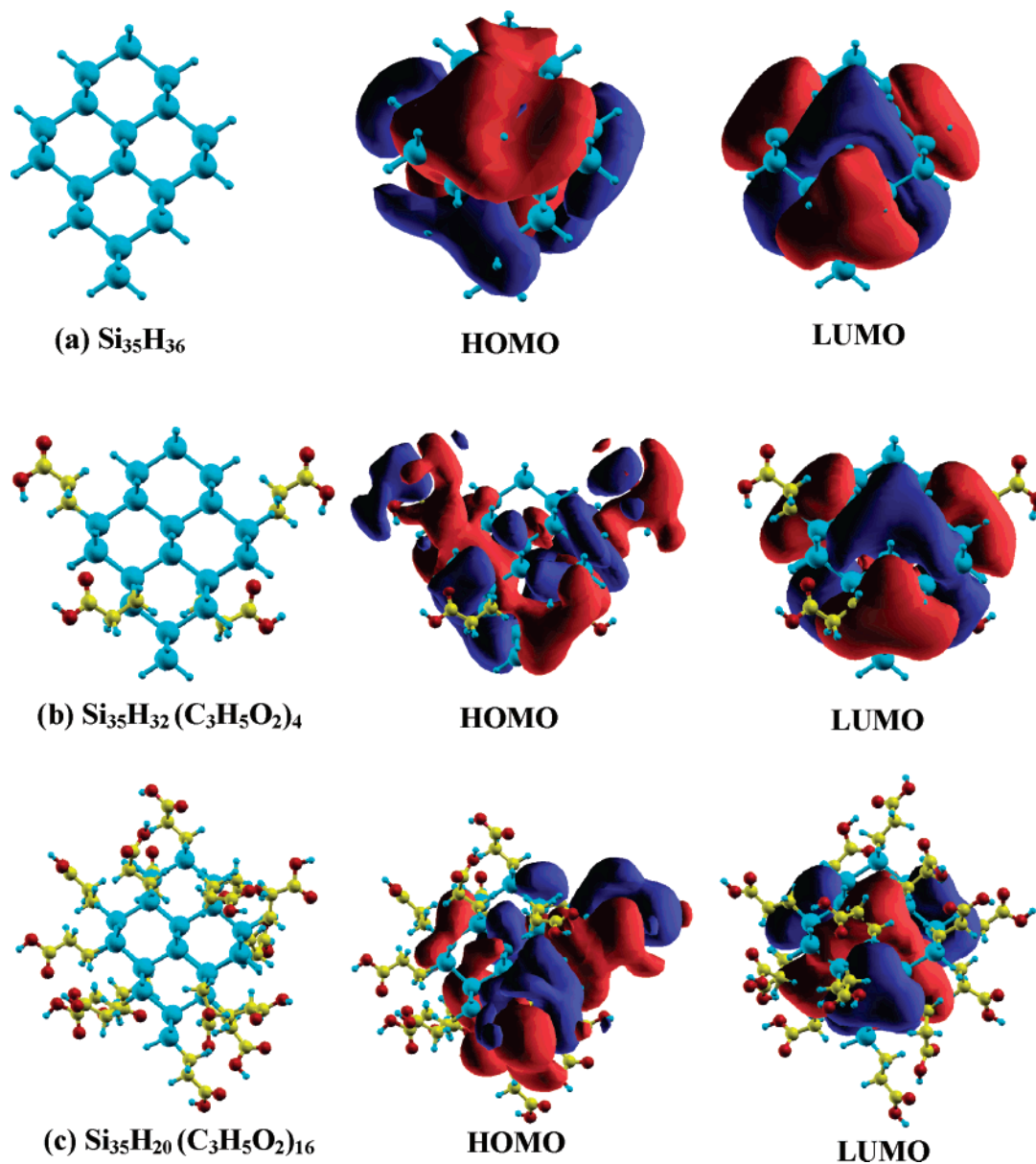


Figure 5. The isosurfaces of the HOMO and LUMO for $\text{Si}_{35}\text{H}_{36}$, $\text{Si}_{35}\text{H}_{32}(\text{C}_3\text{H}_5\text{O}_2)_4$, and $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$ in ground-state configurations.

this theorem has been successfully applied to explore the electronic properties of alkyl-terminated SiQDs.³⁴ In the subsection, we will discuss the HOMO and LUMO energies and relate them to electron affinity and ionization potentials of PA-terminated SiQDs. In Figure 4, we show the electron affinity and ionization potential of SiQDs as a function of the adsorbed PA number. In Figure 4(a), the ionization potential decreases remarkably with an increase in the adsorbed PA number in the excited-state configuration. In detail, the ionization potential is 5.90 eV for the S_1 state $\text{Si}_{35}\text{H}_{36}$, while it decreases to 5.35 and 4.10 eV for the S_1 states $\text{Si}_{35}\text{H}_{28}(\text{C}_3\text{H}_5\text{O}_2)_8$ and $\text{Si}_{35}\text{H}_{16}(\text{C}_3\text{H}_5\text{O}_2)_{20}$, respectively. This implies that PA-terminated SiQDs possess higher reactivity than that of hydrogenated SiQDs. On the other hand, the LUMO energy is slightly affected by PA adsorption except when the PA number exceeds 12 and the SiQDs are in ground-state configuration, as can be seen in Figure 4(b).

The isosurfaces of HOMO and LUMO orbitals in ground-state configurations of $\text{Si}_{35}\text{H}_{36}$, $\text{Si}_{35}\text{H}_{32}(\text{C}_3\text{H}_5\text{O}_2)_4$, and $\text{Si}_{35}\text{H}_{20}$ -

$(\text{C}_3\text{H}_5\text{O}_2)_{16}$ are shown in Figure 5. For $\text{Si}_{35}\text{H}_{36}$, the HOMO is triply degenerated due to the structure with T_d symmetry, while the LUMO is a holosymmetry delocalized orbital belonging to A_1 symmetry. Both HOMO and LUMO of $\text{Si}_{35}\text{H}_{36}$ are delocalized throughout the core of the silicon cluster. The isosurface of frontier orbitals reflects the structure symmetry and implies the isotropic reactivity of hydrogenated SiQDs. In contrast, for PA-terminated SiQDs, such as $\text{Si}_{35}\text{H}_{32}(\text{C}_3\text{H}_5\text{O}_2)_4$ and $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$, the HOMO isosurface is mostly drawn to the surface toward the PA branch, while most of the LUMO isosurface still exists in the core of the cluster, as can be clearly seen in Figure 5(b),-(c). A similar case has been reported by Puzder et al.⁴⁷ in studying the optical properties of silicon nanocrystals with oxygen passivation on the surface.

In order to further clarify the difference in the nature of the electronic states caused by PA adsorption, we schematically present the absorption and emission processes of $\text{Si}_{35}\text{H}_{36}$ and $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$ in Figure 6. In Figure 6(a), for $\text{Si}_{35}\text{H}_{36}$

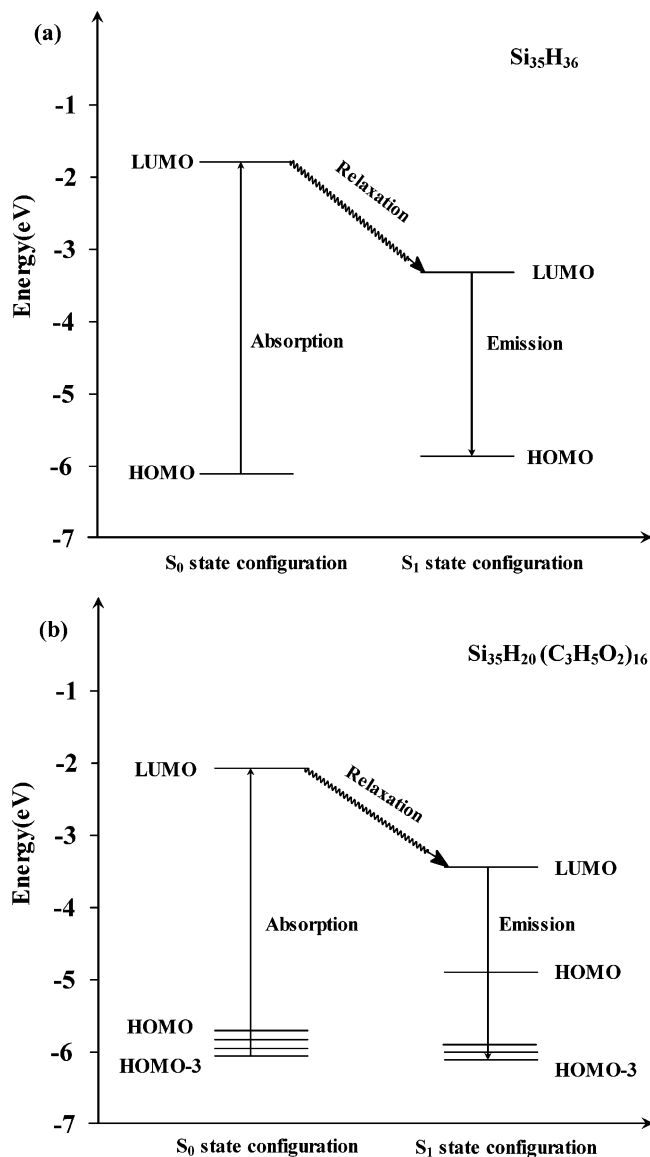


Figure 6. Schematic diagram showing photoabsorption and emission processes for $\text{Si}_{35}\text{H}_{36}$ and $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$.

the absorption spectrum is determined by the electron transfer from HOMO to LUMO. In the structure relaxation progress from ground-state configuration in T_d symmetry to excited-state configuration in C_1 symmetry, the energy of LUMO decreases as much as 1.578 eV, while the HOMO only increases slightly in energy and splits into several nearly degenerated orbitals, denoted by HOMO- n ($n = 0, 1, 2,$ or 3) hereafter. The PL emission occurs when the electron transfers from LUMO to HOMO- n ($n = 0, 1, 2,$ or 3) in the excited-state configuration.

The electronic properties of PA-terminated SiQDs are much more complex than those of hydrogenated SiQDs. Figure 6(b) shows the selected frontier orbital energy levels and the PL process of $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$. Our calculation results reveal that the adsorbed PA molecules result in some new orbitals with adjacent energies around Fermi energy and narrow the HOMO-LUMO energy gap, leading the electronic transition to be much more complex. For example, the absorption corresponding to the peak on the spectrum originates from the electron transfer from HOMO-3 to

LUMO, not HOMO to LUMO. In detail, the oscillator strength is 0.03 for HOMO-3 to LUMO electron transfer in $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$, while the value is only 0.002 for HOMO to LUMO electron transfer. Natural orbital analysis showed that the charge density of $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$ HOMO-3 delocalized in the silicon core and similar to that of $\text{Si}_{35}\text{H}_{36}$ HOMO.

In excited-state configuration, the distribution difference of the frontier orbitals between $\text{Si}_{35}\text{H}_{36}$ and $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$ is much more evident than that in the ground-state configuration, as can be seen in the right panel of Figure 6. The HOMO energy of $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$ in the excited-state configuration increases about 0.9 eV compared with that in the ground-state configuration, while the HOMO energy of $\text{Si}_{35}\text{H}_{36}$ only increases by about 0.15 eV after geometrical relaxation. Figure 7 shows the isosurfaces of the frontier orbitals of $\text{Si}_{35}\text{H}_{36}$ and $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$ in excited-state configurations. It could be clearly seen that most of the HOMO is localized in the PA branch, while the pattern of the $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$ HOMO-3 is generally similar to $\text{Si}_{35}\text{H}_{36}$ HOMO, though the signs (\pm) of some orbitals (e.g., denoted by red and blue in the Figure 7) are reversed. As a matter of fact, electron transfer from LUMO to HOMO-3 is proposed to account for the peak on the emission spectrum of $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$. This means the PA adsorption does not affect the general nature of the optical properties of SiQDs.

Experimentally, the PA adsorption not only increases the dispersibility but also improves the PL stability of the SiQDs against degeneration by water and oxygen.²¹ Although this could be simply explained in terms of steric hindrance, the physical mechanism remains unknown. At a molecular level, our calculations reveal that new frontier orbitals (HOMO, HOMO-1, HOMO-2) with energies adjacent to the Fermi energy appear as a result of PA adsorption. Natural orbital analysis showed that these new frontier orbitals are composed of the atomic orbitals on the PA branch. The modified surface of SiQDs can serve as a reaction substrate to oxygen and solvent molecules, which is responsible for the increase in both PL stability and water solubility. Similarly, it is suggested that alkyl passivation weakly affects optical gaps but leads to new bound states that affect excited-state properties of 1–2 nm silicon nanoclusters.³⁴

Allylamine ($\text{CH}_2=\text{CHCH}_2\text{NH}_2$)²⁰ and acrylic acid ($\text{CH}_2=\text{CHCOOH}$)^{21,23} have been successfully attached to the surface of SiQDs in different experimental conditions, both leading to water-soluble photoluminescent SiQDs. Allylamine and acrylic acid have the same carbon framework and different hydrophilic groups, $-\text{NH}_2$ or $-\text{COOH}$. Compared to N (3.04) and O (3.44) atoms, the electronegativity of C (2.55) is closer to that of the Si (1.90).⁵² According to our calculation results, in $\text{Si}_{35}\text{H}_{35}(\text{OH})$ and $\text{Si}_{35}\text{H}_{35}(\text{NH}_2)$, the adsorption groups carry negative charges of -0.130 and -0.175 , respectively. Both of them obtain more negative charge than that (-0.06) of the PA molecule. Thus, the charge transfer in Si-CR bond is weaker than that in the Si-NR bond or the Si-OR bond. Recent theoretical calculations⁵⁷ also showed that the single Si-C bridges are very stable, and replacing the Si-H bond by alkyl groups (Si-

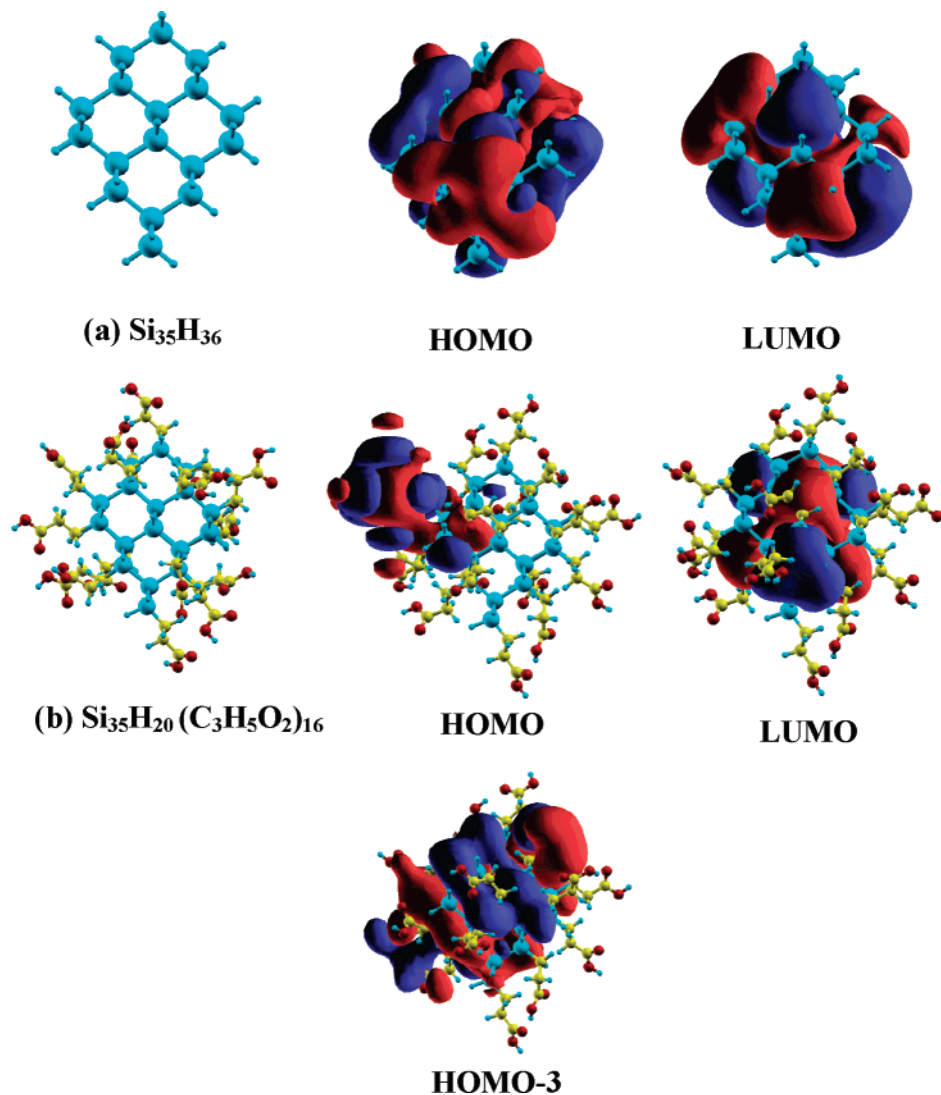


Figure 7. The isosurfaces of the selected frontier orbitals for $\text{Si}_{35}\text{H}_{36}$ and $\text{Si}_{35}\text{H}_{20}(\text{C}_3\text{H}_5\text{O}_2)_{16}$ in excited-state configurations.

C–R) results in a small reduction of the energy gap (0.5 eV) as compared to the large reduction observed with an oxide surface termination (2.3 eV).³⁴ There is no doubt that the formation of the Si–C surface bond will lead to small changes in the electronic and optical characters of the SiQDs. Hence, the PA molecules lead to small changes in the electronic and optical properties of the SiQDs and can be used as ideal branches to help the SiQDs maintain its PL stability.

4. Conclusions

The adsorbed PA molecules slightly affect the geometric structures of the silicon core, while the electron excitation leads to an obvious distortion to the structures. Furthermore, the PA adsorption has little effect on the absorption or emission spectra. It is the size of SiQDs, especially the size of the silicon core that determines the optical properties. The PA adsorption does not change the optical nature of the SiQDs. However, the adsorption substantially decreases the ionization potentials in the excited state and results in new active orbitals with adjacent energies around Fermi energy. Thus, the modified surface of SiQDs can serve as a reaction

substrate to oxygen and solvent molecules, which is responsible for the increase in both PL stability and water solubility.

Finally, our study verifies that surface modification is very important for the band structure engineering. Theoretical calculations can illuminate the physical mechanism and chemical nature of many useful materials at molecular or atomic levels.

Acknowledgment. The work described in this paper is supported by the Research Grants Council of Hong Kong SAR [Project Nos. 8730023, CityU 103106, and CityU 3/04C].

References

- (1) Buriak, J. M. *Chem. Rev.* **2002**, *102*, 1271.
- (2) Hua, F.; Stewart, M. T.; Ruckenstein, E. *Langmuir* **2005**, *21*, 6054.
- (3) Dancil, K. P. S.; Greiner, D. P.; Sailor, M. J. *J. Am. Chem. Soc.* **1999**, *121*, 7925.
- (4) Sohn, H.; Létant, S.; Sailor, J.; Trogler, W. C. *J. Am. Chem. Soc.* **2000**, *122*, 5399.

- (5) Meindl, J. D.; Chen, Q.; Davis, J. E. *Science* **2001**, 293, 2044.
- (6) Veinot, J. G. C. *Chem. Commun.* **2006**, 40, 4160.
- (7) Canham, L. T. *Appl. Phys. Lett.* **1990**, 57, 1046.
- (8) Ruckenstein, E.; Gourisankar, S. V. *J. Colloid. Interface Sci.* **1984**, 101, 436.
- (9) Holmes, J. D.; Ziegler, K. J.; Doty, R. C.; Pell, L. E.; Johnston, K. P.; Korgel, B. A. *J. Am. Chem. Soc.* **2001**, 123, 3743.
- (10) Park, N. M.; Kim, T. S.; Park, S. J. *Appl. Phys. Lett.* **2001**, 78, 2575.
- (11) Kim, B. H.; Cho, C. H.; Kim, T. W.; Park, N. M.; Sung, G. Y. *Appl. Phys. Lett.* **2005**, 86, 091908.
- (12) Nagesha, D. K.; Whitehead, M. A.; Coffey, J. L. *Adv. Mater.* **2005**, 17, 921.
- (13) Canham, L. T.; Reeves, C. L.; Newey, L. P.; Houlton, M. R.; Cox, T. I.; Buriak, J. M.; Stewart, M. P. *Adv. Mater.* **1999**, 11, 1505.
- (14) Delerue, C.; Allan, G.; Lannoo, M. *Phys. Rev. B* **1993**, 48, 11024.
- (15) Seotsanyana-Mokhosi, I.; Kuznetsova, N.; Nyokong, T. *J. Photochem. Photobiol., A* **2001**, 140, 215.
- (16) Wolkin, M. V.; Jorne, J.; Fauchet, P. M. *Phys. Rev. Lett.* **1999**, 82, 197.
- (17) Zhou, Z. Y.; Brus, L.; Friesner, R. *Nano Lett.* **2003**, 3, 163.
- (18) Bruchez, M., Jr.; Moronne, M.; Gin, P.; Weiss, S.; Alivisatos, A. P. *Science* **1998**, 281, 2013.
- (19) Yan, C. S.; Bley, R. A.; Kauzlarich, S. M.; Lee, H. W. H.; Delgado, G. R. *J. Am. Chem. Soc.* **1999**, 121, 5191.
- (20) Warner, J. H.; Hoshino, A.; Yamamoto, K.; Tilley, R. D. *Angew. Chem., Int. Ed.* **2005**, 44, 4550.
- (21) Li, Z. F.; Ruckenstein, E. *Nano Lett.* **2004**, 4, 1463.
- (22) Ruckenstein, E.; Li, Z. F. *Adv. Colloid Interface Sci.* **2005**, 113, 43.
- (23) Sato, S.; Swihart, M. T. *Chem. Mater.* **2006**, 18, 4083.
- (24) Rogozhina, E. V.; Eckhoff, D. A.; Gratton, E.; Braun, P. V. *J. Chem. Mater.* **2006**, 16, 1421.
- (25) Stewart, M. P.; Buriak, J. M. *J. Am. Chem. Soc.* **2001**, 123, 7821.
- (26) Lie, L. H.; Patole, S. N.; Pike, A. R.; Ryder, L. C.; Connolly, B. A.; Ward, A. D.; Tuite, E. M.; Houlton, A.; Horrocks, B. R. *Faraday Discuss.* **2004**, 125, 235.
- (27) Chao, Y.; Krishnamurthy, S.; Montalti, M.; Lie, L. H.; Houlton, A.; Horrocks, B. R.; Kjeldgaard, L.; Dhanak, V. R.; Hunt, M. R. C.; Šiller, L. *J. Appl. Phys.* **2005**, 98, 044316.
- (28) Pettigrew, K. A.; Liu, Q.; Power, P. P.; Kauzlarich, S. M. *Chem. Mater.* **2003**, 15, 4005.
- (29) Li, Z. F.; Kang, E. T.; Neoh, K. G.; Tan, K. L. *Biomaterials* **1998**, 19, 45.
- (30) Franchina, J. G.; Lackowski, W. M.; Dermody, D. L.; Crooks, R. M.; Bergbreiter, D. E. *Anal. Chem.* **1999**, 71, 3133.
- (31) Zhang, R. Q.; Costa, J.; Bertran, E. *Phys. Rev. B* **1996**, 53, 7847.
- (32) Zhou, Z.; Friesner, R. A.; Brus, L. *J. Am. Chem. Soc.* **2003**, 125, 15599.
- (33) Reboredo, F. A.; Schwegler, E.; Galli, G. *J. Am. Chem. Soc.* **2003**, 125, 15243.
- (34) Reboredo, F. A.; Galli, G. *J. Phys. Chem. B* **2005**, 109, 1072.
- (35) Wang, X.; Zhang, R. Q.; Niehaus, T. A.; Frauenheim, Th. *J. Phys. Chem. C* **2007**, 111, 2394.
- (36) Porezag, D.; Frauenheim, Th.; Köhler, Th.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, 51, 947.
- (37) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, 58, 7260.
- (38) Niehaus, T. A.; Suhai, S.; Sala, F. D.; Lugli, P.; Elstner, M.; Seifert, G.; Frauenheim, Th. *Phys. Rev. B* **2001**, 63, 085108.
- (39) Casida, M. E.; Casida, K. C.; Salahub, D. R. *Int. J. Quantum Chem.* **1998**, 70, 933.
- (40) Casida, M. E.; Salahub, D. R. *J. Chem. Phys.* **2000**, 113, 8918.
- (41) Fehér, F. *Research Report of the Federal State North Rhine-Westphalia*; West-deutscher Verlag: Köln, 1977.
- (42) Zdetsis, A. D. *Rev. Adv. Mater. Sci.* **2006**, 11, 56.
- (43) Luppi, E.; Degoli, E.; Cantele, G.; Ossicini, S.; Magri, R.; Ninno, D.; Bisi, O.; Pulci, O.; Onida, G.; Gatti, M.; Ince, A.; Sole, E. D. *Opt. Mater.* **2005**, 27, 1008.
- (44) Williamson, A. J.; Grossman, J. C.; Hood, R. Q.; Puzder, A.; Galli, G. *Phys. Rev. Lett.* **2002**, 89, 196803.
- (45) Vasiliev, I.; Ogut, S.; Chelikowsky, J. R. *Phys. Rev. Lett.* **2001**, 86, 1813.
- (46) Garoufalis, C. S.; Zdetsis, D.; Grimme, S. *Phys. Rev. Lett.* **2001**, 87, 276402.
- (47) Puzder, A.; Williamson, A. J.; Grossman, J. C.; Galli, G. *J. Am. Chem. Soc.* **2003**, 125, 2786.
- (48) Garoufalis, C. S.; Zdetsis, A. D. *Phys. Chem. Chem. Phys.* **2006**, 8, 808.
- (49) Vasiliev, I.; Chelikowsky, J. R.; Martin, R. M. *Phys. Rev. B* **2002**, 65, 121302.
- (50) Luppi, M.; Ossicini, S. *Phys. Rev. B* **2005**, 71, 035340.
- (51) *CRC Handbook of Chemistry and Physics*, 78th ed.; Lide, D. R., Ed.; CRC Press: New York, 1997; pp 9–22.
- (52) Stuhlmann, C.; Bogdányi, G.; Ibach, H. *Phys. Rev. B* **1992**, 45, 6786.
- (53) *The Nature of the Chemical Bond and the Structure of Molecules and Crystals*; Pauling, L., Ed.; Cornell University Press: 1960.
- (54) Schafer, R.; Becker, J. A. *Phys. Rev. B* **1996**, 54, 10296.
- (55) Wilson, W. L.; Szajowski, P. F.; Brus, L. E. *Science* **1993**, 262, 1242.
- (56) Melnikov, D. V.; Chelikowsky, J. R. *Phys. Rev. B* **2004**, 69, 113305.
- (57) Cucinotta, C. S.; Bonferroni, B.; Ferretti, A.; Ruini, A.; Caldas, M. J.; Molinari, E. *Surf. Sci.* **2006**, 600, 3892.

Accurate ab Initio Study on the Hydrogen-Bond Pairs in Protein Secondary Structures

Zhi-Xiang Wang, Chun Wu, Hongxing Lei, and Yong Duan*

Genome Center and Department of Applied Science, University of California, Davis, California 95616

Received January 17, 2007

Abstract: Ab initio calculations up to the MP2/aug-cc-pVQZ//MP2/6-311+G** level have been carried out to characterize the four patterns of hydrogen-bond (H-bond) pairs in protein secondary structures. The unblocked and methyl-blocked glycine dipeptide dimers were arranged to model the H-bond pairs in α -helix (α HH) and antiparallel ($A\beta\beta$ -C₅ and $A\beta\beta$ -C₇) and parallel β -sheet ($P\beta\beta$) secondary structures. The study uncovers that, in addition to the primary CO \cdots NH H-bonds and the crossing secondary interactions, the CH \cdots OC H-bonds and the tertiary effect (as we call it) also contribute substantially. The tertiary effect is due to the interpolarization between the donor and acceptor of a H-bond. This effect, which enhances the dipole–dipole interactions between two nearby H-bonds, stabilizes the β -sheet-like but destabilizes the helix-like H-bond pairs. The MP2 binding energies of the complexes were further refined by extrapolating to the complete basis set limit (CBS) according to Truhlar and co-workers and by a three-basis-set-based method. The best extrapolated CBS(aD-aT-aQ) binding energies of the unblocked dimers are -13.1 (α HH), -11.3 ($A\beta\beta$ -C₅), -19.2 ($A\beta\beta$ -C₇), and -14.8 kcal/mol ($P\beta\beta$). For the methyl-blocked counterparts, the best extrapolated CBS(D-T-Q) binding energies are -14.8 , -13.4 , -20.8 , and -16.7 kcal/mol, respectively. The interactions in the parallel β conformations are very close to the averages of the C₅ and C₇ antiparallel β conformations, and both are stronger than the helical dimers. Because the additive force fields are unable to account for the tertiary effect owing to the lack of polarization, all examined additive force fields significantly overestimate the interaction energies of the helix conformations relative to the β -sheet conformations. Notably, the agreement between molecular mechanical and quantum mechanical binding energies is improved after turning on the polarization. The study provides reference ab initio structures and binding energies for characterizing the backbone H-bonds of the protein secondary structures, which can be used for the parametrization of empirical molecular mechanics force fields.

1. Introduction

Hydrogen bonds (H-bonds), together with other weak interactions, are some of the most important determinants of the three-dimensional structures of proteins.^{1,2} The energy of a single H-bond, ranging from 5.0 to 10.0 kcal/mol, is comparable to the typical folding free energies of proteins. Thus, accurate characterization of these H-bonds is vital for understanding the factors stabilizing protein structures.

Accurate H-bond energies are also crucial reference data for the development of protein molecular mechanics (MM) force fields that have become powerful tools in structural biology.³ Numerous studies^{4–16} have been performed to gain insight into the underlying physical interactions of H-bonds. Among the H-bonds in proteins, the backbone C=O \cdots H–N H-bonds play particularly important roles and are the major driving forces for forming the ordered secondary structures.

Modeling the backbone H-bonds has been one of the major concerns in parametrizing molecular mechanics force fields

* Corresponding author tel.: (530) 754-7632; fax: (530) 754-9658; e-mail: duan@ucdavis.edu.

for protein simulations, and various potential functions has been developed. For the physical-based force fields, because of the lack of experimental data for the backbone H-bonds, *ab initio* values were often used as reference data. As a prototype, the *trans* N-methylacetamide (NMA) dimer^{17–24} has long been used to model such H-bonds and has been compared with the NMA–water complexes to study the relative strength of inter and intra H-bonds. With the advancement in computer hardware and software, the H-bond energies of these model complexes have been updated continuously, from Jorgensen and Swenson's^{23,24} Hartree–Fock (HF)/minimal basis set calculations in 1985 to the most recent work of Langley and Allinger¹⁹ at the MP2/6-311++G(2d,2p) level. Using the continuum solvent model, we²⁵ recently studied the solvent effect on the H-bonds. It is interesting to mention that Kelly and co-workers²⁶ have recently developed the amide-to-ester mutation approach to estimate the contribution of backbone H-bonds. But they are the free energy contributions and cannot be used to parametrize physical-based force fields directly.

A limitation of the *trans* NMA–NMA model is that it only contains one H-bond and is unable to capture the neighboring effect exerted by the nearby H-bonds on protein backbones. Recently, the aesthetic H-bond network in protein secondary structures has attracted attention from both experimentalists^{27,28} and theorists.^{29–34} Highlights of these efforts include the works of Wu^{29,30} and Dannenberg^{31–34} and their respective co-workers. In these cases, the influences on the H-bonds were assessed in the context of H-bond networks, but high-level *ab initio* calculations were difficult to perform due to the large size of the model complexes. Notwithstanding the efforts, ambiguity and controversy exist as to the contributions of the underlying physical interactions, and a detailed and reliable characterization of H-bond pairs in the context of protein secondary structures is unavailable. It is worth mentioning that Hobza and co-workers^{35,36} have delivered highly accurate H-bond energies for nucleic acid base pairs using state-of-the-art computational chemistry methods.

An empirical approach to consider the neighboring effect in H-bond networks has been proposed by Jorgensen and Pranata,³⁷ who found that the effect in the multiple H-bonds of the nucleic acid base pairs could be accounted reasonably by the secondary interactions. This approach has been applied beyond the base pairs; because their model is consistent with additive point charge molecular mechanics force fields, widespread application of the latter implicitly renders their approach as the *de facto* model to account for the neighboring effect. However, as is well-known, the additive force fields are unable to represent the polarization effect. This approach, even for the base pairs, has been questioned by Lukin and Leszczynski³⁸ and by Dannenberg and co-workers³⁹ on the basis of the *ab initio* calculations. The fidelity of their approach in describing the H-bond pairs in protein secondary structures has not been examined despite the countless (implicit) applications.

In this study, we are interested in the typical H-bond pairs existing in protein secondary structures (Figure 1). We attempt (i) to reliably characterize the interactions between the two strands, (ii) to assess the neighboring effect between

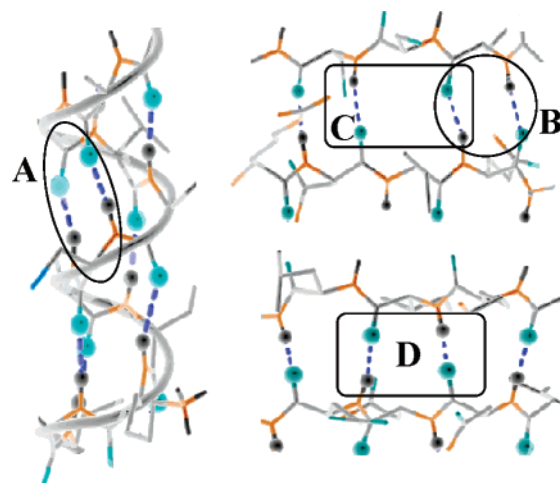


Figure 1. Four patterns of H-bond pairs in protein secondary structures. (A) H-bond pair in α helices; (B) C_5 H-bond pair in antiparallel β sheets; (C) C_7 H-bond pair in antiparallel β sheets; (D) H-bond pair in parallel β sheets. The notations C_5 and C_7 refer to the H-bond pairs in which the H of NH and the O of CO in the same strand are four (C_5) and six (C_7) bonds away, respectively.

two nearby H-bonds in these pairs, (iii) to obtain accurate interaction energies, which can be used to guide the parametrization of force fields, and (iv) to examine the fidelity of the modern force fields with regard to the ability to model the main-chain H-bonds of peptides.

2. Methods

Two sets of glycine dipeptide dimers (referred to as unblocked and methyl-blocked, respectively) were used to model the patterns of H-bond pairs shown in Figure 1. The unblocked set has the advantage to avoid the steric disturbance of the blocking methyl groups, whereas the H-bond donors and acceptors in the methyl-blocked set have the chemical-bonding environment closer to protein peptides. The four dimers in the unblocked set hereafter are referred to as α HH for the H-bond pairs in the α helix, $A\beta\beta-C_5$ and $A\beta\beta-C_7$ for the H-bond pairs in the antiparallel β sheet, and $P\beta\beta$ for the H-bond pairs in the parallel β sheet. Their counterparts in the methyl-blocked set are labeled as α HH', $A\beta\beta-C_5'$, $A\beta\beta-C_7'$, and $P\beta\beta'$, respectively. Here, " C_5 " and " C_7 " denote the H-bond pairs where the hydrogen of the NH donor and the oxygen of the CO acceptor in the same strand are four (C_5) and six bonds (C_7) away, respectively. These dimers and their corresponding monomers are not at the energy minima. To maintain the H-bond pairs to be similar to those in protein secondary structures, we therefore fixed the backbone (Φ , Ψ) torsions at the typical angles in protein secondary structures, that is, (-57.0° , -47.0°) in α HH and α HH'; (-119° , 113°) in $P\beta\beta$ and $P\beta\beta'$; and (-139° , 135°) in $A\beta\beta-C_5$, $A\beta\beta-C_7$, $A\beta\beta-C_5'$, and $A\beta\beta-C_7'$, respectively. It should be noted that, as these energies are applied to calibrate or parametrize empirical force fields, the backbone torsion angles in the force field calculations should be fixed to the same values.

Accurate estimation of nonbonded interactions, including H-bond interactions, has long been a challenge in compu-

Table 1. Energetic Results (in kcal/mol) of Unblocked and Methyl-Blocked H-bond Pairs at Various Levels, Including the BSSE-Uncorrected (ΔE_{uc}) and BSSE-corrected (ΔE_c) bonding energies, together with the BSSE corrections (ΔE_{bsse})^a

unblocked dimers	αHH			$A\beta\beta\text{-C}_5$			$A\beta\beta\text{-C}_7$			$P\beta\beta$			$\Delta\Delta E_{uc} = \Delta E_{uc} - \Delta E_{uc}(\alpha\text{HH})$		
	ΔE_{uc}	ΔE_{bsse}	ΔE_c	ΔE_{uc}	ΔE_{bsse}	ΔE_c	ΔE_{uc}	ΔE_{bsse}	ΔE_c	ΔE_{uc}	ΔE_{bsse}	ΔE_c	$A\beta\beta\text{-C}_5^b$	$A\beta\beta\text{-C}_7^c$	$P\beta\beta^d$
MP2/6-311+G**	-12.4	2.9	-9.5	-13.5	2.7	-9.8	-19.3	3.1	-16.2	-15.5	2.8	-12.7	1.1	-6.9	-3.1
MP2/cc-pVDZ	-14.7	6.7	-8.0	-13.7	5.1	-8.6	-22.2	7.1	-15.0	-17.6	6.1	-11.5	1.0	-7.5	-2.9
MP2/cc-pVTZ	-14.3	3.1	-11.3	-12.5	2.3	-10.3	-20.6	2.8	-17.7	-16.2	2.5	-13.6	1.8	-6.3	-1.9
MP2/cc-pVQZ	-13.9	1.4	-12.5	-12.2	1.0	-11.2	-20.2	1.2	-19.0	-15.8	1.1	-14.6	1.7	-6.3	-1.9
MP2/aug-cc-pVDZ	-15.5	3.6	-11.9	-13.8	3.0	-10.8	-21.8	3.4	-18.4	-17.2	3.1	-14.2	1.7	-6.3	-1.7
MP2/aug-cc-pVTZ	-14.6	1.9	-12.8	-13.0	1.7	-11.3	-21.1	1.9	-19.2	-16.6	1.8	-14.8	1.6	-6.5	-2.0
MP2/aug-cc-pVQZ	-14.0	1.0 ^e	-13.0	-12.3	1.0 ^e	-11.3	-20.4	1.4 ^e	-19.0	-15.9	1.1 ^e	-14.8	1.7	-6.4	-1.9
CBS(D-T) ^f	-15.0			-12.6			-20.7			-16.2			2.4	-5.7	-1.2
CBS(aD-aT) ^f	-14.2			-12.7			-20.9			-16.4			1.5	-6.7	-2.2
CBS(T-Q) ^f	-13.9			-12.1			-20.2			-15.7			1.8	-6.3	-1.8
CBS(D-T-Q) ^g	-13.3			-12.1			-20.1			-15.6			1.2	-6.8	-2.3
CBS(aT-aQ) ^f	-13.3			-11.4			-19.6			-15.0			1.9	-6.3	-1.7
CBS(aD-aT-aQ) ^g	-13.1			-11.3			-19.2			-14.8			1.8	-6.1	-1.7

blocked dimers	$\alpha\text{HH}'$			$A\beta\beta\text{-C}_5'$			$A\beta\beta\text{-C}_7'$			$P\beta\beta$			$\Delta\Delta E_{uc} = \Delta E_{uc} - \Delta E_{uc}(\alpha\text{HH}')$		
	ΔE_{uc}	ΔE_{bsse}	ΔE_c	ΔE_{uc}	ΔE_{bsse}	ΔE_c	ΔE_{uc}	ΔE_{bsse}	ΔE_c	ΔE_{uc}	ΔE_{bsse}	ΔE_c	$A\beta\beta\text{-C}_5'^b$	$A\beta\beta\text{-C}_7'^c$	$P\beta\beta^d$
MP2/6-311+G**	-16.3	4.7	-11.5	-13.8	3.1	-10.7	-20.8	3.9	-16.8	-17.2	3.7	-13.5	2.5	-4.5	-0.9
MP2/cc-pVDZ	-17.9	8.0	-9.9	-16.4	7.0	-9.4	-23.1	7.6	-15.5	-19.5	7.3	-12.2	1.5	-5.2	-1.6
MP2/cc-pVTZ	-16.3	3.4	-12.9	-14.2	2.9	-11.3	-21.5	3.2	-18.3	-17.7	3.1	-14.6	2.1	-5.2	-1.4
MP2/cc-pVQZ	-15.5	1.4	-14.1	-13.6	1.2	-12.4	-21.0	1.4	-19.6	-17.1	1.3	-15.7	1.9	-5.5	-1.6
MP2/aug-cc-pVDZ	-18.8	5.2	-13.5	-16.1	4.1	-12.0	-23.4	4.5	-18.9	-19.5	4.2	-15.2	2.7	-4.6	-0.7
MP2/aug-cc-pVTZ	-16.9			-14.8			-22.3			-18.3			2.1	-5.4	-1.4
CBS(D-T) ^f	-16.5			-13.9			-21.7			-17.7			2.6	-5.2	-1.2
CBS(aD-aT) ^f	-15.8			-14.2			-21.7			-17.7			1.6	-5.9	-1.9
CBS(T-Q) ^f	-15.0			-13.2			-20.8			-16.8			1.8	-5.8	-1.8
CBS(D-T-Q) ^g	-14.8			-13.4			-20.7			-16.7			1.4	-5.9	-1.9

^a All geometries were optimized at MP2/6-311+G**. ^b $\Delta E_{uc}(A\beta\beta\text{-C}_5) - \Delta E_{uc}(\alpha\text{HH})$. ^c $\Delta E_{uc}(A\beta\beta\text{-C}_7) - \Delta E_{uc}(\alpha\text{HH})$. ^d $\Delta E_{uc}(P\beta\beta) - \Delta E_{uc}(\alpha\text{HH})$. ^e Extrapolated using $\text{BSSE}(n) = \text{BSSE}(0) \exp(-\alpha n)$. ^f Extrapolated according to Truhlar and co-workers.^{48,49} ^g Extrapolated using exponential formula based on ΔE_{uc} and ΔE_c calculated at cc-pVXZ or aug-cc-pVXZ (X = D, T, and Q).⁵⁰

tational chemistry. On the basis of a systematic study on a set of nonbonded complexes, Rappe and Bernstein⁴⁰ concluded that low levels of correlation theory such as the second-order Møller–Plesset perturbation theory (MP2) can account for the full range of intermolecular interactions, and the accuracy mainly lies in the convergence with respect to the basis set expansion. In comparison, the DFT method is less reliable because of the lack of an appropriate description of the dispersion effect. According to Rappe and Bernstein⁴⁰ and in consideration of the size of the model complexes (18 heavy atoms and 10 hydrogen atoms for the methyl-blocked dimers) and the available computer resources, we used MP2 theory to account for the correlation energy and focused on the convergence. Because the CCSD(T) calculations even with the 6-31G* basis set are extremely time-consuming, we decided not to account for the correlation at the CCSD(T) level. However, we note that, in the calculations of interaction energies of base pairs of nucleic acids, Hobza and co-workers³⁵ have considered the higher-order correlation at the CCSD(T)/6-31G* level for some of their studied base pairs. They³⁵ found that the CCSD(T) corrections, ranging from 0.0 to -0.6 kcal/mol, only have a marginal effect on the relative stability of base pairs.

The geometries of the complexes were optimized at MP2/6-311+G** without including basis set superposition error^{41,42} (BSSE) correction, and the backbone (Φ , Ψ) torsions

were fixed at the above angles. Interestingly, as indicated in Table 1, the MP2/6-311+G** BSSE-uncorrected H-bond energies are in better agreement with the more reliable estimations than the BSSE-corrected ones. This suggests that the BSSE-uncorrected optimization at the current level might actually give better geometries than the one with the BSSE correction. We note that care should be exercised and further studies with notably higher-level optimization might help to examine this issue.

The energies were then refined by single-point calculations at the MP2/cc-pVXZ and MP2/aug-cc-pVXZ (X = D, T, and Q) levels at the MP2/6-311+G** geometries. The calculations for the unblocked dimers involved up to 1672 basis functions at MP2/aug-cc-pVQZ, and those for methyl-blocked dimers involved up to 1590 basis functions at the MP2/cc-pVQZ level. The parallel Gaussian 03 package⁴³ was used to perform all ab initio calculations, which took about 2 months on the latest SGI Altix computer with 32 Itanium-2 CPUs, 128 GB of memory, and a 2.0 TB hard disk.

The interaction energies, including BSSE-uncorrected (ΔE_{uc}) and BSSE-corrected (ΔE_c), were calculated using $\Delta E_{uc} = E_{\text{dim}} - 2E_{\text{mon}}$ and $\Delta E_c = E_{\text{dim}} - 2E_{\text{mon}} + \Delta E_{\text{BSSE}}$, respectively, where E_{dim} and E_{mon} are the total energies of dimers and the isolated monomers. The BSSE correction energies (ΔE_{BSSE}) were computed using the standard counterpoise (CP) method^{41,42} at the dimer geometries. Because

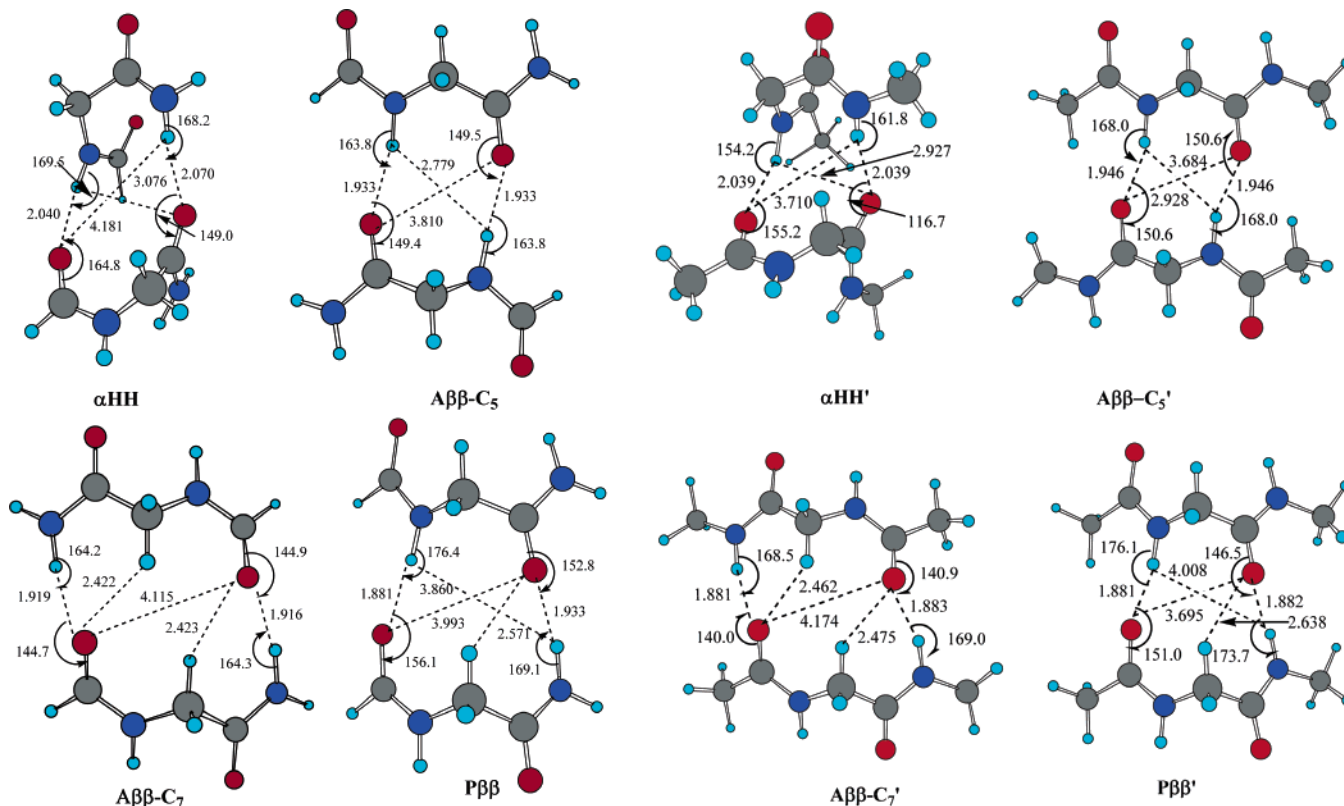


Figure 2. MP2/6-311+G** geometries of unblocked and methyl-blocked dimers (angles are in degrees and bond lengths are in angstroms). The C₅ and C₇ conformations refer to the two patterns of H-bond pairs in antiparallel β sheets (see Figure 1 caption for explanation). The atoms are colored as red (O), blue (N), gray (C), and cyan (H).

the interactions are dominated by H-bonds in these complexes, we will refer to the interaction energies as the H-bond energies.

In the present study, we focus on the interaction energies of the H-bonds in peptides with patterns similar to those in the protein secondary structures, rather than their relative stabilities of the dimers and monomers. In addition, because all of the dimers and monomers are not minima, the zero-point energies are not meaningful and are excluded in our defined binding energies. The structures and energies of alanine and glycine dipeptide monomers have been extensively studied previously.^{44–47}

The HF and MP2 single-point energies were further extrapolated to the complete basis set (CBS) energies following Truhlar's scheme.⁴⁸ The HF and correlation CBS energies were estimated using $E_{\text{CBS}}^{\text{HF}} = E^{\text{HF}}(n) - A^{\text{HF}}n^{-\alpha}$ and $E_{\text{CBS}}^{\text{corr}} = E^{\text{corr}}(n) - A^{\text{corr}}n^{-\beta}$, respectively, where n represents the highest angular momentum in the basis set, that is, $n = 2, 3,$ and 4 corresponding to D, T, and Q basis sets, respectively, and $\alpha = 3.39$ and $\beta = 1.91$, which were optimized by Truhlar and co-workers.⁴⁹ These energies are termed as CBS(X-Y), where X and Y can be D, T, and Q, representing the basis sets used in MP2 calculations (or aD, aT, and aQ for the augmented basis sets).

Following Dixon and co-workers,⁵⁰ the interaction energies were alternatively extrapolated to the complete basis set limit by the exponential relation, $\Delta E_{\text{CBS}} = \Delta E(n) - A \exp(-Bn)$, where n is the same as in Truhlar and co-workers' formula and ΔE_{CBS} is the extrapolated interaction energies. The constants, A and B , were determined by the least-square fit

to six energy points including both BSSE-corrected and BSSE-uncorrected binding energies. Because these extrapolations involve energies calculated at three levels of basis sets, they are referred to as the three-basis-set exponential extrapolations and are noted as either CBS(D-T-Q) or CBS(aD-aT-aQ) for the unaugmented and augmented basis sets, respectively.

The MM energies were calculated using AMBER,^{51–54} CHARMM,^{55,56} and OPLS force fields.^{57,58} The AMBER MM calculations were carried out by the AMBER 8.0 package,⁵⁹ and calculations with CHARMM and OPLS force fields were done with the TINKER program.⁶⁰ In all MM energy calculations, the backbone (Φ, Ψ) torsion angles were restrained to the corresponding values in the quantum mechanical (QM) calculations, and other geometrical parameters were optimized.

3. Results and Discussion

The MP2/6-311+G** optimized structures of the complexes, together with the key geometrical parameters, are displayed in Figure 2. Table 1 compiles the energetic results, including the BSSE-corrected (ΔE_c) and BSSE-uncorrected binding energies (ΔE_{uc}) and the BSSE corrections (ΔE_{BSSE}). The extrapolated binding energies using Truhlar's⁴⁸ method are termed as CBS(X-Y), and those using the three-basis-set exponential extrapolations⁵⁰ are noted as CBS(X-Y-Z) (see notes in "Methods" for explanation). Also listed in Table 1 are the binding energy differences at various levels relative to the α -helical dimers.

As shown in Table 1, the extrapolated CBS H-bond energies for the unblocked dimers generally decrease with the increasing basis set functions, with CBS(D-T) being the largest and CBS(aD-aT-aQ) the smallest. The two best extrapolated energy sets, CBS(aT-aQ) and CBS(aD-aT-aQ), are very close to each other, and the average differences are less than 0.3 kcal/mol. This holds true for the CBS(T-Q) and CBS(D-T-Q) energies of the methyl-blocked dimers. Although, with our available computer power, we were unable to perform MP2/aug-cc-pVQZ calculations for the methyl-blocked dimers, the trends shown in Table 1 indicate that the binding energies reached satisfactory convergence at the current level and the uncertainties are below the room-temperature thermal energies. On the basis of convergence and consistency, we consider the three-basis-set extrapolated interaction energies to be the most reliable values.

The binding energy differences relative to the helical dimers (last three columns in Table 1) are reasonably consistent at various basis sets, although the individual binding energies differ notably from the more reliable calculations. A comparison of the binding energies at the various levels with the best estimates suggests that a reasonable assessment of the relative strengths of the H-bonds could be achieved at the MP2/cc-pVTZ level. However, reliable absolute binding energies require large basis sets and extrapolation, as indicated by the trend shown in Table 1.

At the CBS(aD-aT-aQ) level, relative to αHH , the binding in $A\beta\beta\text{-C}_5$ is 1.8 kcal/mol weaker, in $A\beta\beta\text{-C}_7$ is 6.1 kcal/mol stronger, and in $P\beta\beta$ is 1.7 kcal/mol stronger. The average binding energy (15.3 kcal/mol) of the two antiparallel β -sheet conformations, $A\beta\beta\text{-C}_5$ and $A\beta\beta\text{-C}_7$, which coexist in the antiparallel β sheet, is 2.2 kcal/mol larger than that of αHH . Consistently, for the blocked dimers, relative to $\alpha\text{HH}'$, the binding in $A\beta\beta\text{-C}_5'$ is 1.4 kcal/mol weaker at the CBS(D-T-Q) level, in $A\beta\beta\text{-C}_7'$ is 5.9 kcal/mol stronger, and in $P\beta\beta$ is 1.9 kcal/mol stronger. The average (17.1 kcal/mol) of $A\beta\beta\text{-C}_5'$ and $A\beta\beta\text{-C}_7'$ is 2.3 kcal/mol larger. The energetic results of both sets of model complexes indicate that the interactions in the helices are weaker than those in parallel and antiparallel β sheets and the interactions in the antiparallel β sheet is comparable to that in the parallel β sheets. In comparison to the unlocked dimers, the presence of the blocking groups in the methyl-blocked dipeptide dimers strengthens the H-bonds by 1.5–2.1 kcal/mol.

Other consistent trends include the systematic decrease of BSSE-uncorrected MP2 binding energies (ΔE_{uc}) as the basis sets expanded from DZ to TZ to QZ and, conversely, the systematic increase of the BSSE-corrected binding energies (ΔE_{c}). The opposite convergence trends of ΔE_{c} and ΔE_{uc} are indicative of the overestimation of BSSE by the CP method,^{41,42} particularly with the small basis sets. For the unblocked dimers, the MP2/aug-cc-pVQZ binding energies with the extrapolated BSSEs are very close to the final convergent CBS(aD-aT-aQ) values, but the uncorrected ones are overestimated.

3.1. Interaction Energies in H-Bond Pairs. The binding energies of αHH at CBS(aD-aT-aQ) and CBS(D-T-Q) are -13.1 and -13.3 kcal/mol, respectively, which are slightly less than the CBS(aT-aQ) and CBS(T-Q) values, -13.3 and

-13.9 kcal/mol, respectively. However, the CBS(D-T) and CBS(aD-aT) values, -15.0 and -14.2 kcal/mol, respectively, are less reliable, which may be due to the less superior DZ basis set. Relative to the most sophisticated CBS(aD-aT-aQ) energy, all BSSE-uncorrected MP2 binding energies are overestimated, but the energies with large basis sets, -14.0 (aug-cc-pVQZ) and -13.9 kcal/mol (cc-pVQZ), are in reasonable agreement with the best extrapolated value (-13.1 kcal/mol) at CBS(aD-aT-aQ). On the other hand, after BSSE corrections, the binding energies are all underestimated owing to the overestimation of BSSE error by the CP method.^{41,42} It should be noted that the BSSE-corrected ΔE_{c} with very large basis sets, -13.0 (aug-cc-pVQZ), -12.8 (aug-cc-pVTZ), and -12.5 (cc-pVQZ), are quite close to the CBS(aD-aT-aQ) value, but the values with basis sets smaller than cc-pVTZ (-11.3 kcal/mol) appear to be too small. In general, for basis sets larger than cc-pVTZ, the BSSE-corrected binding energies are in better agreement with the reliable extrapolated values than the uncorrected ones. For the methyl-blocked helical dimer ($\alpha\text{HH}'$), the CBS(T-Q) and CBS(D-T-Q) energies are very close, being -15.0 and -14.8 kcal/mol, respectively. These values are between the BSSE-corrected and -uncorrected MP2/cc-pVQZ values, -14.1 and -15.5 kcal/mol, respectively.

The average H \cdots O H-bond distances in αHH and $\alpha\text{HH}'$ are close (2.050 and 2.039 Å, respectively). However, the steric effect between the nearby methyl groups in $\alpha\text{HH}'$ prevents the alignment of the two H-bonds from being “parallel”, which is indicated by the four H-bond angles ($\angle\text{NHO} = 154.2^\circ$ and $\angle\text{COH} = 155.2^\circ$ for the left-hand side and $\angle\text{NHO} = 161.8^\circ$ and $\angle\text{COH} = 116.7^\circ$ for the right-hand side H-bond). Although the left-hand $\angle\text{CON}$ angle (not shown), 148.0° , is close to the average 155.0° obtained from the survey of X-ray structures in the Protein Data Bank (PDB),²⁷ the right-hand $\angle\text{CON}$ angle (not shown), 114.3° , deviates significantly. In contrast, the H-bond alignment in αHH is closer to that in the protein α helices, and both $\angle\text{CON}$ angles, 162.8° and 152.4° , respectively, are close to the PDB survey value (155.0°).²⁷ We note that an α hydrogen in $\alpha\text{HH}'$ approaches a nitrogen in another strand, and the distance between the two atoms is 2.58 Å. This does not occur in the unblocked helical complex αHH .

The $A\beta\beta\text{-C}_5$ ($A\beta\beta\text{-C}_5'$) and $A\beta\beta\text{-C}_7$ ($A\beta\beta\text{-C}_7'$) conformations represent the two types of H-bond pairs in the antiparallel β sheets. All levels of calculations show that the C_7 forms have much stronger binding energies than the C_5 forms. The extrapolated binding energies of the two unblocked dimers at the CBS(aD-aT-aQ) level are -11.3 and -19.2 kcal/mol, respectively, in comparison with -11.4 and -19.6 kcal/mol, respectively, at the CBS(aT-aQ) level. For $A\beta\beta\text{-C}_5'$ and $A\beta\beta\text{-C}_7'$, the CBS(D-T-Q) binding energies are -13.4 and -20.7 kcal/mol, respectively, which are almost identical to the CBS(T-Q) values of -13.2 and -20.8 kcal/mol.

The dimer geometries are consistent with the binding energies; the dimers with larger binding energies (i.e., $A\beta\beta\text{-C}_7$ and $A\beta\beta\text{-C}_7'$) have shorter H \cdots N H-bonds than $A\beta\beta\text{-C}_5$ and $A\beta\beta\text{-C}_5'$ (see Figure 2), respectively. Compared to the helical $\alpha\text{HH}'$ dimer, the methyl groups in $A\beta\beta\text{-C}_5'$ and $A\beta\beta\text{-C}_7'$

C_7' have limited influence on the alignment of two H-bonds, and the H-bond angles in the two antiparallel β dimers are very close (see Figure 2).

The complexes similar to $A\beta\beta-C_5'$ and $A\beta\beta-C_7'$ have been studied previously. For the convenience of direct comparison, the following discussion about the antiparallel conformations is based on the two methyl-blocked dimers, and they can be applied to their unblocked counterparts.

The substantial binding energy difference between $a\beta\beta-C_7'$ and $a\beta\beta-C_5'$, -7.3 kcal/mol at CBS(D-T-Q), has also been observed by others in spite of substantial differences in magnitude. However, the explanations have been somewhat controversial. Zhao and Wu³⁰ attributed the difference primarily to the two weak $C-H\cdots O=C$ H-bonds in $a\beta\beta-C_7'$ and the destabilization in $a\beta\beta-C_5'$ due to the repulsive O/O (representing the two O atoms of CO groups in the two paired H-bonds) and H/H (representing the two H atoms of NH groups in the two paired H-bonds) secondary interactions in the H-bond pairs across the two monomers. In contrast, Dannenberg and co-workers³⁴ attributed the difference mainly to the intrastrand $C_5 O\cdots H$ interaction and considered that the reported $H\cdots O$ distance of the $C-H\cdots O=C$ H-bond (2.855 Å in Zhao and Wu's work³⁰) seems too long for such a H-bond. However, in the present MP2/6-311+G** structure of $A\beta\beta-C_7'$, the two $C-H\cdots O=C$ H-bond distances, 2.462 and 2.475 Å, respectively, are substantially shorter than the 2.855 Å at the HF/6-31G* level reported previously,³⁰ strongly indicating the existence of the $C-H\cdots O=C$ H-bonds.

Vargas et al.^{21,61} computationally estimated the $C-H\cdots O=C$ H-bond energy to be about 2.1 kcal/mol. Thus, the two $C-H\cdots O=C$ H-bonds in $A\beta\beta-C_7'$ may strengthen the binding by about 4.2 kcal/mol, which could be one of the sources of the ~ 7.3 kcal/mol difference at the CBS(D-T-Q) level observed in this study. On the other hand, the intrastrand $C_5 O\cdots H$ distance in $A\beta\beta-C_7'$ (2.391 Å) is substantially shorter than the 2.701 Å found in $A\beta\beta-C_5$, suggesting the non-negligible role of the $C_5 O\cdots H$, although we are unable to quantify the "pure" $C_5 O\cdots H$ interaction energy from the QM calculations. The energy difference between the two monomers in $A\beta\beta-C_5'$ and $A\beta\beta-C_7'$ is 0.3 kcal/mol, but the small difference does not imply that the $C_5 O\cdots H$ interaction plays a minor role because the favorable $C_5 O\cdots H$ interaction could be canceled by other unfavorable factors. The deformation energies (the energy difference between the monomers in the complex and the isolated one) of $A\beta\beta-C_5'$ and $A\beta\beta-C_7'$ are 1.2 and 0.6 kcal/mol, respectively, which contribute 0.6 kcal/mol to the difference. Dannenberg and co-workers³⁴ also studied the complexes similar to $a\beta\beta-C_5'$ and $a\beta\beta-C_7'$ but with a restrained C_s symmetry at the level of B3LYP/D95(D,P) + BSSE correction. The reported H-bond energies, -4.9 and -14.0 kcal/mol, respectively, are significantly different from our CBS(D-T-Q) values of -13.4 and -20.7 kcal/mol.

The difference between our results and those of others^{30,34} could originate from two main sources. The first could be the different levels of theory used in the geometry optimizations and energy calculations. The second could arise from the symmetry restraint used by Zhao and Wu³⁰ and by

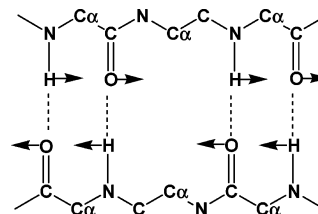


Figure 3. Moving tendency of atoms as the antiparallel β sheets are formed.

Dannenberg and co-workers.³⁴ For computational efficiency, they constrained the (Φ , Ψ) main-chain torsion angles to 180.0° to maintain the C_s symmetry, which might have contributed to their too-long $C-H\cdots O=C$ distance (2.855 Å). In comparison, the (Φ , Ψ) torsion angles in our model complexes were constrained to the typical values adopted in the antiparallel β sheets in proteins. The overestimation of BSSE corrections and the lack of accuracy of the B3LYP/D95(D,P) method in accounting for the nonbonded interactions further contribute to the much smaller binding energies reported by Dannenberg and co-workers.³⁴

In the methyl-blocked $A\beta\beta-C_5'$ dimer, because the crossing O/O secondary repulsion is larger than the crossing H/H repulsion, the two carbonyl O's inevitably moved away and two amidic H's moved closer. As a consequence, the O/O distance, 3.684 Å, is larger than the 2.928 Å of the H/H distance. In contrast, because the H-bonds in $A\beta\beta-C_7'$ are separated by four bonds (the crossing O/O distance is 4.174 Å), the secondary repulsions are weaker, allowing the carbonyl groups to move closer to one of the α -hydrogen atoms in the crossing strand to form a $C-H\cdots O=C$ H-bond. It is interesting to note that the movements in $a\beta\beta-C_5'$ and $a\beta\beta-C_7'$ are concerted (Figure 3), and therefore, the two patterns of H-bond pairs can be combined without introducing excess strain as they coexist in a long antiparallel β sheet.

The unblocked $P\beta\beta$ complex represents the unique H-bond pattern in the parallel β sheets. Its extrapolated binding energies are -15.7 kcal/mol [CBS(T-Q)], -15.6 kcal/mol [CBS(D-T-Q)], and -15.0 kcal/mol [CBS(aT-aQ)], converging to -14.8 kcal/mol at the CBS(aD-aT-aQ) level. The average H-bond length in $P\beta\beta$, 1.907 Å, is compared with 1.918 Å in $A\beta\beta-C_7$ and 1.933 Å in $A\beta\beta-C_5$. The MP2/6-311+G** optimized structures apparently indicate the existence of $C-H\cdots O=C$ H-bonds in the parallel β sheet. But the longer $C-H\cdots O=C$ distances, 2.571 Å in $P\beta\beta$ and 2.638 Å in $P\beta\beta'$, than the 2.423 Å and 2.422 Å in $A\beta\beta-C_7$ and 2.462 Å and the 2.475 Å in $A\beta\beta-C_7'$, respectively, suggest that the $C-H\cdots O=C$ H-bond in the parallel β may be weaker than those in the antiparallel β sheets. The ordering of the binding energies of the three β -sheet dimers at the CBS-(aD-aT-aQ) level, -11.3 ($A\beta\beta-C_5$), -14.8 kcal/mol ($A\beta\beta-C_7$), and -19.2 kcal/mol ($P\beta\beta$), is consistent with the fact that there are zero, one, and two $C-H\cdots O=C$ H-bonds in $A\beta\beta-C_5$, $P\beta\beta$, and $A\beta\beta-C_7$, respectively. The methyl-blocked β -sheet dimers follow the same ordering, -13.4 ($A\beta\beta-C_5'$), -16.7 ($P\beta\beta'$), and -20.7 kcal/mol ($A\beta\beta-C_7'$), respectively, at the CBS(D-T-Q) level.

In the study on the cooperativity of H-bonds in the β sheets, Zhao and Wu³⁰ have optimized the C_s -restrained

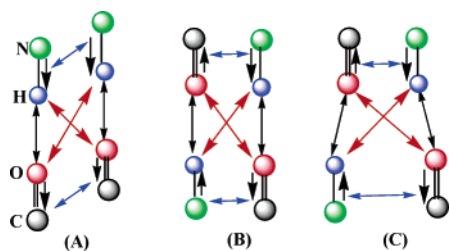


Figure 4. Schematic illustration of primary, secondary, and tertiary interactions in helix (A), antiparallel (B), and parallel (C) β -sheet dimers. The black \rightarrow lines represent the induced dipoles. The black, red, and blue \leftrightarrow lines represent the primary, secondary, and tertiary interactions, respectively.

ACE-(GLY)₂-NH₂ dimers in which Φ and Ψ were fixed at 180.0° to model the H-bonds in the parallel and antiparallel β sheets. Their HF/6-31G* optimization led to significantly different H-bond lengths in the antiparallel and parallel β -sheet conformations; the two types of H-bonds in the parallel- β -like dimer, 2.618 Å and 2.653 Å, respectively, were substantially longer than 2.150 Å in the antiparallel- β -like dimer, leading the authors to conclude that the H-bonds in the parallel β sheet may be quite weaker.³⁰ In contrast, the geometries optimized at the current MP2/6-311+G** level reveal that the H-bond lengths in the parallel β conformations are comparable to those in the C₅ and C₇ conformations of the antiparallel β sheets (the differences are less than 0.05 Å). It should be noted that the survey of protein crystal structures in PDB showed that the average O/N distance in the parallel β sheet H-bonds, 2.905 Å, is actually slightly shorter than 2.925 Å in the antiparallel β sheets. This is consistent with our results at the correlation level that the average O/N distance in P $\beta\beta'$, 2.896 Å, is also slightly shorter than the average (2.918 Å) in A $\beta\beta$ -C_{5'} and A $\beta\beta$ -C_{7'}. The calculated binding energies show no obvious preference for the H-bond in antiparallel β sheets; the parallel β H-bonds are weaker than those in the antiparallel β C₇ forms but are stronger than those in the antiparallel C₅ form and are comparable to the average of the two antiparallel forms.

On the basis of the survey of the X-ray structures of proteins, Derewenda et al.⁶² also suggested the existence of the C–H \cdots O=C H-bonds in parallel β sheets. This is supported by the present study at the correlation level. In contrast, the earlier HF/6-31G* study was not able to uncover such interactions in the C₅-restrained ACE-(GLY)₂-NH₂ dimers. In spite of the differences between ours and previous model complexes, the different pictures that emerged from the studies with regard to the structures and interaction energies underscore the need of high levels of theory for reliably characterizing the H-bonds.

The carbonyl and amide groups in both strands of A $\beta\beta$ -C₅ (or A $\beta\beta$ -C_{5'}) are separated by two bonds. Intuitively, one might anticipate that this characteristic enhances the H-bond directionality and strength. In comparison, the carbonyl and amide groups in P $\beta\beta$ (or P $\beta\beta'$) are separated by four bonds in one strand and by two in the other, making them possibly misaligned. Therefore, the A $\beta\beta$ -C₅ (or A $\beta\beta$ -C_{5'}) H-bonds could be conceivably stronger than the P $\beta\beta$ (or P $\beta\beta'$)

H-bonds. This view has led to the popular belief that the perceived distortion to the H-bond geometry in the parallel β sheet weakens its H-bonds.⁶³ As discussed above, the current study reveals a more complex picture; the N–H \cdots O=C H-bonds, C–H \cdots O=C H-bonds, crossing O/O and H/H repulsions (i.e., the secondary interactions), and the aforementioned tertiary effect all contribute. Because antiparallel β C₅ and C₇ coexist in antiparallel β sheets, as an approximation, we compare the average binding energies of the two antiparallel β conformations with those of parallel β ones. At the CBS(aD-aT-aQ) level, the average energy, –15.3 kcal/mol, is only slightly larger than the –14.9 kcal/mol of P $\beta\beta$. Similarly, the average value of –17.1 kcal/mol of A $\beta\beta$ -C_{5'} and A $\beta\beta$ -C_{7'} at CBS(D-T-Q) is also comparable to the –16.7 kcal/mol of P $\beta\beta$. Although the H-bonds in C₅ forms have the potential to form “ideal” H-bonds, the secondary O/O and H/H repulsions prevent them from doing so. In contrast, the inherent long crossing O/O and H/H distances in parallel β forms let them suffer less unfavorable secondary repulsions, and the H-bonds are able to improve the linearity with minor adjustment without introducing repulsive secondary interactions; the two groups that are separated by two bonds move outward while the groups connected by four bonds move inward. As a consequence, the H-bond linearity in the parallel β sheet is actually better than that in the antiparallel β and C₅ forms, as indicated by the N–H \cdots O H-bond angles shown in Figure 2.

3.2. Quality of Additive Force Fields in Modeling H-Bond Pairs. The large size of biological molecules (e.g., proteins) and the complexity of biological processes (e.g., protein folding) impose a tremendous limitation to the application of quantum-mechanics-based methods in the studies of the biological systems and processes. Molecular-mechanics-based modeling is an affordable alternative. In molecular mechanics calculations, the force fields underlie all modeling approaches and their quality is essential. In the development of protein force fields, appropriate representation of the backbone H-bonds is one of the most important concerns. While the individual backbone H-bond often exists as components of an ordered H-bond network, the typical approach in the parameter development and calibration is to compare with the single H-bond in the NMA–NMA dimer which, as mentioned above, is incapable of representing the neighboring effect between nearby H-bonds. Although the additive force fields do include the crossing O/O and H/H secondary electrostatic interactions, the suitability of this approach in describing the H-bond network in protein secondary structures has not been clarified. The results reported in this study can serve as the benchmarks to examine the existing additive force fields and reference data for future force field development.

Table 2 compares the ab initio energies of the two sets of dimers with the various empirical force fields. The partial charges for the unblocked dimers in AMBER ff94 force fields¹⁸ were refitted using the same strategy as in ff94; they were fitted to the electrostatic potentials of the monomer calculated at the HF/6-31G**//HF/6-31G* level by using the restricted electrostatic potential (RESP) approach.⁶⁴ The other parameters, including those for bonds, angles, torsions, and

Table 2. H-Bond Energies (in kcal/mol) of H-Bond Pair Complexes, Their Relative Values to That of α Helical Conformations, and the Contributions (E_{pol}) Due to Polarization in the Polarizable Force Fields^a

	$\Delta E(\alpha\text{HH})$	$\Delta E(A\beta\beta\text{-}C_5)$	$\Delta E(A\beta\beta\text{-}C_7)$	$\Delta E(P\beta\beta)$	$\frac{\Delta E(A\beta\beta\text{-}C_5) - \Delta E(\alpha\text{HH})}{\Delta E(\alpha\text{HH})}$	$\frac{\Delta E(A\beta\beta\text{-}C_7) - \Delta E(\alpha\text{HH})}{\Delta E(\alpha\text{HH})}$	$\frac{\Delta E(P\beta\beta) - \Delta E(\alpha\text{HH})}{\Delta E(\alpha\text{HH})}$	$\frac{\Delta E(A\beta\beta\text{-}C_7) - \Delta E(A\beta\beta\text{-}C_5)}{\Delta E(A\beta\beta\text{-}C_5)}$
QM	-13.1	-11.3	-19.2	-14.8	1.8	-6.1	-1.7	-7.9
AMBER-FF94	-16.5 (-3.4)	-12.8 (-1.5)	-17.9 (1.3)	-15.0 (-1.0)	3.8	-0.6	1.7	-5.1
OPLS_AA	-17.4 (-4.3)	-10.9 (0.4)	-18.8 (0.4)	-14.8 (0.0)	6.5	-1.4	2.6	-7.9
CHARM19-UA	-17.5 (-4.4)	-13.6 (-2.3)	-14.9 (4.3)	-14.1 (0.7)	3.9	-2.5	3.4	-1.3
FF94+Pol	-14.6 (-1.5)	-13.9 (-2.6)	-17.6 (1.6)	-15.9 (-1.1)	0.7	-3.0	-1.3	-3.7
E_{Pol}	1.7	-1.1	-0.2	-1.1				
QM	-14.8	-13.4	-20.7	-16.7	1.4	-5.9	-1.9	-7.3
FF94	-18.6 (-3.8)	-14.0 (-0.6)	-17.8 (2.9)	-15.7 (1.0)	4.6	-0.8	2.9	-3.8
FF03	-17.2 (-2.4)	-12.1 (1.3)	-17.2 (3.5)	-14.2 (2.5)	5.1	-0.0	3.0	-5.1
CHARM19	-19.0 (-4.2)	-15.2 (-1.8)	-16.2 (4.5)	-16.6 (0.1)	3.8	2.8	2.4	-1.0
CHARM27	-18.6 (-3.8)	-13.5 (-0.1)	-17.2 (3.5)	-15.4 (1.3)	5.1	1.4	3.2	-3.7
OPLS-AA	-18.2 (-3.4)	-12.9 (0.5)	-17.5 (3.2)	-15.2 (1.5)	5.3	0.7	3.0	-4.6
OPLS-UA	-20.9 (-6.1)	-14.7 (-1.3)	-18.7 (2.0)	-17.3 (-0.6)	6.2	2.2	3.6	-4.0
FF02 (Pol)	-17.6 (2.8)	-14.8 (-1.4)	-19.5 (1.2)	-17.5 (-0.8)	2.8	-1.9	0.1	-4.7
E_{Pol}	1.7	-1.0	-0.6	-1.0				
FF94+Pol	-17.3 (2.5)	-15.2 (-1.8)	-17.0 (3.7)	-16.7 (0.0)	2.1	0.3	0.6	-1.8
E_{Pol}	1.4	-1.2	-0.7	-1.0				
FF03+Pol	-15.7 (0.9)	-13.8 (-0.4)	-16.8 (4.1)	-15.2 (1.5)	1.9	-1.1	0.5	-3.0
E_{Pol}	1.0	-0.8	-0.7	-0.8				

^a The differences relative to the QM energies are given in the parentheses.

Lennard-Jones, were taken from AMBER ff94, whereas the charges for the unblocked dimers in AMBER ff02 and ff03 force fields were refitted using the strategies consistent with those of ff02⁵³ and ff03.⁵² Despite some striking agreements between the ab initio and the MM binding energies in Table 2, attention should be paid to the balance among different conformations. It is often the case that, for a given force field, good agreement with the ab initio values can be found for some dimers but not for the rest. However, in spite of the different behavior of these force fields, they all share one common feature: significantly overestimating the binding energy in the helical dimers. For example, the ab initio data shows that the binding energy of $A\beta\beta\text{-}C_5$ is 1.8 kcal/mol stronger than that of αHH . Yet, all force fields favor the helical conformation by 3.8–6.5 kcal/mol. The same holds for the methyl-blocked complexes. Given the fact that these force fields were developed by different groups on the basis of different strategies, we attribute the common feature to the inherent deficiency of the additive (point charge) molecular mechanics models.

Figure 4 schematically illustrates the major contributions to the binding energies accounted for by an additive force field. In addition to the contributions due to van der Waals interactions and deformations, the electrostatic interactions are the dominant components of the binding energies. The helical form has two primary (represented by the black double-headed lines in Figure 4) and two favorable secondary (represented the red double-headed lines in Figure 4) interactions, but the two secondary interactions in the β forms are unfavorable. Because both helical and β -sheet dimers have similar amidic H-bond donors and acceptors, the primary interactions are approximately the same. Therefore, the interactions in helical dimers tends to be stronger than that in the corresponding β forms if no other interactions are involved (e.g., the $\text{C-H}\cdots\text{O}=\text{C}$ H-bond). This is in

qualitative agreement with the ab initio results; the binding energy of αHH is 1.8 kcal/mol larger than that of $A\beta\beta\text{-}C_5$. But the substantial overestimations by force fields clearly indicate the limitation of the additive force field. We next consider the contribution due to the deformation. At the MP2/aug-cc-pVQZ level, the deformation energies of αHH and $A\beta\beta\text{-}C_5$ are 0.8 and 1.4 kcal/mol, respectively. Therefore, if the deformation contribution was excluded, the binding energy difference between the two conformations would become even smaller, which indicates severe overestimation by the various force fields.

In the following, we suggest possible explanations for the significant disagreement. Because the antiparallel- β C_7 and parallel- β conformations involve $\text{C-H}\cdots\text{O}=\text{C}$ H-bonds, which complicates the analysis, they are excluded in the following discussion.

A well-known defect in the additive force fields is the omission of the instantaneous polarization. As illustrated in Figure 4, when a H-bond forms, the donor and acceptor of the H-bond polarize their partners, making the polar $\text{C}=\text{O}$ and N-H groups more polar in comparison to those in the monomer. In helical conformation, the enhanced polarization increases the energetically unfavorable intrastrand repulsions between the CO and CO groups in one strand and between NH and NH groups in the other (indicated by the blue double-headed lines in Figure 4). In contrast, this effect strengthens the favorable intrastrand attractions between the NH and CO in both strands of the $A\beta\beta\text{-}C_5$. Obviously, the neglect of the effect results in the overestimation of binding energy in αHH and the underestimation of binding energy in $A\beta\beta\text{-}C_5$, which could be one of the sources for the large disagreement between MM and ab initio data. Following Jorgensen and Pranata,³⁷ we call this effect tertiary interaction. Due to the overestimation, the point-charge force fields also give wrong relative binding energies of αHH to $P\beta\beta$.

Because of a favored C–H···O=C H-bond in the parallel β conformation, as predicted by ab initio results, the binding of the parallel β conformation is 1.7 kcal/mol stronger than the helical conformation α HH at the CBS(aD-aT-aQ) level. In contrast, all force fields predict the former to be 1.7–3.4 kcal/mol weaker than the latter. The relative binding energies of $A\beta\beta$ - C_7 to the helix form are also substantially underestimated (see the column 7).

Because the torsions were fixed in both monomers and complexes, the binding energies do not reflect the contribution of (Φ , Ψ) torsion energies. The similar behavior of various additive force fields only applies to the electrostatic (H-bond) interactions and does not reflect the overall behavior of these force fields. Because (Φ , Ψ) torsion energies are also important for the conformational preference, the overestimation of the binding energies of helical over β conformations does not necessarily imply that all additive force fields are biased to the helical conformations. Nevertheless, appropriate modeling of the H-bond pairs is necessary for accurate and balanced protein force fields.

Various nonadditive polarizable force fields have been pursued in several groups to account for the instantaneous polarization.^{53,54,65–73} Because most of them are not publicly available, Table 2 only gives the binding energies calculated by the AMBER force fields including one specifically designed as a polarizable force field (ff02)⁵³ and those developed as fixed-charge models with ad hoc addition of the polarizability (ff94+pol, ff03+pol); the latter two are included solely for the purposes of comparison. Table 2 also gives the contributions of polarization to the binding energies.

After turning on the polarization in the additive AMBER force fields, the polarization effect systematically weakens the interactions in α HH and α HH' dimers and strengthens the binding in the β -sheet dimers. This is consistent with the positive polarization contributions in the former and negative ones in the latter. The unanimous positive contributions of polarization (ca. 1.7 kcal/mol) in the helical conformations indicate that the polarization is energetically unfavorable to the H-bond, which is consistent with the earlier reasoning that the tertiary effect is energetically unfavorable in helical conformations. In contrast, the polarization effect in all β forms is energetically favorable (–0.2~–1.2 kcal/mol), which is consistent with the notion that the tertiary effect enhances the binding in β dimers. As a consequence, the balance between the helix and the β -sheet conformers is notably improved in comparison with the ab initio results. For the unblocked dimers, the binding energy differences (relative to α HH) of $A\beta\beta$ - C_5 , $A\beta\beta$ - C_7 , and $P\beta\beta$, being 0.7, –3.0, and –1.3 kcal/mol, calculated by polarizable force field (ff94+pol), are compared to the ab initio values, 1.8, –6.1, and –1.7 kcal/mol, respectively, and are notably better than 3.8, –0.6, and 1.7 kcal/mol, calculated by additive force field ff94. Similar improvements can be observed among the blocked dimers after turning on the polarization in ff94 and ff03. It is noteworthy that both ff94 and ff03 were designed as the fixed-charge models (i.e., without the polarizability). Thus, the improvement is indicative of the positive roles that the polarizability plays in describing the main-chain H-bonds. Furthermore, because the two H-bonds

in antiparallel C_7 forms are farther apart than those in antiparallel C_5 and parallel β forms, the polarization contributes more in the latter than in the former. The polarization energies of the former C_7 is less than that of the former C_5 's (see Table 2). The detailed comparisons between the polarizable force field (ff02) and the ab initio data also indicate that further improvements are necessary to obtain the correct absolute and relative binding energies.

Because of the favorable C–H···O=C H-bonds in the antiparallel β C_7 forms and the unfavorable crossing secondary interactions in the C_5 forms, the binding energy of the C_7 forms are much larger (more than 7.0 kcal/mol) than those of the C_5 forms. However, except for the OPLS-AA, all other force fields, including both additive and polarizable force fields, underestimate the relative binding energies of the unblocked C_7 to C_5 dimers. Among many possible factors, a lack of consideration of the C–H···O=C H-bonds in parametrizations may take the main responsibility.

4. Conclusions

The unblocked and methyl-blocked glycine dipeptide dimers, which were arranged to model the four patterns of backbone H-bond pairs in the protein secondary structures, have been investigated by ab initio calculations. The study provides reference structures and energetics for characterizing the protein backbone H-bonds. On the basis of the structures optimized at the MP2/6-311+G** level and the energies at various high levels, the following conclusions can be drawn.

In addition to the conventionally concerned primary N–H···O=C H-bonds and the crossing secondary interactions, the C–H···O=C H-bonds and the other neighboring effect (e.g., tertiary effect) also contribute substantially. Unlike previous HF/6-31G* optimization in which the C–H···O=C H-bond can only be observed in the antiparallel β -sheet-like complex, the current MP2/6-311+G** optimization demonstrates that the C–H···O=C H-bonds exist in both parallel and antiparallel β -sheet-like conformers, which is in agreement with the PDB survey study.

The best extrapolated binding energies [CBS(aD-aT-aQ)] of unblocked dimers are –13.1, –11.3, –19.2, and –14.8 kcal/mol, and the best values [CBS(D-T-Q)] for the methyl-blocked dimers are –14.8, –13.4, –20.7, and –16.7 kcal/mol, respectively. Because the binding energies of parallel β -sheet conformations are only marginally weaker than the average of the two antiparallel β -sheet conformations, we conclude that the H-bond energies in the parallel and antiparallel β sheets are comparable. Consistently, the H-bond lengths in the two types of conformations are very close. This conclusion is different from the previous views, which concluded that the H-bond interaction in the parallel β sheet could be weaker than that in the antiparallel β sheets on the basis of the HF/6-31G** optimization.

The secondary interactions, which are included in the additive force fields, are not able to account for the neighboring effects completely. Because other neighboring effects such as tertiary effect are also important, all additive force fields significantly overestimate the interactions in the helical conformations with respect to the β -sheet conformations. For instance, the energy difference between α HH and

$\text{A}\beta\text{C}_5$, ranging from 3.8 to 6.5 kcal/mol, estimated by various force fields, is much larger than the ab initio value 1.8 kcal/mol. However, after inclusion of the polarization in the AMBER conventional force fields, the agreement with ab initio results is notably improved, which shows the promise of polarizable force fields to account for such interactions.

Acknowledgment. This work was supported by research grants from NIH (GM64458, GM67168, GM079383 to Y.D.). Dr. P. Cieplak provided the ff02 partial charges of unblocked glycine dipeptide.

References

- (1) Carl, B.; Toozee, J. *Introduction to Progein Structure*, 2nd ed.; Garland Publishing, Inc: New York, 1999.
- (2) Nelson, D. L.; Cox, M. M. *Principles of Biochemistry*, 4th ed.; W. H. Freeman and Company: New York, 2005.
- (3) Karplus, M. *Acc. Chem. Res.* **2002**, *35*, 321.
- (4) Dessent, C. E. H.; Muller-Dethlefs, K. *Chem. Rev.* **2000**, *100*, 3999.
- (5) Dykstra, C. E. *Chem. Rev.* **1993**, *93*, 2339.
- (6) Hobza, P.; Havlas, Z. *Chem. Rev.* **2000**, *100*, 4253.
- (7) Kollman, P.; Allen, L. C. *Chem. Rev.* **1972**, *72*, 283.
- (8) Muller-Dethlefs, K.; Hobza, P. *Chem. Rev.* **2000**, *100*, 143.
- (9) Neusser, H. J.; Siglow, K. *Chem. Rev.* **2000**, *100*, 3921.
- (10) Schermann, J. P.; Carles, S.; Desfrancois, C. *Chem. Rev.* **2000**, *100*, 3943.
- (11) Neusser, H. J.; Siglow, K. *Chem. Rev.* **2000**, *100*, 3921.
- (12) Sherrington, D. C.; Taskinen, K. A. *Chem. Soc. Rev.* **2001**, *30*, 83.
- (13) Jeffrey, G. A. *An Introduction to Hydrogen Bonding*; Oxford University Press: New York, 1997.
- (14) Scheiner, S. *Hydrogen Bonding; A Theoretical Perspective*; Oxford University Press: New York, 1997.
- (15) Perczel, A.; Imre, J.; Csizmadia, I. G. *Chem.—Eur. J.* **2003**, *9*, 5332.
- (16) Perczel, A.; Gaspari, Z.; Csizmadia, I. G. *J. Comput. Chem.* **2005**, *26*, 1155–1168.
- (17) Mackerell, A. D.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 10559.
- (18) 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.
- (19) Langley, C. H.; Allinger, N. L. *J. Phys. Chem. A* **2003**, *107*, 5208.
- (20) Dixon, D. A.; Dobbs, K. D.; Valentini, J. J. *J. Phys. Chem.* **1994**, *98*, 13435.
- (21) Vargas, R.; Garza, J.; Friesner, R. A.; Stern, H.; Hay, B. P.; Dixon, D. A. *J. Phys. Chem. A* **2001**, *105*, 4963.
- (22) Guo, H.; Karplus, M. *J. Phys. Chem.* **1994**, *98*, 7104.
- (23) Jorgensen, W. L.; Swenson, C. J. *J. Am. Chem. Soc.* **1985**, *107*, 1489.
- (24) Jorgensen, W. L.; Swenson, C. J. *J. Am. Chem. Soc.* **1985**, *107*, 569.
- (25) Wang, Z.-X.; Duan, Y. *J. Theor. Comput. Chem.* **2005**, *4*, 689.
- (26) Deechongkit, S.; Dawson, P. E.; Kelly, J. W. *J. Am. Chem. Soc.* **2004**, *126*, 16762.
- (27) Koch, O.; Bocola, M.; Klebe, G. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 310.
- (28) Lario, P. I.; Vrieling, A. *J. Am. Chem. Soc.* **2003**, *125*, 12787.
- (29) Wu, Y. D.; Zhao, Y. L. *J. Am. Chem. Soc.* **2001**, *123*, 5313.
- (30) Zhao, Y. L.; Wu, Y. D. *J. Am. Chem. Soc.* **2002**, *124*, 1570.
- (31) Wiczorek, R.; Dannenberg, J. J. *J. Am. Chem. Soc.* **2003**, *125*, 8124.
- (32) Wiczorek, R.; Dannenberg, J. J. *J. Am. Chem. Soc.* **2003**, *125*, 14065.
- (33) Kobko, N.; Dannenberg, J. J. *J. Phys. Chem. A* **2003**, *107*, 10389.
- (34) Viswanathan, R.; Asensio, A.; Dannenberg, J. J. *J. Phys. Chem. A* **2004**, *108*, 9205.
- (35) Spomer, J.; Jurecka, P.; Hobza, P. *J. Am. Chem. Soc.* **2004**, *126*, 10142.
- (36) Jurecka, P.; Hobza, P. *J. Am. Chem. Soc.* **2003**, *125*, 15608.
- (37) Jorgensen, W. L.; Pranata, J. *J. Am. Chem. Soc.* **1990**, *112*, 2008.
- (38) Lukin, O.; Leszczynski, J. *J. Phys. Chem. A* **2002**, *106*, 6775.
- (39) Asensio, A.; Kobko, N.; Dannenberg, J. J. *J. Phys. Chem. A* **2003**, *107*, 6441.
- (40) Rappe, A. K.; Bernstein, E. R. *J. Phys. Chem. A* **2000**, *104*, 6117.
- (41) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.
- (42) Simon, S.; Duran, M.; Dannenberg, J. J. *J. Chem. Phys.* **1996**, *105*, 11024.
- (43) 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.
- (44) Head-Gordon, T.; Head-Gordon, M.; Frisch, M. J.; Brooks, C. L.; Pople, J. A. *J. Am. Chem. Soc.* **1991**, *113*, 5989.
- (45) Iwaoka, M.; Okada, M.; Tomoda, S. *THEOCHEM* **2002**, *586*, 111.
- (46) Mackerell, A. D.; Feig, M.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1400.
- (47) Wang, Z. X.; Duan, Y. *J. Comput. Chem.* **2004**, *25*, 1699.

- (48) Truhlar, D. G. *Chem. Phys. Lett.* **1998**, 294, 45.
- (49) Fast, P. L.; Sanchez, M. L.; Truhlar, D. G. *J. Chem. Phys.* **1999**, 111, 2921.
- (50) Feyereisen, M. W.; Feller, D.; Dixon, D. A. *J. Phys. Chem.* **1996**, 100, 2993.
- (51) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *J. Am. Chem. Soc.* **1993**, 115, 9620.
- (52) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. *J. Comput. Chem.* **2003**, 24, 1999.
- (53) Cieplak, P.; Caldwell, J.; Kollman, P. A. *J. Comput. Chem.* **2001**, 22, 1048.
- (54) Wang, Z.-X.; Zhang, W.; Wu, C. H. L.; Cieplak, P.; Duan, Y. *J. Comput. Chem.* **2006**, 27, 781.
- (55) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, 4, 187.
- (56) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, 102, 3586.
- (57) Jorgensen, W. L.; Tiradorives, J. *J. Am. Chem. Soc.* **1988**, 110, 1666.
- (58) Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. *J. Am. Chem. Soc.* **1996**, 118, 11225.
- (59) Case, D. A.; Darden, T. A. T. E.; Cheatham, I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 8*; University of California: San Francisco, 2004.
- (60) Ponder, J. W. *Software Tools for Molecular Design*, 4.2 ed.; Washington University School of Medicine: St. Louis, MO, 2004.
- (61) Vargas, R.; Garza, J.; Dixon, D. A.; Hay, B. P. *J. Am. Chem. Soc.* **2000**, 122, 4750.
- (62) Derewenda, Z. S.; Lee, L.; Derewenda, U. *J. Mol. Biol.* **1995**, 252, 248.
- (63) Voet, D.; Voet, J. G. *Biochemistry*, 2nd ed.; John Wiley & Sons: New York, 1995; p 150.
- (64) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, 97, 10269.
- (65) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, 101, 6141.
- (66) Stern, H. A.; Kaminski, G. A.; Banks, J. L.; Zhou, R. H.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. B* **1999**, 103, 4730.
- (67) Stuart, S. J.; Berne, B. J. *J. Phys. Chem.* **1996**, 100, 11934.
- (68) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X. X.; Murphy, R. B.; Zhou, R. H.; Halgren, T. A. *J. Comput. Chem.* **2002**, 23, 1515.
- (69) Patel, S.; Brooks, C. L. *J. Comput. Chem.* **2004**, 25, 1.
- (70) Patel, S.; Brooks, C. L. *J. Chem. Phys.* **2005**, 122, PAGE.
- (71) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D. *J. Chem. Theory Comput.* **2005**.
- (72) Ren, P. Y.; Ponder, J. W. *J. Comput. Chem.* **2002**, 23, 1497.
- (73) Ren, P. Y.; Ponder, J. W. *J. Phys. Chem. B* **2003**, 107, 5933.

CT700021F

Establishing Effective Simulation Protocols for β - and α/β -Mixed Peptides. I. QM and QM/MM Models

Xiao Zhu, Arun Yethiraj,* and Qiang Cui*

Department of Chemistry and Theoretical Chemistry Institute, University of Wisconsin—Madison, 1101 University Avenue, Madison, Wisconsin 53706

Received December 18, 2006

Abstract: A quantum mechanical (QM) model for non-natural β - and α/β -mixed peptides is investigated using an approximate density functional method (called SCC-DFTB). In the gas phase the predictions of the model for cyclic and acyclic dipeptides and several acyclic heptapeptides are compared to ab initio B3LYP and LMP2 calculations. The SCC-DFTB reproduces the global minimum of the configurations with the root-mean-square (rms) error in the key dihedral angles of less than 14 degrees. The relative energies of different conformers are also well described in general, with the typical rms error of 2–3 kcal/mol relative to LMP2 energies at either B3LYP or LMP2 optimized structures. The dipole moments are reproduced with a systematic underestimate of less than 15%. The QM model is also used with a molecular mechanical (MM) model of the solvent. For a tetrameric α/β -peptide in water, the SCC-DFTB/MM energies are well correlated with B3LYP/6-31+G**/MM single point energies for a wide range of structures sampled in 2 ns of SCC-DFTB/MM molecular dynamics. For an octameric α/β -peptide in methanol the predicted structures are in qualitative agreement with experimental NOE data. These results suggest that the SCC-DFTB model provides a fairly accurate representation of the structure and thermodynamics of these peptides.

I. Introduction

The primary building blocks of naturally occurring α -peptides are amino acid residues which consist of a peptide bond and an α -carbon atom which can have a side chain. In β -peptides there is an additional carbon atom along the peptide backbone. The presence of two carbon atoms allows one to introduce cyclic residues along the backbone, something that is not possible in α -peptides. This class of relatively new materials has attracted enormous interest lately. This interest derives partly from their unique structural properties and partly from their potential in biomedical and material applications. β -Peptides and α/β -mixed peptides are interesting from a structural perspective because they form secondary structures (e.g., helices, sheets, reverse turns) more readily than natural α -peptides.^{1–10} They provide an interesting alternative to conventional peptides in many applications and have the advantage that there is no mechanism in the

body for their degradation. Non-natural peptides may have applications as antimicrobial materials^{11–16} and gene delivery agents¹⁷ and are possible candidates for lung surfactant mimics. Barron and co-workers¹⁸ have shown that designed peptoid oligomers (N-alkyl-glycine oligomers) can function as lung surfactant mimics. From a fundamental standpoint, the ability to control the chemical composition of non-natural peptides provides a unique opportunity for exploring the role of microscopic properties (e.g., chain stiffness) in determining the phase behavior and other macroscopic properties of polymeric materials.

A central question is the relation between structure and property in these materials. Characterizing the structure of these molecules experimentally is challenging, and quantitative structural information is scarce. There are significant thermal fluctuations in the structure, and this makes the interpretation of, e.g., Nuclear Overhauser Effect (NOE) and Circular Dichroism (CD) experiments far from straightforward.¹⁹ Therefore, much remains to be learned regarding how

* Corresponding author e-mail: yethiraj@chem.wisc.edu (A.Y.) and cui@chem.wisc.edu (Q.C.).

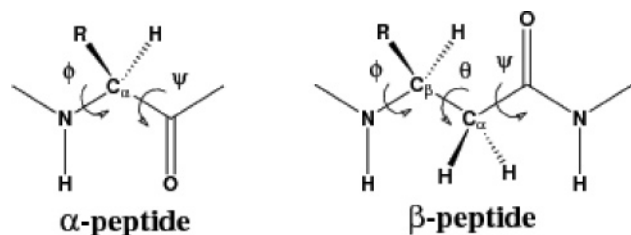


Figure 1. Comparison of α -(left) and β -(right) peptides. The torsional angles essential for characterizing the backbone structure of the system are ϕ and ψ for α -peptides and ϕ , θ , and ψ for β -peptides.

the structure and dynamics of β -peptides depend on their sequence. The goal of our research is to develop accurate models that can predict the structure of peptides in solution from first principles calculations. In this paper we investigate an approximate quantum mechanical (QM) model for non-natural peptides in the gas phase and, using this result, study a hybrid quantum mechanical/classical mechanical (QM/MM) model for peptides in solution.

There have been several experimental studies on the structure of β -peptides^{4–8,14,20–23} and α/β -mixed peptides,^{9,10} and these have led to some tentative empirical rules relating sequence to structure. For example, the 14-helix (the number 14 refers to the number of atoms between the two hydrogen-bonded moieties) is a preferred conformation for β^3 -peptides in organic solvents,^{7,20} and cyclic residues in general can greatly enhance the helical propensity of non-natural peptides.^{4–6,10} Electrostatic interactions in the forms of intrahelical salt-bridges^{21,22} and helical dipoles²⁴ enhance the stability of 14-helix structures in water. Despite these advances, for the reasons mentioned above, quantitative experimental data are not common, and some of these rules are not universally accepted. For example, β^3 -residues that bear a side-chain branch point adjacent to the backbone, such as β^3 -Val, have been suggested^{24,25} to promote helix formation, but this idea has also been challenged.²⁶ For α/β -mixed peptides, even less systematic work has been carried out.²⁷

The goal of this research is to use molecular simulation to provide a deeper understanding of the sequence-structure–property relations of β - and α/β -peptides. We would like to be able to predict the structure of short β - and α/β -mixed peptides based only on sequence information, and, once the structure of individual peptide is known, we would like to be able to predict the macroscopic material properties of solutions of these peptides. Ultimately, we hope to use computational methods to aid the design of non-natural peptides with desired structure and function, e.g., as lung surfactant mimics.

There have been a few computational studies of β -peptides using ab initio QM calculations^{28–30} and molecular dynamics (MD) simulations using classical force fields.^{31–35} The ab initio calculations are valuable benchmarks but far too computationally intensive to map out the conformational and sequence space of interest. Classical MD simulations are computationally convenient but employ empirical force fields the parameters of which must be determined by comparison to either high level ab initio QM calculations or experiment. Most previous classical MD simulations employed force field

parameters developed for the α -amino acids which are probably not transferable to non-natural peptides. Therefore, although those studies provide valuable qualitative insight, they are not expected to be quantitatively reliable. Indeed, quantitatively validating the force field for non-natural peptides is not straightforward because experimental data are limited, and high level QM calculations are available only for the gas phase; most force fields are developed for peptides in solution.³⁶

A useful compromise, which we have decided to follow, is a hierarchical protocol in which both hybrid quantum mechanical/classical mechanical (QM/MM)^{37–40} and classical molecular mechanics (MM) simulations are used. In particular, we use QM/MM simulations as the reference to facilitate the development of a reliable MM force field for peptides that contain β -amino acids. In cases where larger-scale simulations (e.g., for the study of phase behaviors) are needed, the all-atom MM simulations can be used to parametrize an effective coarse-grained model. The reason to use a QM/MM model is that a QM model, in contrast to a MM model, can be directly calibrated against high-level ab initio calculations in the gas phase, which makes QM/MM simulations a uniquely meaningful reference.

Computational considerations force us to use an approximate QM model, and we use the Self-Consistent-Charge Density Functional Tight Binding (SCC-DFTB) method.⁴¹ Our choice is motivated by the computational efficiency of this method (comparable to widely used semiempirical methods such as AM1 and PM3) coupled with its reasonable accuracy, especially concerning the treatment of hydrogen bonding interactions.⁴² The SCC-DFTB method has been applied successfully to a range of problems involving biomolecules, such as conformational energies of natural peptides^{43–45} and catalysis in several enzymes;^{46–49} a recent review can be found in ref 50. Furthermore, the SCC-DFTB approach has been benchmarked for reaction energies, geometries, and vibrational frequencies for small molecules in comparison to the G2 approach.⁵¹ An empirical dispersion correction has also been developed,⁵² which was found to be crucial for predicting reliable nucleic acid base-stacking interactions⁵² and the relative stability of α and 3_{10} helices in proteins.⁵³

In this work, we set out to check the validity of SCC-DFTB as an appropriate QM method for β - and α/β -peptides with a diverse set of benchmark calculations, which include analyses in both the gas phase and in solution (methanol, water). In the gas phase, we test the method by comparison to high level ab initio calculations; in solution, we test the method by comparison to available experiments or QM/MM simulations with a high-level QM method. We find that the SCC-DFTB method is an acceptable alternative to high-level QM methods. The optimized structures are consistent with high level calculations, the barriers in the torsional potential energy are different by ~ 2 – 4 kcal/mol, and the dipole moments are within 15% of the high level calculations. Similar agreement is also found in solution.

The rest of the paper is organized as follows. The computational methods are described in section 2, the results

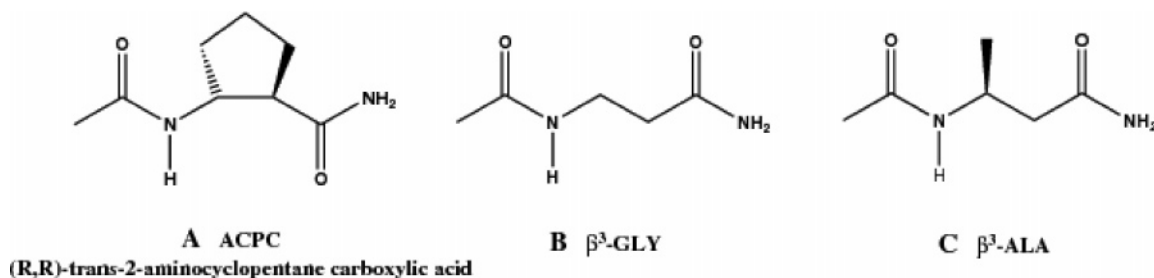


Figure 2. Dipeptide models studied here in the gas phase: **A**(ACPC), **B**(β^3 -GLY), and **C**(β^3 -ALA).

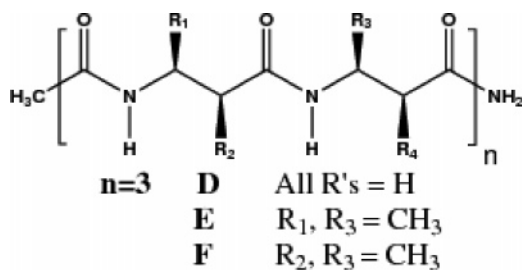


Figure 3. The heptapeptide models studied here in the gas phase: **D–F**.

are presented and discussed in section 3, and some conclusions are presented in section 4.

II. Methods

We test the performance of the SCC-DFTB method for peptides containing β -amino acids with benchmark calculations in the gas phase and in solution. The SCC-DFTB results for β -peptides in the gas phase are compared to high-level density functional theory (DFT)⁵⁴ and local MP2 calculations (LMP2).⁵⁵ The predictions of the method, implemented in a QM/MM framework,⁴⁶ for α/β -mixed peptides in water and methanol are compared to DFT/MM simulations and experiments, respectively. This section describes the nomenclature and methodology used in these calculations.

A. β -Peptides in the Gas Phase. 1. *One-Dimensional Adiabatic Mapping.* Since β -amino acids have an additional carbon atom along the backbone, compared to α -amino acids, we require one more torsional angle to describe the backbone conformation. In addition to the usual ϕ - ψ torsional angles, we require the angle (denoted θ) for the rotation along the C_α - C_β bond (Figure 1). As the first set of benchmark calculations, one-dimensional adiabatic mapping along these three degrees of freedom is carried out for two simple β dipeptides (**A** and **C** in Figure 2) at different levels of quantum mechanical theories; these include the standard SCC-DFTB,⁴¹ Hartree–Fock (HF) with the 6-31G* basis set,⁵⁶ B3LYP^{57–59} with the 6-31+G** basis set, and LMP2 with a larger 6-311G** basis set.⁶⁰ Diffuse functions have been included in the B3LYP calculations, because it has been shown that basis set superposition error (BSSE) for hydrogen-bonding interactions and conformational energies is reduced significantly with diffuse functions.^{61,62} The LMP2 rather than the canonical MP2 calculations are chosen because of the lower computational cost and reduced BSSE for the LMP2 approach;⁶³ previous calculations^{64,65} showed that LMP2 with triplet-zeta plus polarization basis sets perform accurately for conformational energies of natural peptides. The adiabatic

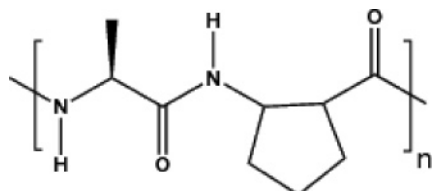
energy profile is calculated every 18° for each torsional angle, which spans the entire 360° range except that θ in compound **C** only goes from -180° to 0° due to stereochemical constraints.

2. *Conformational Studies of Model Peptides.* To sample the conformations of dipeptides more extensively, a systematic conformational search with 30° increments for the three characteristic dihedral angles (ϕ , ψ , θ) is performed for β^3 -GLY and β^3 -ALA (**B** and **C** in Figure 2). This generates a total of 1728 starting structures for each compound, which are then subject to minimization at different levels of theories: SCC-DFTB, B3LYP/6-31+G**, and LMP2/6-31G**; for SCC-DFTB, in addition to the standard parametrization, a recently proposed model for improving hydrogen-bonding interactions (referred to as “Hbond” parametrization in tables^{66,67}) and an empirical dispersion correction⁶⁸ have also been tested. Following the structural optimizations, single point energies are evaluated at the LMP2 level with a larger 6-311G** basis set;⁶⁰ dipole moments are also calculated with the standard SCC-DFTB and B3LYP/6-31+G**. The motivation of including the local MP2 results is that dispersion interactions, which have been found necessary for properly describing the relative stability of α -helix and 3_{10} -helix of natural peptides,⁶⁹ are poorly described in the popular DFT methods such as B3LYP.^{68–70} It is possible that dispersion could possibly make an important contribution to the relative stability of different conformers for β -peptides.

In addition to the simple dipeptides, several more complex heptapeptide molecules that have been analyzed in previous studies are also considered (Figure 3). The quantity of interest is the relative stability of the 14-helix and two different 10/12 mixed helical structures.² These conformations are fully optimized at both the SCC-DFTB and B3LYP/6-31G* levels. Similar to the dipeptide studies, single point energies of the optimized conformers are also calculated at the LMP2/6-311G** level, and the dipole moments are calculated with the standard SCC-DFTB and B3LYP/6-31+G**.

All HF and DFT calculations are carried out using the GAUSSIAN 03 package⁷¹ and LMP2 calculations using Jaguar.⁷² In all SCC-DFTB optimizations, the Adapted Basis Newton–Raphson (ABNR) approach in CHARMM⁷³ is used with a gradient tolerance of 0.0001 kcal/(mol·Å) for the dipeptides and 0.001 kcal/(mol·Å) for the heptapeptides.

B. α/β -Mixed Peptides in the Condensed Phase. Two sequences of α/β -mixed peptides are studied in the condensed phase, with the QM/MM approach, and these are depicted in Figure 4. The tetrapeptide (ACPC-A-ACPC-A, where



tetrapeptide $n=2$
octapeptide $n=4$

Figure 4. The α/β -mixed peptide models studied here in solution; the tetra- ($n = 2$) and octapeptides ($n = 4$), which are studied in water and methanol, respectively.

ACPC is (*R,R*)-*trans*-2-aminocyclopentane carboxylic acid) is studied in aqueous solution, and the initial structure is built from the crystal structure for a slightly different system¹⁰ where the α -amino acids are Aib. The octapeptide (ACPC-A-ACPC-A-ACPC-A-ACPC-A) is studied in methanol, and the starting configuration is built from the NOE derived structure reported by Schmitt et al.¹⁰ In both studies, the peptide is treated with QM, and the solvent (water or methanol) is treated with a MM model.

1. Tetrapeptide (ACPC-A-ACPC-A) in Explicit Water. Molecular dynamics simulations are carried out with SCC-DFTB/MM⁷⁴ at 300 K. The peptide is treated with the standard SCC-DFTB, while the solvent molecules are treated classically using the TIP3P model.⁷⁵ The peptide is solvated in an 18 Å water sphere, and the generalized solvent boundary potential (GSBP)^{76,77} is used for the boundary condition with a 2 Å water exclusion radius.⁷⁶ The system contains a total of 549 TIP3P water molecules and 65 peptide atoms. Two nanoseconds of equilibrium QM/MM simulation are carried out in which atoms in the spherical shell from 13 to 16 Å are treated with Langevin dynamics while the rest with Newtonian dynamics.⁷⁸ A time step of 1.0 fs is used, and all bonds to hydrogen atoms are constrained with SHAKE.⁷⁹

To evaluate the accuracy of the SCC-DFTB/MM hybrid potential, ~ 70 snapshots are taken from the 2 ns QM/MM simulation, and single point energy calculations are performed at the B3LYP/6-31+G**/MM level with the GAMESS package⁸⁰ interfaced with CHARMM.

2. Octapeptide (ACPC-A-ACPC-ACPC-A-ACPC-A) in Methanol. In the octapeptide-methanol simulation, the peptide is treated with the standard SCC-DFTB, while the methanol molecules are treated classically using the MEOH model in the CHARMM 22 all-atom force field;³⁶ calculations show that this methanol model describes the bulk property rather well with both periodic boundary and GSBP simulations (see the Supporting Information). The peptide is solvated in a 20 Å methanol sphere around its center of mass with a 2.5 Å of methanol exclusion shell associated with the GSBP setup. The system contains a total of 357 MEOH methanol molecules and 117 peptide atoms.

For the octapeptide, the issue of interest is the relative stability of the 14/15 and 11 helical structures. NOE data in methanol suggest that both conformers appear with likely similar stability.¹⁰ Motivated by this observation, the potential of mean force (PMF) associated with the conversion between

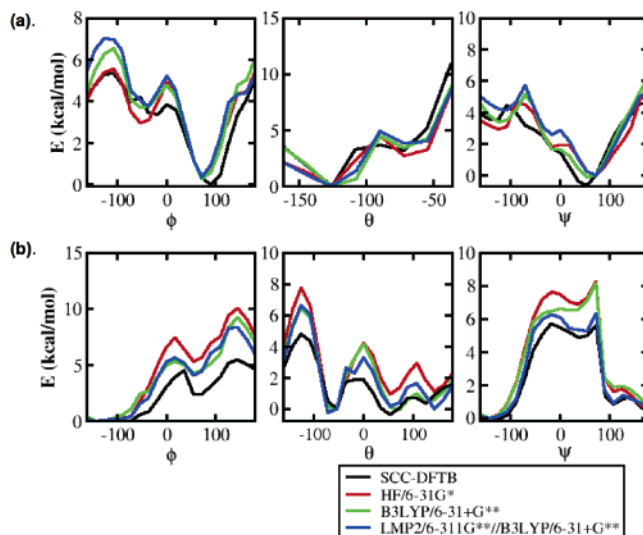


Figure 5. One-dimensional adiabatic energy map along the torsional angles ϕ , θ , and ψ for the (a) cyclic dipeptide **A** and (b) acyclic dipeptide **C**. Black: standard SCC-DFTB; red: HF/6-31G*; green: B3LYP/6-31+G**; blue: LMP2/6-311G**//B3LYP/6-31+G**.

the two helical forms is calculated using umbrella sampling.⁸¹ The reaction coordinate is chosen to be the end-to-end distance between the amide nitrogen in the first residue and the carbonyl carbon in the last residue, and the sampled range is between 10.6 and 18.0 Å. To prepare the initial structures for the umbrella samplings, a 14/15-helix is pulled toward the 11-helix in the gas phase with the reaction coordinate and intermediate structures collected every 0.2 Å, which are then used as the starting peptide conformation for each window. Totally 37 windows are sampled, each including 25 ps of equilibration and 75 ps of production calculations. The Weighted Histogram Analysis Method (WHAM)⁸² is used to analyze the data to obtain the PMF.

III. Results and Discussions

A. β -Peptides in the Gas Phase. 1. One-Dimensional Adiabatic Mapping of a Dipeptide. The SCC-DFTB reproduces the general features of the energy profiles along the ϕ , θ , and ψ angles for both the cyclic and acyclic dipeptides. Figure 5 (a),(b) compares the SCC-DFTB predictions to ab initio (HF, B3LYP, and LMP2) calculations for cyclic and acyclic peptides, respectively. The positions of local minima and maxima are the same in all theories, but the SCC-DFTB tends to underestimate the barriers between different minima, especially in the case of β^3 -Ala (which has a more flexible backbone than ACPC). The largest errors are found for the energy profile along ϕ and can be as large as ~ 3 – 4 kcal/mol compared to the LMP2 results. In the other cases, the errors are typically smaller than 2 kcal/mol. The implication is that the structure of β -peptides may be too flexible in SCC-DFTB and SCC-DFTB/MM simulations, although the theory is expected to be reliable for the structural propensity.

2. Conformational Studies of Dipeptides. For the acyclic dipeptides, β^3 -Gly (**B**) and β^3 -Ala (**C**), systematic conformational searches have identified six and twelve low-energy conformers, respectively (not including mirror image struc-

Table 1. Dihedral Angles, Dipole Moments, and Relative Energies for Optimized Conformers of Model **B** at Different QM Levels^a

conformer ^b	type ^c	dihedral angles (SCC ^d)			ΔE					dipole moment	
		ϕ	θ	ψ	SCC ^d	SCC-dispersion ^e	SCC-Hbond ^f	B3LYP ^g	LMP2 ^h	SCC ^d	B3LYP ^g
B1	C ₈	-109.6	56.8	35.4	0.0	0.0	0.0	0.0	0.0	3.6	5.3
B2	C ₆	117.5	58.1	125.6	0.3	0.5	0.5	-0.5	0.2	4.8	4.4
B3	C ₈	-83.3	124.8	-56.0	0.9	0.9	0.9	1.0	0.9	5.2	6.1
B4	C ₈	-46.4	-51.3	110.3	1.4	1.4	1.6	-0.5	0.6	3.1	4.4
B5		-80.5	167.3	-67.6	2.2	0.9	2.8	1.6	0.9	2.9	2.3
B6		-92.9	-171.7	-144.6	2.6	0.9	3.1	1.6	1.7	3.4	2.3

^a For structure, see Figure 2. Energies are in kcal/mol, dihedral angles in degrees, dipole moment in Debyes. ^b The mirror image conformers are energetically equivalent, and their dihedral angles differ only by sign. ^c C_x: hydrogen-bonded cycles with *x* atoms. ^d The standard parametrization of SCC-DFTB.⁴¹ ^e SCC-DFTB with the empirical dispersion interaction;⁶⁸ both **B5** and **B6** convert to **B3**. ^f SCC-DFTB with the hydrogen-bonding interaction correction.^{66,67} ^g Fully optimized at the B3LYP/6-31+G** level; **B4** converts to the mirror image conformer of **B2** and **B5** converts to **B6**. ^h LMP2/6-311G** single point energies at the LMP2/6-31G** optimized structures; **B5** converts to **B3**.

Table 2. Dihedral Angles, Dipole Moments, and Relative Energies for Optimized Conformers of Model **C** at Different QM Levels^a

conformer	type ^b	dihedral angles (SCC ^c)			ΔE					dipole moment	
		ϕ	θ	ψ	SCC ^c	SCC-dispersion ^d	SCC-Hbond ^e	B3LYP ^f	LMP2 ^g	SCC ^c	B3LYP ^f
C1	C ₈	-109.8	53.0	39.8	0.0	0.0	0.0	0.0	0.0	3.6	5.1
C2	C ₆	-106.7	-59.7	-128.6	0.2	0.1	0.4	-0.2	-0.5	4.6	5.4
C3	C ₈	-78.2	129.3	-61.4	0.7	0.5	0.6	0.4	0.1	5.1	5.8
C4	C ₈	-49.6	-48.6	111.0	1.1	1.1	1.3	1.7	0.4	2.8	2.6
C5	C ₆	-167.1	54.3	112.9	1.3	1.3	1.4	1.6	0.9	4.3	4.6
C6	C ₆	57.8	58.3	178.4	1.3	1.4	1.5	1.3	2.3	2.0	0.9
C7	C ₈	50.2	48.4	-111.2	1.4	1.6	1.6	1.7	0.4	3.0	3.0
C8		-64.0	171.6	-175.9	2.6	0.5	2.9	0.4	0.1	3.5	5.8
C9	C ₈	63.7	-126.5	83.4	2.8	2.6	2.8	4.0	3.3	5.0	6.0
C10		60.0	169.3	160.2	2.8	3.0	3.3	3.0	2.6	1.7	1.3
C11	H _{10/12}	71.5	-37.0	-94.4	3.3	0.5	3.8	7.8	5.7	2.5	3.5
C12		-173.9	-72.5	42.3	5.1	5.0	5.4	6.3	4.6	3.5	3.0

^a For structure, see Figure 2. Energies are in kcal/mol, dihedral angles in degrees, dipole moment in Debyes. ^b C_x hydrogen-bonded cycles with *x* atoms; H_{xy} mixed helix pattern. ^c The standard parametrization of SCC-DFTB.⁴¹ ^d SCC-DFTB with the empirical dispersion interaction;⁶⁸ **C8** converts to **C3** and **C11** to **C5**. ^e SCC-DFTB with the hydrogen-bonding interaction correction.^{66,67} ^f Fully optimized B3LYP/6-31+G**, **C8** converts to **C3**. ^g LMP2/6-311G** single point energies at the LMP2/6-31G** optimized structures; **C8** converts to **C3**.

Table 3. rms Differences in Optimized Dihedral Angles in Various β -Peptides Compared to the Standard SCC-DFTB Results^a

model	SCC-dispersion ^c	SCC-Hbond ^d	B3LYP	LMP2 ^g
dipeptide	8.8	6.6	13.4 ^e	9.9
heptapeptide ^b	6.7	0.7	9.0 ^f	

^a All dihedral angles rms differences are in degrees. ^b Only the 10/12 mixed helices are included; for the 14-helix, see Figures 6–8. ^c SCC-DFTB with the empirical dispersion interaction.⁶⁸ ^d SCC-DFTB with the hydrogen-bonding interaction correction.^{66,67} ^e B3LYP/6-31+G**. ^f B3LYP/6-31G*. ^g LMP2/6-31G**.

tures for **B**); the structure, energy, and dipole moment of these conformers are summarized in Tables 1 and 2.

For both β -dipeptides, SCC-DFTB is able to reproduce most of the structures predicted by previous ab initio calculations^{28,30} and has identified new locally stable structures not reported before. For **B**, conformers **B1**, **B2**, **B3**, and **B6** are consistent with those found in previous HF/6-31G** calculations,²⁸ and **B4** and **B5** have not been reported before. Instead, Wu et al.²⁸ reported two other conformers, which are not local minima and convert to the lower energy conformer **B1** at both SCC-DFTB and B3LYP/6-31+G** levels. For **C**, SCC-DFTB reproduces 8 conformers (**C1**, **C2**, **C4**, **C5**, **C6**, **C7**, **C8**, **C10**) out of the ten that have been

Table 4. rms Differences in Relative Energetics for Various β -Peptides Compared to LMP2/6-311G** Single Point Energies^a

model	SCC ^d	SCC-dispersion ^e	SCC-Hbond ^f	B3LYP ^g
dipeptide ^b	1.2	1.3	1.3	0.9
heptapeptide ^b	3.3	4.5	3.6	2.2
heptapeptide ^c	2.0			2.6

^a All energies are in kcal/mol. ^b With the standard SCC-DFTB optimized structures; for the heptapeptides, only the 10/12 mixed helices and the fully optimized 14-helices are included. ^c With B3LYP/6-31G* optimized structures. ^d The standard parametrization of SCC-DFTB.⁴¹ ^e SCC-DFTB with the empirical dispersion interaction.⁶⁸ ^f SCC-DFTB with the hydrogen-bonding interaction correction.^{66,67} ^g B3LYP/6-31+G**.

reported by Wu and Wang.²⁸ Other conformers (**C3**, **C9**, **C11**, **C12**) were found in calculations of Möhle et al.,³⁰ although their model is N-methylated. One of the two conformers found by Wu et al.²⁸ converts to **C1** at both SCC-DFTB and B3LYP/6-31+G** levels. The other one converts to **C5** with SCC-DFTB; it is a local minimum at the B3LYP/6-31+G** level although it is much higher in energy than most of the conformers in Table 2 (except for **C11** and **C12**). Overall, the optimized dihedral angles in these low-energy conformers at the SCC-DFTB level are fairly close to B3LYP

Table 5. Average Dihedral Angles and Dipole Moments of Various Conformers of Models **D–F** at Different QM Levels^a

model	structure	dipole moment		av dihedral angles (SCC ^b)					
		SCC ^b	B3LYP ^c	ϕ_1	θ_1	ψ_1	ϕ_2	θ_2	ψ_2
D	10/12/10/12/10	4.7	4.9	88.7	60.2	-107.1	-100.9	59.0	74.2
	12/10/12/10/12	1.2	1.8	-104.9	57.6	86.8	96.7	59.3	-125.8
	14-helix ^d	25.1	28.5	-134.3	60.0	-139.9			
	14-helix ^e	26.2	29.8	-141.9	61.4	-137.4			
	14-helix ^f	7.9	9.0						
E	10/12/10/12/10	3.5	3.7	71.7	62.9	-106.1	-103.7	60.4	85.0
	12/10/12/10/12	1.9	2.0	-104.1	59.6	99.1	75.4	64.3	-131.6
	14-helix ^d	25.0	27.9	-134.3	60.0	-139.9			
	14-helix ^e	25.8	28.8	-144.4	59.9	-135.3			
	14-helix ^f	16.5	18.2						
F	10/12/10/12/10	5.1	5.5	89.2	59.6	-106.4	-101.0	58.2	74.2
	12/10/12/10/12	2.1	2.8	110.0	-52.9	-88.9	-104.1	-56.7	113.6
	14-helix ^d	25.0	28.4	-134.3	60.0	-139.9			
	14-helix ^e	25.8	29.1	-144.1	59.2	-134.0			
	14-helix ^f	16.5	18.5						

^a Dihedral angles in degrees and dipole moments in Debyes. ^b The standard parametrization of SCC-DFTB.⁴¹ ^c B3LYP/6-31+G** at standard SCC-DFTB optimized structure. ^d Optimized with the backbone dihedral angles constrained to be ideal 14-helix values. ^e Optimized with the backbone dihedral angles constrained to be HF/6-31G* optimized values of Wu et al.²⁹ ^f Fully optimized without any constraints; for these structures, the average dihedral angles are less meaningful due to the substantial deviation from the ideal helical form and, therefore, not given.

Table 6. Relative Energies of Various Conformers of Models **D–F** at Different QM Levels^a

model	structure	SCC ^b	SCC-dispersion ^c	SCC-Hbond ^d	B3LYP ^e	LMP2 ^f
D	10/12/10/12/10	0.0 (0.0)	0.0	0.0	0.0 (0.0)	0.0 (0.0)
	12/10/12/10/12	0.8 (0.1)	1.0	1.1	-3.9 (-1.3)	-3.5 (-1.4)
	14-helix ^g	32.5 (19.4)	35.8	34.8	21.0 (22.6)	21.6 (22.5)
	14-helix ^h	25.4 (19.4)	27.1	27.2	15.5 (22.6)	15.7 (22.5)
	14-helix ⁱ	9.4 (9.9)	7.4	10.0	12.5 (12.2)	12.4 (10.9)
E	10/12/10/12/10	0.0 (0.0)	0.0	0.0	0.0 (0.0)	0.0 (0.0)
	12/10/12/10/12	1.6 (1.4)	1.9	1.8	0.3 (-1.3)	-1.8 (0.04)
	14-helix ^g	26.2 (13.9)	28.9	27.9	11.4 (12.2)	5.4 (9.6)
	14-helix ^h	20.0 (13.9)	22.1	21.7	5.8 (12.2)	0.6 (9.6)
	14-helix ⁱ	13.0 (11.9)	16.1	14.0	9.2 (11.0)	5.2 (9.7)
F	10/12/10/12/10	0.0 (0.0)	0.0	0.0	0.0 (0.0)	0.0 (0.0)
	12/10/12/10/12	2.5 (3.8)	1.4	2.2	6.9 (8.5)	2.8 (4.0)
	14-helix ^g	31.9 (20.5)	34.1	33.8	21.3 (22.9)	17.4 (18.7)
	14-helix ^h	25.1 (20.5)	26.7	27.0	15.1 (22.9)	11.7 (18.7)
	14-helix ⁱ	17.2 (16.8)	18.7	18.3	20.5 (19.5)	17.7 (13.9)

^a Energies are in kcal/mol; values in parentheses are computed using the B3LYP/6-31G* optimized structures, and others are computed using the standard SCC-DFTB optimized structures. ^b The standard parametrization of SCC-DFTB.⁴¹ ^c SCC-DFTB with the empirical dispersion interaction.⁶⁸ ^d SCC-DFTB with the hydrogen-bonding interaction correction.^{66,67} ^e B3LYP/6-31+G** energies. ^f LMP2/6-311G** energies. ^g Optimized with the backbone dihedral angles constrained to be ideal 14-helix values. ^h Optimized with the backbone dihedral angles constrained to be HF/6-31G* optimized values of Wu et al.²⁹ ⁱ Fully optimized without any constraints.

and LMP2 values; the rms errors of 10 degrees (Table 3) are expected considering the rather flat energy profiles along these torsional angles. The hydrogen-bonding correction for SCC-DFTB^{66,67} changes the optimized structure only slightly with a rms in the dihedral angles of 6.6°. Dispersion also has only little effect on the geometries, based on results from both SCC-DFTB calculations that include the empirical dispersion⁶⁸ and LMP2 calculations.

For both **B** and **C**, the central torsional angle, θ , prefers the *gauche* conformation, and conformers with 8- and 6-membered hydrogen-bonding cycles are most stable. At all SCC-DFTB levels (i.e., regardless of hydrogen-bonding and dispersion effects), the global minimum is a C₈ conformer (**B1**, **C1**), although a C₆ conformer (**B2**, **C2**) is only slightly higher in energy by 0.2–0.6 kcal/mol. B3LYP and

LMP2 calculations tend to favor the C₆ conformer, although the energy preference is again very small, on the order of 0.2–0.4 kcal/mol (see Tables 1 and 2).

Overall, there is very good agreement between SCC-DFTB, B3LYP, and LMP2 relative energies. Both hydrogen-bonding correction and dispersion have, in general, little effect on the SCC-DFTB results; in a few cases, however, including the dispersion causes certain conformers to disappear as local minima, often in agreement with the LMP2 result (see footnotes of Tables 1 and 2). The rms errors of the various SCC-DFTB models relative to the LMP2 results are 1.2–1.3 kcal/mol, only slightly larger than the value of 0.9 kcal/mol for B3LYP (Table 4). The major exception is **C11**, which adopts the 10/12 helical structure; it is 7.6 kcal/mol and 10 kcal/mol higher than the global minimum at the

LMP2 and B3LYP level, respectively, but is only ~ 3 kcal/mol higher than the global minimum at all SCC-DFTB levels.

For dipole moments, the SCC-DFTB results deviate from the B3LYP/6-31+G** values by $\sim 15\%$, and the rms error is 0.9 Debye.

3. Conformational Studies of Heptapeptides. For the three heptapeptides, **D–F**, we focus on the three typical helical forms that have been studied in the previous work of Wu and Wang:²⁹ the 10/12 mixed helices (10/12/10/12/10, 12/10/12/10/12) and the 14-helix. For the 10/12 mixed helices, SCC-DFTB optimizations give structures (Table 5) in close agreement with both previous HF/6-31G*²⁹ and current B3LYP calculations, regardless of the hydrogen-bonding correction and inclusion of dispersion. For example, the rms difference in the dihedral angles is about 9 degrees relative to B3LYP/6-31G* calculations (Table 3), and the rmsd of backbone atoms is normally less than 0.2 Å. In addition, single point energies at both the B3LYP and LMP2 levels at the fully optimized SCC-DFTB structures are generally very similar to those at the B3LYP optimized structures (Table 6), indicating that the SCC-DFTB geometries are close to the B3LYP ones. As a result, as seen in Table 5, the conformational dependences in the dipole moment for all three heptapeptides are well reproduced at the SCC-DFTB level with a systematic underestimate of about 10%. The dipole moments of 14-helices are much greater than those of the 10/12 mixed helices, which makes the former better stabilized in polar solvent as found experimentally.^{4,6,7}

Regarding energetics, SCC-DFTB systematically predicts the 10/12/10/12/10 helix to be lower in energy, although the preference over the 12/10/12/10/12 structure is smaller than 2 kcal/mol. At the B3LYP and LMP2 levels, the 12/10/12/10/12 form is lower in energy for **D** and **E**, although the preference is also very small and around 1 kcal/mol. For **F**, the LMP2 result is in fact closer to the SCC-DFTB results compared to the B3LYP value.

The situation for the 14 helical form is more complex. Figures 6–8 compare the structures obtained from various methods. For all three heptapeptides, the ideal 14-helix is not stable at the standard SCC-DFTB level and either partially converts to a 10- or 12-membered-ring hydrogen-bonding pattern or adopts bifurcated hydrogen-bonding at the two termini. Indeed, partial SCC-DFTB optimizations with the backbone dihedrals constrained to either the values in an ideal 14-helix or those in the reported HF/6-31G* structures of Wu and Wang²⁹ give structures of substantially higher energy than the fully optimized SCC-DFTB structures (Table 4). At the B3LYP/6-31G* level, the HF structures of Wu and Wang²⁹ do exist as stable local minima, although their energies are also higher than the structures optimized using the SCC-DFTB result as the starting configuration; this trend is maintained with LMP2/6-311G** single point energy calculations at the B3LYP geometries (Table 6). The B3LYP and SCC-DFTB optimized structures are generally very similar with backbone rmsd less than 0.4 Å, although B3LYP and LMP2 single point energies at the fully optimized SCC-DFTB structures tend to be *higher* than those at the partially optimized SCC-DFTB structures with backbone dihedrals constrained to the HF/6-31G* values, in

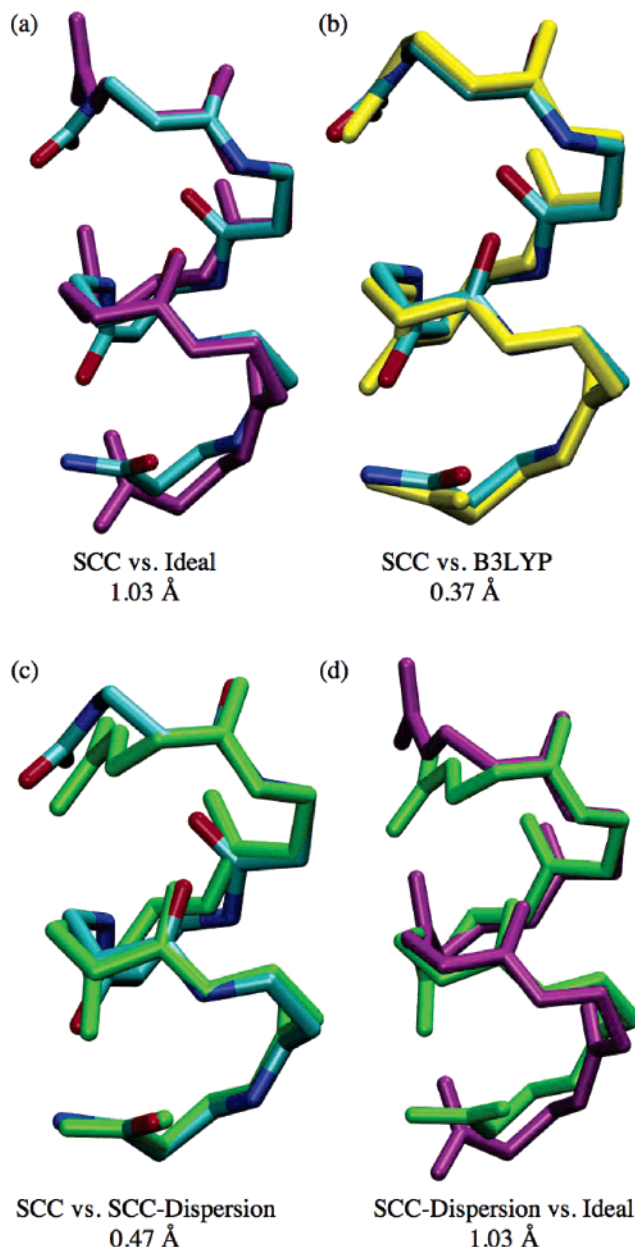


Figure 6. Comparison of different structures of model **D** from different calculations (the number below each superposition is the backbone rmsd): (a) SCC-DFTB optimized structure (CPK color) vs ideal 14-helix (purple); (b) SCC-DFTB optimized structure (CPK color) vs B3LYP/6-31G* optimized structure (yellow); (c) SCC-DFTB optimized structure (CPK color) vs SCC-DFTB+dispersion optimized structure (green); and (d) SCC-DFTB+dispersion optimized structure (green) vs ideal 14-helix (purple).

contrast to both SCC-DFTB energies and high-level single point energies at the B3LYP optimized structures. This subtlety suggests that there are still non-negligible errors in the SCC-DFTB geometries, and caution has to be exercised when attempting to improve the energetics by performing high-level energy calculations at the SCC-DFTB structures.

In general, the hydrogen-bonding correction to SCC-DFTB does not change the geometry or energetics of the heptapeptides studied here. By contrast, including dispersion gives more variations. Although dispersion does not affect the geometries of the 10/12 mixed helices, notable effects on

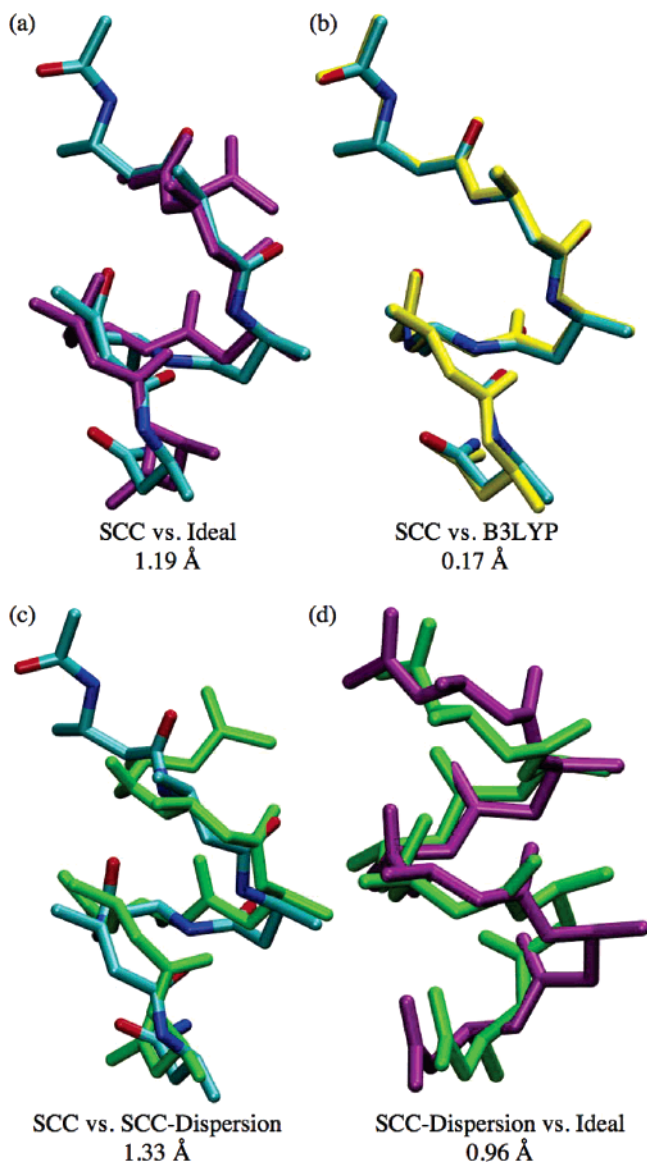


Figure 7. Comparison of different structures of model **E** from different calculations (the number below each superposition is the backbone rmsd): (a) SCC-DFTB optimized structure (CPK color) vs ideal 14-helix (purple); (b) SCC-DFTB optimized structure (CPK color) vs B3LYP/6-31G* optimized structure (yellow); (c) SCC-DFTB optimized structure (CPK color) vs SCC-DFTB+dispersion optimized structure (green); and (d) SCC-DFTB+dispersion optimized structure (green) vs ideal 14-helix (purple).

the structure of the 14-helix is observed except for **D**. For the β -substituted model **E**, with dispersion included in the SCC-DFTB, 14-membered-ring hydrogen bonds are formed except at the C terminus; the average dihedral angles for the 4 middle residues are $\phi = -160.5^\circ$, $\theta = 61.0^\circ$, and $\psi = -127.8^\circ$, very close to the values in an ideal 14-helix.² Without dispersion, the two hydrogen bonds close to the C-terminus are completely lost, which leads to a structure with a rmsd value larger than 1.1 Å relative to the ideal 14-helix (Figure 7). For the mixed-substituted model **F**, by contrast, dispersion does not stabilize the ideal 14-helical structure (Figure 8). On average (Table 6), including dispersion *raises* the energy of the 14-helix relative to the 10/12

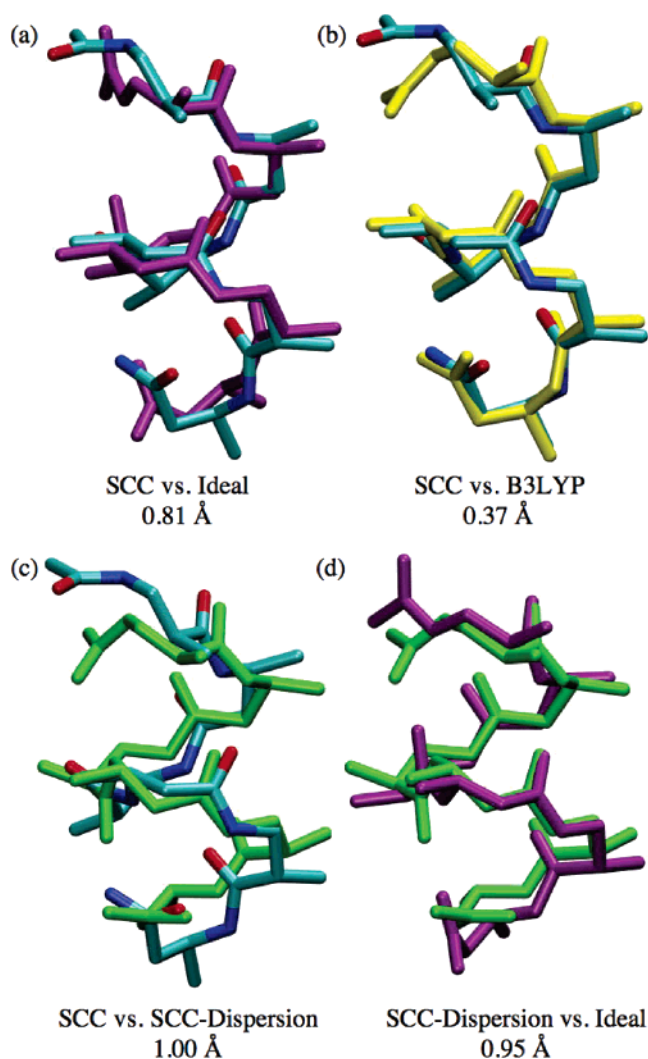


Figure 8. Comparison of different structures of model **F** from different calculations (the number below each superposition is the backbone rmsd): (a) SCC-DFTB optimized structure (CPK color) vs ideal 14-helix (purple); (b) SCC-DFTB optimized structure (CPK color) vs B3LYP/6-31G* optimized structure (yellow); (c) SCC-DFTB optimized structure (CPK color) vs SCC-DFTB+dispersion optimized structure (green); and (d) SCC-DFTB+dispersion optimized structure (green) vs ideal 14-helix (purple).

mixed helices by a small amount, which is somewhat unexpected based on the previous findings for natural peptides that dispersion stabilizes the wider α -helix than the thinner 3_{10} -helix.⁶⁹ This is probably because the larger number of atoms in the β -amino acids causes dispersion to saturate more quickly as a function of the number of residues compared to natural peptides, which leads to a smaller effect on the relative energies of different helical forms.

Overall, the findings from the heptapeptide calculations are similar to that for the dipeptides, which show that the standard SCC-DFTB parametrization gives rather reliable structures and relative energetics (rms error on the order of 2–3 kcal/mol) for various conformers as compared to B3LYP and LMP2 calculations. Therefore, it seems that the standard SCC-DFTB, even without the hydrogen-bonding correction^{66,67} and dispersion interactions,⁶⁸ can describe the

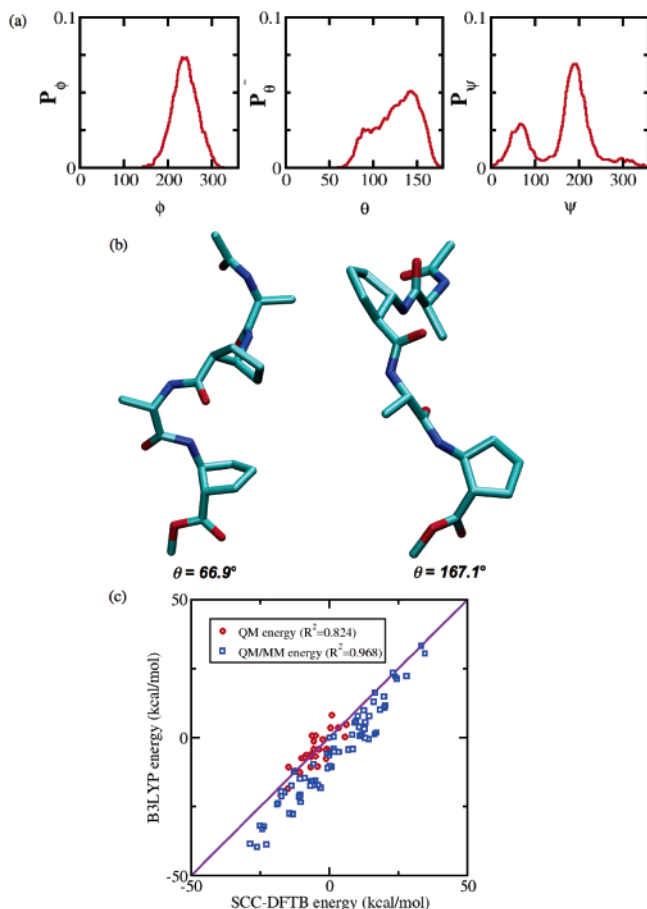


Figure 9. SCC-DFTB/MM simulations for the tetrapeptide (ACPC-A-ACPC-A) in water solution: (a) dihedral angle ϕ , θ , and ψ distributions of the central β -residue (ACPC); (b) two representative snapshots with large and small θ angles; and (c) the correlation between SCC-DFTB/MM and B3LYP/6-31+G**/MM energies for ~ 70 snapshots taken from 2 ns SCC-DFTB/MM simulation (blue squares); gas-phase SCC-DFTB and B3LYP/6-31+G** energy correlations at those structures (red circles) are also shown for comparison.

intrinsic structural-energy relation of the helical forms of β -peptides to a satisfactory degree.

B. α/β -Mixed Peptides in the Condensed Phase. 1. Tetrapeptide (ACPC-A-ACPC-A) in Explicit Water. The SCC-DFTB/MM simulations predict that the tetrapeptide is rather flexible, despite the five-membered ring in the β -amino acids: A wide range of θ values are sampled within a time duration of 2 ns. The conformational properties of the tetrapeptide and comparison with ab initio calculations are presented in Figure 9. The probability distribution function for θ is weakly bimodal with two peaks (Figure 9a). Characteristic structures corresponding to the values of θ at the peak in the distribution are rather different (Figure 9b). The ϕ and ψ predominately sample the region that corresponds to the β -sheet structure of α -peptides in the Ramachandran plot.⁸³ Evidently, different ϕ - ψ ranges characterize secondary structures in non-natural peptides.

To further validate the SCC-DFTB/MM model, a number of randomly chosen snapshots (~ 70) are taken from the SCC-DFTB/MM trajectory, and single point energies are calculated at the B3LYP/6-31+G**/MM level. As seen in Figure

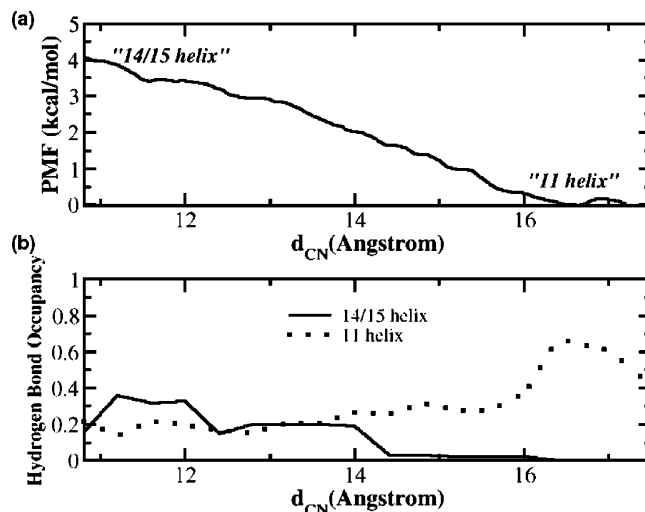


Figure 10. SCC-DFTB/MM simulations for octapeptide (ACPC-A-ACPC-A-ACPC-A-ACPC-A) in methanol solution: (a) PMF for the conversion between the 14/15-helix and 11-helix of octapeptide and (b) hydrogen-bonding occupancy analysis (solid line: 14/15-helix; dashed line: 11-helix).

9c, the SCC-DFTB/MM energies show a good correlation with the B3LYP/MM results with a correlation coefficient of 0.968, although SCC-DFTB/MM systematically underestimates the relative energies. Interestingly, the correlation is weaker when only gas-phase energies at the same structures are considered (i.e., without the interaction with the MM water molecules), which has a correlation coefficient of 0.824. Apparently, there is some degree of cancellation between the gas-phase errors in SCC-DFTB and the errors associated with the QM/MM interactions in the SCC-DFTB/MM framework; the SCC-DFTB/MM interaction is treated with the same Mulliken approximation as for the electrostatic interactions between SCC-DFTB atoms, while the QM/MM interactions are calculated with exact one-electron integrals in the B3LYP/MM calculations.⁷⁴ Considering the wide range of conformations sampled in the SCC-DFTB/MM simulations, the general agreement with B3LYP/MM energies is encouraging.

2. Octapeptide (ACPC-A-ACPC-ACPC-A-ACPC-A) in Methanol. For the octapeptide in methanol, the key question is the relative stability of the 14/15 and 11 helices. Experiments¹⁰ suggest that both of these structures are present with detectable populations, although a quantitative characterization was not reported. Figure 10(a) depicts the potential of mean force for the conversion between the 14/15-helix and the 11-helix. The SCC-DFTB/MM simulations (Figure 10a) show that the 11-helix is more stable than the 14/15-helix by about 3 kcal/mol. If we consider that the gas-phase calculations discussed above suggest that SCC-DFTB tends to underestimate the stability of shorter and wider helices (e.g., the stability of the 14-helix is underestimated compared to 10/12 mixed helices for heptapeptides in Table 4), the PMF result implies that the two helical forms are even closer in free energy (14/15-helix is shorter and wider than the 11-helix) than 3 kcal/mol, which is qualitatively consistent with the experimental NOE data.¹⁰

The order parameter used in the PMF calculations, d_{CN} , appears to be a valid one since it can effectively distinguish

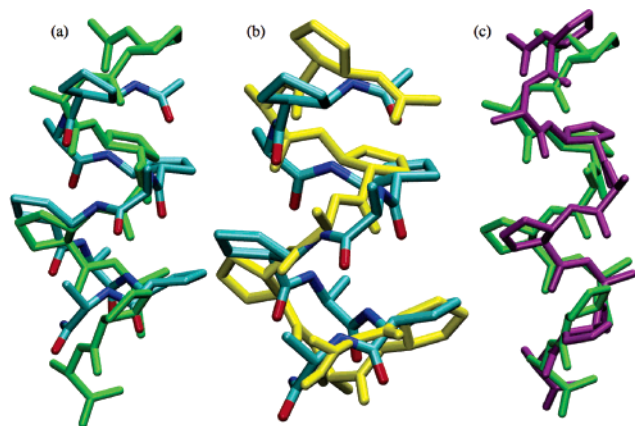


Figure 11. Structures of 14/15 and 11 helices of octapeptide (ACPC-A-ACPC-A-ACPC-A-ACPC-A) in methanol: (a) superposition of the ideal 14/15-helix (CPK color) and 11-helix (green); (b) superposition of the ideal 14/15-helix (CPK color) and the average structure of the window $d_{\text{CN}} = 11.4 \text{ \AA}$; and (c) superposition of the ideal 11-helix (green) and the average structure of the window $d_{\text{CN}} = 17.0 \text{ \AA}$ (purple).

the two helical forms based on the average structures from different windows (Figure 11). There is an interesting difference in the stability of the backbone hydrogen-bonding interactions between the two helical forms (Figure 10b). Although the occupancy of the backbone hydrogen bonds is rather high (~ 0.7) for the 11-helix, the value is substantially lower (< 0.4) for the 14/15-helix, implying a very dynamical structure for the latter.

IV. Conclusions and Outlook

Carrying out molecular simulations of non-natural peptides is both exciting and challenging. On one hand, the lack of extensive amount of experimental data means that molecular simulations can play a major role in understanding the structural and dynamical properties of these novel materials. On the other hand, there is only a limited amount of data available for establishing a robust molecular model and simulation protocol. Our long-term goal is to use QM/MM models to guide the development of classical models for β - and α/β -peptides at both the all-atom and coarse-grained levels, such that the relationship between the sequence of these peptides and their structural as well as material properties can be analyzed. This is motivated by the fact that the reliability of a QM model can be meaningfully tested in both the gas phase and condensed phase by comparing to high-level QM calculations and available experimental data, respectively.

In this paper, the reliability of an approximate density functional theory, SCC-DFTB, as the QM model for β - and α/β -mixed peptides has been tested by both gas-phase and condensed-phase calculations for several rather different systems. In the gas phase, both cyclic and acyclic dipeptides and three acyclic heptapeptides have been studied at several QM levels, including the standard SCC-DFTB, B3LYP, and LMP2; the effect of two recent enhancements to SCC-DFTB, which deal with hydrogen-bonding interactions and dispersion interactions, has also been tested. Overall, the standard SCC-DFTB approach has been shown to reproduce the

B3LYP structures with an rms error in the key dihedral angles of less than 14 degrees. Importantly, SCC-DFTB is able to capture the lowest-energy conformers for all dipeptides and heptapeptides studied here, although several local minima of higher energy, such as the ideal 14-helical form for the heptapeptides, are missed at the SCC-DFTB level. The relative energies of different conformers are also well described in general, with typical rms errors of 2–3 kcal/mol relative to LMP2 single points at the B3LYP structures. The dipole moments are reproduced with a systematic underestimate of approximately 15%. The effect of including the hydrogen-bonding correction or empirical dispersion in the SCC-DFTB calculations is generally small, although including dispersion in several cases leads to rather different structures; the effect of those corrections is expected to be more significant when comparing folded (compact) structures and unfolded structures, as found in the folding simulation of β -peptides (Zhu, X. et al., work in progress). In addition to the gas-phase studies, SCC-DFTB/MM simulations have been carried out for a tetra- α/β -mixed peptide in water and for an octamer in methanol. For the tetrameric system, the SCC-DFTB/MM energies are well correlated with B3LYP/6-31+G**/MM single point energies for a wide range of structures sampled in 2 ns of SCC-DFTB/MM molecular dynamics trajectory. For the octamer, PMF calculations indicate that the 14/15 and 11 helices are within 3 kcal/mol in free energy with the latter being more stable, which is in qualitative agreement with available NOE data.

With all these results taken together, it is established that although SCC-DFTB has non-negligible errors in structures and energetics compared to high-level DFT and ab initio methods, it is expected to capture the most important configurations for β - and α/β -mixed peptides. Considering their computational efficiency, we conclude that SCC-DFTB and SCC-DFTB/MM are effective methods for describing the structure-energy properties of these non-natural peptides in the gas phase and solution (water, methanol) phase, respectively. In particular, SCC-DFTB/MM simulations can be used as the unique reference for developing useful MM models at multiple resolutions. Such studies are in progress.

Acknowledgment. This work is supported from the National Science Foundation (CRC-CHE-0404704). Discussions with Prof. S. Gellman, Dr. M. Schmitt, and Mr. S. H. Choi are greatly appreciated. Q.C. also acknowledges an Alfred P. Sloan Research Fellowship. Computational resources from the National Center for Supercomputing Applications at the University of Illinois are greatly appreciated.

Supporting Information Available: Calculations for the bulk properties of methanol using two different classical force field models, comparison to experimental results, and the complete ref 36. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Gellman, S. H. *Acc. Chem. Res.* **1998**, *31*, 173–180.
- (2) Cheng, R. P.; Gellman, S. H.; DeGrado, W. F. *Chem. Rev.* **2001**, *101*, 3219–3232.

- (3) DeGrado, W. F.; Schneider, J. P.; Hamuro, Y. *J. Peptide Res.* **1999**, *54*, 206–217.
- (4) Appella, D. H.; Christianson, L. A.; Karle, I. L.; Powell, D. R.; Gellman, S. H. *J. Am. Chem. Soc.* **1996**, *118*, 13071–13072.
- (5) Appella, D. H.; Christianson, L. A.; Klein, D. A.; Richards, M. R.; Powell, D. R.; Gellman, S. H. *J. Am. Chem. Soc.* **1999**, *121*, 7574–7581.
- (6) Appella, D. H.; Christianson, L. A.; Karle, I. L.; Powell, D. R.; Gellman, S. H. *J. Am. Chem. Soc.* **1999**, *121*, 6206–6212.
- (7) Seebach, D.; Overhand, M.; Kühnle, F. N. M.; Martinoni, B.; Oberer, L.; Hommel, U.; Widmer, H. *Helv. Chim. Acta* **1996**, *79*, 913–941.
- (8) Seebach, D.; Gademann, K.; Schreiber, J. V.; Matthews, J. L.; Hintermann, T.; Jaun, B.; Oberer, L.; Hommel, U.; Widmer, H. *Helv. Chim. Acta* **1997**, *80*, 2033–2038.
- (9) Hayen, A.; Schmitt, M. A.; Ngassa, F. N.; Thomasson, K. A.; Gellman, S. H. *Angew. Chem., Int. Ed.* **2004**, *43*, 505–510.
- (10) Schmitt, M. A.; Choi, S. H.; Guzei, I. A.; Gellman, S. H. *J. Am. Chem. Soc.* **2005**, *127*, 13130–13131.
- (11) Porter, E. A.; Wang, X. F.; Lee, H. S.; Weisblum, B.; Gellman, S. H. *Nature* **2000**, *404*, 565–565.
- (12) Porter, E. A.; Weisblum, B.; Gellman, S. H. *J. Am. Chem. Soc.* **2002**, *124*, 7324–7330.
- (13) Raguse, T. L.; Porter, E. A.; Weisblum, B.; Gellman, S. H. *J. Am. Chem. Soc.* **2002**, *124*, 12774–12785.
- (14) Hamuro, Y.; Schneider, J. P.; DeGrado, W. F. *J. Am. Chem. Soc.* **1999**, *121*, 12200–12201.
- (15) Liu, D. H.; DeGrado, W. F. *J. Am. Chem. Soc.* **2001**, *123*, 7553–7559.
- (16) Schmitt, M. A.; Weisblum, B.; Gellman, S. H. *J. Am. Chem. Soc.* **2004**, *126*, 6848–6849.
- (17) Eldred, S. E.; Pancost, M. R.; Otte, K. M.; Rozema, D.; Stahl, S. S.; Gellman, S. H. *Bioconjugate Chem.* **2005**, *16*, 694–699.
- (18) Wu, C. W.; Seuryneck, S. L.; Lee, K. Y. C.; Barron, A. E. *Chem. Biol.* **2003**, *11*, 1057–1063.
- (19) Glättli, A.; Daura, X.; Seebach, D.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **2002**, *124*, 12972–12978.
- (20) Seebach, D.; Ciceri, P. E.; Overhand, M.; Jaun, B.; Rigo, D.; Oberer, L.; Hommel, U.; Amstutz, R.; Widmer, H. *Helv. Chim. Acta* **1996**, *79*, 2043–2066.
- (21) Arvidsson, P. I.; Rueping, M.; Seebach, D. *Chem. Commun.* **2001**, 649–650.
- (22) Cheng, R. P.; DeGrado, W. F. *J. Am. Chem. Soc.* **2001**, *123*, 5162–5163.
- (23) Cheng, R. P.; DeGrado, W. F. *J. Am. Chem. Soc.* **2002**, *124*, 11564–11565.
- (24) Hart, S. A.; Bahadour, A. B. F.; Matthews, E. E.; Qiu, X. Y. J.; Schepartz, A. *J. Am. Chem. Soc.* **2003**, *125*, 4022–4023.
- (25) Raguse, T. L.; Lai, J. R.; Gellman, S. H. *Helv. Chim. Acta* **2002**, *85*, 4154–4164.
- (26) Glättli, A.; Seebach, D.; van Gunsteren, W. F. *Helv. Chim. Acta* **2004**, *87*, 2487–2506.
- (27) Baldauf, C.; Günther, R.; Hofmann, H. J. *Biopolymers* **2006**, *84*, 408–413.
- (28) Wu, Y. D.; Wang, D. P. *J. Am. Chem. Soc.* **1998**, *120*, 13485–13493.
- (29) Wu, Y. D.; Wang, D. P. *J. Am. Chem. Soc.* **1999**, *121*, 9352–9362.
- (30) Möhle, K.; Günther, R.; Thormann, M.; Sewald, N.; Hofmann, H. J. *Biopolymers* **1999**, *50*, 167–184.
- (31) Appella, D. H.; Christianson, L. A.; Klein, D. A.; Powell, D. R.; Huang, X. L.; Barchi, J. J.; Gellman, S. H. *Nature* **1997**, *387*, 381–384.
- (32) Daura, X.; van Gunsteren, W. F.; Rigo, D.; Jaun, B.; Seebach, D. *Chem. Eur. J.* **1997**, *3*, 1410–1417.
- (33) Daura, X.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *J. Mol. Biol.* **1998**, *280*, 925–932.
- (34) Daura, X.; van Gunsteren, W. F.; Mark, A. E. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 269–280.
- (35) Glättli, A.; Daura, X.; Bindschadler, P.; Jaun, B.; Mahajan, Y. R.; Mathad, R. I.; Rueping, M.; Seebach, D.; van Gunsteren, W. F. *Chem. Eur. J.* **2005**, *11*, 7276–7293.
- (36) MacKerell, A. D., Jr.; et al.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- (37) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700–733.
- (38) Gao, J. *Methods and Applications of Combined Quantum Mechanical and Molecular Mechanical Potentials*; VCH: New York, 1995; Vol. 7 of *Reviews in Computational Chemistry*.
- (39) Shurki, A.; Warshel, A. *Adv. Prot. Chem.* **2003**, *66*, 249.
- (40) Friesner, R. A.; Guallar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389–427.
- (41) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.
- (42) Kolb, M.; Thiel, W. *J. Comput. Chem.* **1993**, *14*, 775–789.
- (43) Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. *Chem. Phys.* **2001**, *263*, 203–219.
- (44) Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. *Chem. Phys.* **2000**, *256*, 15–27.
- (45) Hu, H.; Elstner, M.; Hermans, J. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 451–463.
- (46) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 569–585.
- (47) Zhang, X.; Harrison, D.; Cui, Q. *J. Am. Chem. Soc.* **2002**, *124*, 14871–14878.
- (48) Li, G.; Cui, Q. *J. Am. Chem. Soc.* **2003**, *125*, 15028–15038.
- (49) Bondar, A. N.; Fischer, S.; Smith, J. C.; Elstner, M.; Suhai, S. *J. Am. Chem. Soc.* **2004**, *126*, 14668–14677.
- (50) Elstner, M.; Frauenheim, T.; Suhai, S. *THEOCHEM* **2003**, *632*, 29.
- (51) Kruger, T.; Elstner, M.; Schiffels, P.; Frauenheim, T. *J. Chem. Phys.* **2005**, *122*, 114110.
- (52) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149–5155.

- (53) Liu, H.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Hermans, J.; Yang, W. *Proteins* **2001**, *44*, 484.
- (54) Parr, R. G.; Yang, W. T. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.
- (55) Saebo, S.; Pulay, P. *Annu. Rev. Phys. Chem.* **1993**, *44*, 213–236.
- (56) Petersson, G. A.; Bennett, A.; Tensfeldt, T. G.; Allaham, M. A.; Shirley, W. A.; Mantzaris, J. *J. Chem. Phys.* **1988**, *89*, 2193–2218.
- (57) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (58) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (59) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (60) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650–654.
- (61) Florian, J.; Johnson, B. G. *J. Phys. Chem.* **1995**, *99*, 5899–5908.
- (62) Lii, J. H.; Ma, B. Y.; Allinger, N. L. *J. Comput. Chem.* **1999**, *20*, 1593–1603.
- (63) Saebo, S.; Tong, W.; Pulay, P. *J. Chem. Phys.* **1993**, *98*, 2170–2175.
- (64) Murphy, R. B.; Beachy, M. D.; Friesner, R. A.; Ringnalda, M. N. *J. Chem. Phys.* **1995**, *103*, 1481–1490.
- (65) Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. *J. Am. Chem. Soc.* **1997**, *119*, 5908–5920.
- (66) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; König, P.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, *110*, 6458–6469.
- (67) Elstner, M. *Theor. Chem. Acc.* **2006**, *116*, 316–325.
- (68) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149–5155.
- (69) Wu, Q.; Yang, W. T. *J. Chem. Phys.* **2002**, *116*, 515–524.
- (70) Kristyan, S.; Pulay, P. *Chem. Phys. Lett.* **1994**, *229*, 175–180.
- (71) Frisch, M. J.; Pople, J. A. et al. *Gaussian 03, Revision B.05*; Gaussian, Inc.: Wallingford, CT, 2004.
- (72) Schrödinger, L. L. C. *Jaguar 5.5*; New York, 1991–2006.
- (73) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (74) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 569–585.
- (75) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (76) Im, W.; Berneche, S.; Roux, B. *J. Chem. Phys.* **2001**, *114*, 2924–2937.
- (77) Schaefer, P.; Riccardi, D.; Cui, Q. *J. Chem. Phys.* **2005**, *123*, 014905.
- (78) Brooks, C. L.; Karplus, M. *J. Chem. Phys.* **1983**, *79*, 6312–6325.
- (79) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (80) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.
- (81) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.
- (82) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
- (83) Ramachandran, G. N.; V. S. *Adv. Prot. Chem.* **1968**, *23*, 283–438.

CT600352E

Molecular Dynamics Simulations of Proteins: Can the Explicit Water Model Be Varied?

David R. Nutt^{*,†} and Jeremy C. Smith^{†,‡}

Computational Molecular Biophysics, IWR, Im Neuenheimer Feld 368, University of Heidelberg, 69120 Heidelberg, Germany, and Center for Molecular Biophysics, Oak Ridge National Laboratory/University of Tennessee, P.O. Box 2008, 1 Bethel Valley Road, Oak Ridge, Tennessee 37831

Received March 6, 2007

Abstract: In molecular mechanics simulations of biological systems, the solvation water is typically represented by a default water model which is an integral part of the force field. Indeed, protein nonbonding parameters are chosen in order to obtain a balance between water–water and protein–water interactions and hence a reliable description of protein solvation. However, less attention has been paid to the question of whether the water model provides a reliable description of the water properties under the chosen simulation conditions, for which more accurate water models often exist. Here we consider the case of the CHARMM protein force field, which was parametrized for use with a modified TIP3P model. Using quantum mechanical and molecular mechanical calculations, we investigate whether the CHARMM force field can be used with other water models: TIP4P and TIP5P. Solvation properties of N-methylacetamide (NMA), other small solute molecules, and a small protein are examined. The results indicate differences in binding energies and minimum energy geometries, especially for TIP5P, but the overall description of solvation is found to be similar for all models tested. The results provide an indication that molecular mechanics simulations with the CHARMM force field can be performed with water models other than TIP3P, thus enabling an improved description of the solvent water properties.

Introduction

Water plays an essential role in all living organisms. Enzymes, for example, require a certain level of hydration before they can perform their biological function.¹ Therefore, it is important that biomolecular simulations should try and recreate the aqueous environment of biomolecules as accurately as possible by including the effects of water, either implicitly or explicitly.

Most modern molecular mechanics (MM) force fields, such as CHARMM,² AMBER,³ and GROMOS,⁴ are typically designed for use with a specific water model, chosen during

the parametrization process. The OPLS-AA force field is slightly different, since it is designed to be compatible with three different water models.⁵ The majority of biomolecular simulations are then performed using the default water model. However, for any given simulation, the question arises as to whether the water model is suitable under the precise simulation conditions. In the case of biological systems under physiological conditions, the default description of water is probably satisfactory. However, this is probably not the case for simulations at low temperatures or high pressure, for example.

Ideally, one should be able to select the most appropriate water model for the chosen simulation conditions in order to obtain the best possible description of the system being studied. However, in practice, as the water model is an integral part of the force field, it is possible that simply

* Corresponding author phone: +49-6221-548805; fax: +49-6221-548868; e-mail: david.nutt@iwr.uni-heidelberg.de.

† University of Heidelberg.

‡ Oak Ridge National Laboratory/University of Tennessee.

changing the water model will lead to an imbalance in the protein–water and water–water interactions. However, this issue has not yet been investigated.

In the main part of this paper, we investigate the effect of using different water models on the solvation properties of a number of small molecules and a small protein described with the CHARMM force field. Before this, it is instructive to consider water models themselves and the central role of the water model in the parametrization of biomolecular force fields.

Models for Water. In the development and parametrization of water models, two separate approaches can be identified. The first supposes that if the details of the interactions (electrostatic, repulsive, and dispersion interactions, for example) between two atoms or molecules can be correctly described (for example, by reproducing high-level *ab initio* calculations), the resulting potential will also be suitable for describing larger clusters as well as all possible phases. An example of such a potential is provided by the family of anisotropic site–site potentials (ASP) for water developed by Stone and co-workers.^{6–8} Although these so-called *ab initio* interaction potentials have yielded insight into the structure and properties of small water clusters⁹ and the adsorption of water on surfaces,^{10–12} they suffer from being computationally intensive, leading to limitations in their fields of application (at least with current computational resources).

Alternatively, the potential can be parametrized in order to reproduce bulk properties for a chosen phase under given conditions. This empirical approach gave rise to the family of Transferable Intermolecular Potentials (TIP) developed by Jorgensen and co-workers: TIP3P,¹³ TIP4P,¹³ and TIP5P.¹⁴ Other widely used empirical water models include the Simple Point Charge models developed by Berendsen and co-workers (SPC¹⁵ and SPC/E¹⁶) and the Stillinger and Rahman model (ST2).¹⁷ These models describe bulk liquid water with varying degrees of accuracy.¹⁸ However, they are not capable of correctly describing small clusters or other phases since they were neither designed nor parametrized for these purposes. Biomolecular simulations typically employ empirical water models due to their computational efficiency.

Water as a Solvent in Biological Systems. Early biomolecular simulations were carried out either in vacuum or in an environment of fixed dielectric constant in order to reduce the computational expense. In most modern simulations, however, water is explicitly included in order to describe the system as completely as possible. In some cases, such as very large protein systems, it sometimes remains necessary to use one of a range of implicit solvent models, such as those based on Generalized Born approaches.^{19–22}

In simulations involving explicit water, it is crucial that a balance exists between water–protein and water–water interactions in order to describe correctly the water–protein interface. This balance is ensured by careful parametrization. In the following we consider the parametrization of the CHARMM force field, although other force fields have used similar approaches.

Water and the CHARMM Force Field. In both the CHARMM19^{23,24} and CHARMM22²⁵ force fields, the basis

was chosen to be a modified version of the TIP3P water model,¹³ since this model provides a satisfactory description of first-shell hydration and the energetics of liquid water while remaining computationally inexpensive. The modification of the original model involved the addition of van der Waals interaction sites to the H atoms,^{23,24} but the effects of these additional sites on the properties of the modified water model relative to the original TIP3P model have been found to be small.²⁶ To differentiate between the original TIP3P model and the CHARMM-modified TIP3P model, we use the acronym mTIP3P to indicate the latter.

The second stage of the parametrization was to determine peptide backbone parameters using the model compound N-methyl acetamide (NMA). The nonbonding parameters for the atoms in NMA (partial charges and van der Waals parameters) were chosen in order to reproduce the binding energy and minimum energy structure of NMA–water and NMA–NMA dimers as determined from *ab initio* calculations at the HF/6-31G* level. mTIP3P water and NMA can therefore be considered as the foundations on which the rest of the CHARMM force field is built.

Tests of the parametrization included calculation of the molecular volume and heat of solvation of NMA, which were found to be in good agreement with experiment.²⁵ This agreement was taken to indicate that the solute–solute and solute–solvent interactions are appropriately balanced in the CHARMM22 parameter set.

Recognizing that the solvation energetics are of critical importance for many biomolecular processes, such as protein folding and biomolecular association, it is also interesting to note that the latest parameter sets for the GROMOS force field²⁷ have been parametrized explicitly to reproduce the experimental solvation free enthalpies of a range of small polar molecules in cyclohexane and in water. More recently, the transferability of these parameters to the calculation of solvation properties in other solvents has been demonstrated.²⁸

Recently, there has been a growing interest in comparing force fields and determining which combination(s) of biomolecular force field and water model gives the most satisfactory results.^{29–33} These studies have mainly employed thermodynamic criteria in their assessment, such as solvation free energies. Although it is indeed necessary that thermodynamic properties are correctly reproduced, so far little work has been done on examining the details of the water–solute interactions at the atomic level.

Limitations of the Current Force Field. TIP3P (original or modified) is not a perfect water model. In particular, TIP3P water is found to display too little structuring, with the second peak in the oxygen–oxygen radial distribution function (g_{OO}) almost completely absent.¹³ The isothermal compressibility is too low, and the coefficient of thermal expansion is too high. It must also be remembered that the model was parametrized for 1 atm and 25 °C. Away from these conditions it must be used with caution. This point is reinforced by noting that the freezing temperature of the TIP3P model has recently been calculated to be 146 K.³⁴

The TIP3P model was originally designed for bulk water simulations. Because of this, many-body interactions and

polarization effects are described in an empirical way by increasing the dipole moment of each water molecule relative to the gas phase; the dipole moment of a TIP3P water is 2.35 D, compared to 1.85 D determined experimentally for the gas phase.³⁵ Because of this, TIP3P is unable to describe correctly the water dimer minimum energy structure; the binding energy is too high; and the O...O distance too short. This suggests that the TIP3P model is not ideal for investigating the details of protein–water interactions, for example, in the case of buried water molecules or water molecules at the protein surface, since it can be expected that polarization effects will be significant in such situations.

It is therefore of significant interest to investigate the behavior of the CHARMM22 force field with water models other than the modified TIP3P model. Once this behavior is understood, it will be possible to choose the most appropriate water model for any given simulation being carried out. For example, this might be the TIP4P/Ew model when Ewald summation is being used,³⁶ the TIP5P model if the simulations are being carried out close to the water density maximum at 4 °C,¹⁴ TIP4P/Ice if simulations are to be carried out with ice–water coexistence,³⁷ the Gaussian charge polarizable model (GCPM) for systems under high pressure,³⁸ or the ab initio, anisotropically polarizable ASP-W2K model if an accurate description of a small number of surface or buried water molecules are of interest.⁸

In the rest of this paper we fix ourselves a more modest aim—to investigate the effect of using the TIP4P and TIP5P models to describe solvent water in biomolecular simulations that use the CHARMM22 force field. The use of the CHARMM22 parameter set with water models other than TIP3P may in principle lead to inconsistencies because the water–protein and protein–protein intermolecular parametrization may not be well balanced. Here, we investigate and assess these potential inconsistencies by examining thermodynamic and structural properties of the different models in a range of test cases.

The work is presented as follows. First, the gas-phase interactions of the water models with themselves and with NMA are investigated in order to assess their ability to reproduce ab initio data and to illustrate the importance of a balanced potential. The free energy of solvation for NMA in the different water models is calculated, together with some structural properties of water around NMA. In addition to NMA, we also consider the solvation of a small number of other model compounds which are representative of amino acid side chains: ethane, benzene, acetate, and guanadinium as well as a small protein, crambin. In light of the results obtained, we then draw some conclusions.

Methods

Geometry optimizations and molecular dynamics simulations were performed using the CHARMM program, version c31b2,² and ab initio calculations were performed using the CADPAC suite of programs.³⁹

The topology and parameters for NMA and the other solute molecules were taken from the CHARMM22 parameter set.²⁵ Geometry optimizations were performed using the Conjugate Gradient or Adopted-Basis Newton–Raphson (ABNR) al-

gorithms with a tolerance of 1×10^{-5} kcal/mol Å⁻¹ unless otherwise stated. Molecular dynamics simulations were performed in the NVE ensemble at 300 K, following heating and sufficient equilibration. A time step of 1 fs was used. The SHAKE algorithm was used to constrain all bonds to hydrogen,⁴⁰ and the water models were treated as rigid.

The solvation free energies were determined as described in ref 41, where the solvation free energies (ΔA) of N-methylacetamide and methylamine in CHARMM-modified TIP3P water were calculated. Briefly, the NMA molecule was positioned at the center of a sphere of water molecules with radius 16 Å, taken from a 70 Å cubic box equilibrated at 300 K. Water molecules within 2.8 Å of any NMA atom were deleted. The NMA molecule was then constrained to the center of the sphere with a harmonic force constant of 1 kcal/mol. The system was simulated using the stochastic boundary method⁴² with a reaction region of radius 12 Å, in which the system is propagated using Newtonian dynamics, and a buffer region of radius 4 Å around the reaction region in which the motion is simulated using Langevin dynamics. The Langevin friction coefficient for the oxygen atoms in the buffer region was set to 62 ps⁻¹.

For the remaining solute molecules, the setup was performed in the same way. In each case, the atom constrained to the center of the sphere was as follows: acetate—the carbon atom of the CO₂ group, ethane and benzene—one of the carbon atoms, and guanadinium—the central carbon atom.

Other simulation techniques can also be used to calculate solvation free energies. Price and Brooks used a Monte Carlo method to determine solvation free energies of 40 mono- and disubstituted benzenes,⁴³ whereas Shirts et al. favored molecular dynamics with periodic boundary conditions.³⁰ The stochastic boundary method was chosen in this work for direct comparison with the results of ref 41. In addition, this protocol provides satisfactory accuracy while reducing the computational effort required.

Following ref 41, solvation free energies were calculated by performing simulations at λ values of 0.02, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.98, where λ represents a coupling coefficient between the solvent and solute which can take values between 0 (full coupling) and 1 (no coupling). At each λ , 100 ps of equilibration was followed by production dynamics for either 1900 ps (for $\lambda = 0.02, 0.98$) or 900 ps (for all other λ).

The solvation free energies (ΔA) were then calculated using either the exponential formula (also known as thermodynamic perturbation, TP)⁴⁴ with double-wide sampling or thermodynamic integration (TI).⁴⁵ Errors were estimated by calculating the energy over 10 ps batches and obtaining the mean and standard deviation, as in ref 41.

For the simulations of crambin, the high resolution (0.54 Å) X-ray crystal structure was taken from the PDB (1ejg)⁴⁶ and imported into the CHARMM program. The protein was solvated in a cubic box of TIP n P water (with box length 50 Å), previously equilibrated at 300 K. Water molecules within 2.8 Å of a protein heavy atom were deleted, leaving between 3924 and 3936 water molecules. One hundred steps of steepest descent minimization were used to remove bad contacts. The system was then equilibrated at 300 K for 100

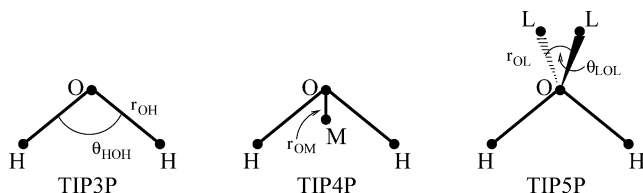


Figure 1. Structural features of the TIP3P, TIP4P, and TIP5P water models.

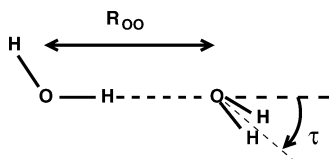


Figure 2. Structure of the linear water dimer.

ps with the protein held fixed, followed by a further 300 ps with no constraints. Two to three ns of dynamics were then calculated. Such simulation lengths are typical of studies involving the comparison of different simulation protocols.^{29,47} Simulations were performed in the NVE ensemble with periodic boundary conditions. SHAKE was used to constrain bonds to hydrogen,⁴⁰ and the water molecules were treated as rigid. The nonbonded interactions were truncated at 12 Å, with a switch function for the van der Waals interactions and a shift function for the electrostatics.

The truncation of the nonbonded interactions in this work is consistent with the original parametrization procedure used in the development of the TIP water models.^{13,14} In comparing and assessing the solvation behavior of different water models, it is necessary to reduce the number of potential sources of error. It is already known that the kinetic and thermodynamic properties of TIP4P water are changed when Ewald summation is included,³⁶ and this is also likely to occur for TIP3P and TIP5P. However, one cannot assume that such changes will be similar in each case nor even go in the same direction. For these reasons, the nonbonded interactions were truncated as described above.

The TIP n P Water Potentials. The TIP n P family of water potentials represents a useful test set for investigating the compatibility of alternative water models with the CHARMM22 force field, since they possess many features common to other water potentials, for example, an interaction site at the center of mass or at positions corresponding to lone pairs, but all have the same geometry. They have the general form

$$E_{ab} = \sum_{ij} \frac{q_i q_j e^2}{r_{ij}} + 4\epsilon_0 \left[\left(\frac{\sigma_0}{r_{OO}} \right)^{12} - \left(\frac{\sigma_0}{r_{OO}} \right)^6 \right] \quad (1)$$

where i and j are the charged sites on molecules a and b separated by a distance r_{ij} , ϵ_0 and σ_0 are the van der Waals parameters between the two oxygen sites, which are separated by r_{OO} . The O–H bond lengths are fixed at 0.9572 Å and the H–O–H angle is 105.42°, corresponding to the experimental gas-phase values. In the TIP3P model, charges are placed at the O and H atom sites, with a single van der Waals site on O. In the TIP4P model, there is no longer a charge on the O site; instead it is placed at a position corresponding to the molecular center of mass (M), 0.15 Å from the oxygen

Table 1. Monomer Geometry and Parameters for the TIP n P Potential Functions for use with CHARMM^a

	TIP3P	mTIP3P	TIP4P	TIP5P
q_H/e	0.417	0.417	0.520	0.241
q_O/e	−0.834	−0.834		
q_M/e			−1.04	
q_L/e				−0.241
$\sigma_{OO}/\text{Å}$	3.5364	3.5364	3.5399	3.5021
$\epsilon_O/\text{kcal/mol}$	0.1521	0.1521	0.1550	0.16
$\sigma_{HH}/\text{Å}$		0.4490		
$\epsilon_H/\text{kcal/mol}$		0.0460		
$r_{OH}/\text{Å}$	0.9572	0.9572	0.9572	0.9572
$\theta_{HOH}/^\circ$	104.52	104.52	104.52	104.52
$r_{OM}/\text{Å}$			0.15	
$r_{OL}/\text{Å}$				0.70
$\theta_{LOL}/^\circ$				109.47

^a Because of the different definition of the van der Waals interaction energy in CHARMM compared to eq 1, the σ_0 parameters differ from those presented in the original papers by a factor of 2^{1/6}.

Table 2. Geometry and Dimerization Energy for Optimized Linear Water Dimers^a

donor	acceptor	$R_{OO}/\text{Å}$	$\tau/^\circ$	$\Delta E/\text{kcal/mol}$	E_{vdw}
Homodimers					
TIP3P	TIP3P	2.75	27.3	−6.50	1.74
mTIP3P	mTIP3P	2.77	27.4	−6.55	1.50
TIP4P	TIP4P	2.75	46.2	−6.23	1.80
TIP5P	TIP5P	2.68	51.4	−6.78	2.37
Heterodimers					
mTIP3P	TIP4P	2.79	50.3	−5.88	1.32
TIP4P	mTIP3P	2.72	21.0	−7.05	2.10
mTIP3P	TIP5P	2.63	51.7	−9.06	3.50
TIP5P	mTIP3P	2.80	30.3	−5.27	1.18
TIP4P	TIP5P	2.53	51.4	−10.60	6.03
TIP5P	TIP4P	2.83	48.8	−4.74	1.00
HF/6-31G*		2.98	56.2	−5.65	
expt ^b		2.98 ± 0.02	57. ± 10	−5.4 ± 0.5	

^a See Figure 2 for the definition of the structural parameters. In the mixed water dimers, the van der Waals parameters are obtained using the standard combining rules ($\sigma_{AB} = 1/2(\sigma_{AA} + \sigma_{BB})$, $\epsilon_{AB} = \sqrt{(\epsilon_A \epsilon_B)}$). ^b Reference 59 for (D₂O)₂.

along the bisector of the H–O–H angle. In the TIP5P model, charges are placed at sites corresponding to lone pair positions (L), 0.7 Å from the oxygen. The three models are shown schematically in Figure 1, and the monomer geometry and parameters are summarized in Table 1.

Results

Water–Water Interactions. In the original papers for the TIP n P water models,^{13,14} the minimum energy structure of the linear water dimer (constrained to C_s symmetry, with a linear O–H···O hydrogen bond, see Figure 2) was investigated. To demonstrate the importance of a balanced potential, it is instructive to investigate the structure and binding energy of linear water dimers where the donor and acceptor molecules are described with different water models. The results are given in Table 2 and clearly illustrate that such a description is “unbalanced”.

It is interesting to note the significant asymmetry in the results when the donor and acceptor molecules in the

Table 3. Interaction Energy Scaling Factors (Ab Initio → Model) for TIP n P Water Models

model	scaling factor
mTIP3P	1.16
TIP4P	1.10
TIP5P	1.20

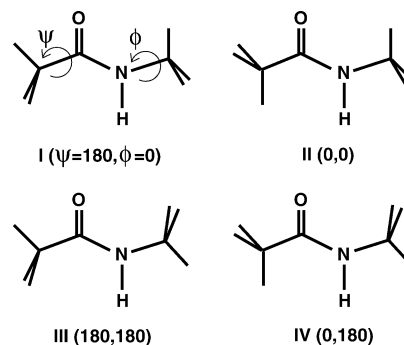
heterodimers are exchanged, in particular with dimers involving a TIP5P molecule. The interaction energies become too large in several cases, coupled to a significantly shortened O...O distance, and vice versa. For the homodimers, the optimized O...O distances are 0.2–0.3 Å shorter than the experimental distance. This difference is due partly to the fact that the TIP n P water models were parametrized for bulk water and therefore were not designed to reproduce the water dimer minimum energy geometry.

During the development of the CHARMM22 force field, the water dimer geometry from the mTIP3P model was then compared to the equivalent linear minimum energy conformation obtained from ab initio calculations at the HF/6-31G* level of theory.⁴⁸ It was found that the HF/6-31G* interaction energy was smaller than the model interaction energy and the minimum O...O distance longer, effects which were attributed to the small basis set, the low-level of theory used (which neglects correlation contributions), the use of fixed geometries, and the omission of a correction for Basis Set Superposition Error (BSSE). In order to make the energies from ab initio and model calculations directly comparable, the ratio $E_{\text{model}}/E_{\text{ab initio}} = -6.55/-5.65 = 1.16$ was used to scale the ab initio interaction energies.⁴⁸ In the same way, it was also assumed that intermolecular distances from model calculations should be 0.2 Å shorter than intermolecular distances obtained from HF/6-31G* calculations (see Table 2).⁴⁸ (In this approach, the model calculations were taken as the reference, since the correlation effects, etc. are already included in the water model in an average way.) Had the CHARMM22 force field been parametrized with a different water model, different scaling factors would have been obtained. These are presented in Table 3 and use the data from Table 2. However, as we shall see later, these scale factors are in fact of little use in considering how to use the CHARMM force field with water models other than mTIP3P.

The Interaction of TIP n P Water Molecules with NMA.

NMA has been widely used as the simplest model of a peptide backbone, and the hydrogen-bonding between water and NMA has been extensively studied.^{49–53} Since the first steps of the parametrization of the CHARMM22 force field involved the determination of the geometry and interaction energy of isomers of NMA–water complexes, calculation of equivalent data for alternative water models will give a first indication of whether there are likely to be effects due to an unbalanced potential. However, before considering the NMA–water complexes, it is first useful to consider the conformers of NMA in order to decide which should be used in the solvation calculations.

The Structure of NMA. Considering only the most favorable trans conformation of the peptide bond, and assuming C_s symmetry, a total of four possible conformers

**Figure 3.** The four conformations of NMA obtained by rotating about the angles ψ and ϕ .**Table 4.** Relative Potential Energy as Calculated with CHARMM for Four NMA Conformations, before and after Minimization

conformation	$E_{\text{before}}/\text{kcal/mol}$	$E_{\text{after}}/\text{kcal/mol}$
I	0.59	0.52
II	0.86	0.79
III	0.0	0.0
IV	0.28	0.26

are obtained, which can be interconverted through rotation of the methyl groups. These four conformers are illustrated in Figure 3, using the same numbering scheme as in ref 53.

All four conformers were constructed automatically in the CHARMM program, and their energy was minimized. The relative energies before and after minimization are given in Table 4. It can be seen that CHARMM finds isomer III to be the most stable, followed by IV, I, and II, both before and after minimization. Isomer III corresponds to the structure as reported by gas-phase electron diffraction in 1973.⁵⁴ Although this experimental structure was obtained with good precision, several of the assumptions which were necessary to enable interpretation of the experimental data have since been found to be incorrect.^{49,53} Ab initio calculations (MP2/6-31G*,⁴⁹ B3LYP/6-311++G(2d,2p),⁵¹ B3LYP/6-31+G(d,p)⁵³) have consistently shown that isomer III is not the most stable conformation, with isomers II and IV being the isoenergetic minimum energy states for isolated NMA. For hydrated NMA, isomer IV is found to be the minimum energy conformation.^{51,53} However, for completeness, we have studied water–NMA complexes involving all four conformations.

Water–NMA Complexes. We have determined the interaction energy and minimum energy geometry for water–NMA complexes in a way as consistent as possible with the method described in the CHARMM22 paper.²⁵ Three water–NMA isomers are considered here and are depicted in Figure 4 (the original CHARMM paper considered only two conformers, OHO1 and NHO). In each case, the NMA moiety was fixed in the appropriate CHARMM-optimized geometry (the water models are rigid by design), and two intermolecular degrees of freedom were optimized; namely, the H...O hydrogen bond distance, denoted R (where H and O can belong to either water or NMA), and the C–O...O or N...O–(HOH bisector) angle, θ . The hydrogen bond was

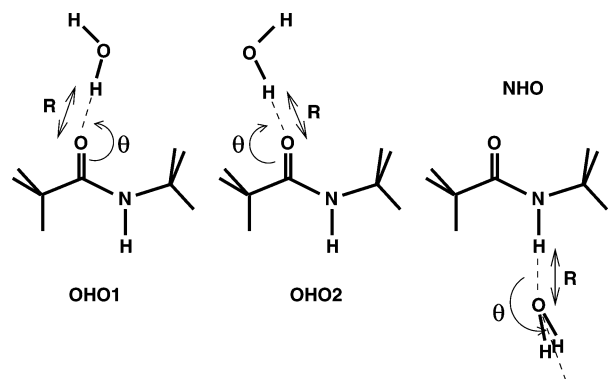


Figure 4. The three NMA–water conformations considered, denoted as OHO1, OHO2, and NHO, with the optimized degrees of freedom R and θ .

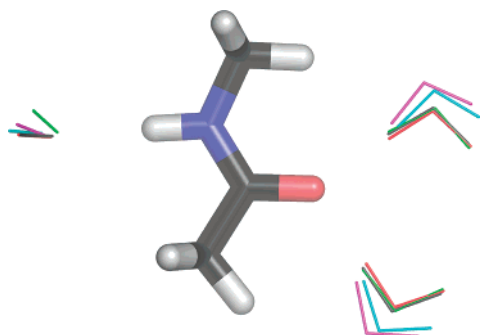


Figure 5. Superposition of NMA...water complexes after optimization of all intermolecular degrees of freedom from model and ab initio calculations. The water molecules are color-coded as follows: mTIP3P – black, TIP4P – red, TIP5P – green, HF/6-31G* – cyan, MP2/6-31G* – magenta. For clarity, the positions of the lone pairs (TIP4P and TIP5P) are not shown. All structures have C_s symmetry.

constrained to be linear, and all calculations were performed with C_s symmetry.

To assess the quality of the structures obtained, the same two degrees of freedom were optimized using ab initio calculations at the HF/6-31G* level (as in the original CHARMM paper) and at the MP2/6-31G* level. The NMA and water geometries were fixed in the optimized geometries given by CHARMM. In Table 5 we present the results for NMA–mTIP3P, together with ab initio data for comparison. As in the original CHARMM parametrization procedure, ab initio interaction energies were not corrected for BSSE.

Upon inspection of Table 5, several features are apparent. First the choice of NMA conformation (I,II,III,IV) does not significantly affect the energy or structural details of the model calculations. The complexes OHO1 and OHO2 are close in energy, with the OHO1 structure slightly more stable than OHO2 for all NMA conformations except conformation III, in which OHO1 and OHO2 are almost isoenergetic. The binding energies of the OHO conformations are found to be between 0.9 and 1.7 kcal/mol larger than the binding energy of the NHO conformation.

There is good agreement between the structural features of the model and HF calculations. The distance R is in general overestimated by around 0.2 Å in the HF calculations, for the reasons described above. The angle θ is reproduced

to within 9°. Although the ordering of stability is not always reproduced by the HF calculations, the OHO conformations are still found to be close in energy, while the binding energy of the NHO complex is approximately 2 kcal/mol smaller. In these calculations, the scaled energy (using a factor of 1.16, as described above) is not consistently closer to the model energy than the unscaled energy, except for the NHO complexes. The fact that the scaled energies presented here are not in exact agreement with the scaled energies presented in the original CHARMM22 paper²⁵ is due to the use of slightly different geometries for the NMA moiety.

In moving from the HF calculations to the MP2 level of theory, the distance R becomes shorter, as expected. The angles in the OHO complexes remain fairly close to those from the HF and model calculations. On the other hand, θ in the NHO complexes moves significantly (up to 35°) away from the approximately linear geometry found in the model and HF calculations. This strongly suggests that the mTIP3P model is indeed not capable of reproducing the details of gas-phase interactions.

Equivalent calculations were carried out for the TIP4P and TIP5P water models, and the results are presented in Table 6. Some interesting observations can be made. For all three water models, the minimized geometries are very similar, with θ varying by a maximum of 9° and R by a maximum of 0.07 Å as the water model is changed. When mTIP3P is changed to TIP4P, the binding energies of the OHO complexes increase by 0.2–0.3 kcal/mol, while the binding energies of the NHO complexes are reduced by around 0.6 kcal/mol. However, the binding energies with the TIP5P model behave quite differently. The OHO complexes have a binding energy between 6 and 7 kcal/mol, significantly smaller than found for mTIP3P and TIP4P. On the other hand, the NHO complexes are more strongly bound than with mTIP3P and TIP4P and are even more strongly bound than some of the OHO complexes. This is in disagreement with the trends found in the ab initio calculations (both HF and MP2). Since there is no systematic difference between the water models, this also indicates that the use of a scale factor calculated for a given water model in order to bring ab initio binding energies into agreement with energies from model force field calculations cannot be generalized to other water models.

The energy minimizations with two degrees of freedom presented above are useful for direct comparisons between models, but it is possible that these structures differ significantly from those found when all the intermolecular degrees of freedom are optimized. For this reason, all intermolecular degrees of freedom were optimized (in principle there are six intermolecular degrees of freedom, but the presence of a symmetry plane in all cases reduces this number to three) while keeping the monomers in their CHARMM-optimized geometry. For comparison, we also performed equivalent ab initio calculations at the HF/6-31G* and MP2/6-31G* levels.

The binding energies obtained for NMA conformation I (the others NMA conformations lead to very similar results) are given in Table 7, and the optimized structures are superimposed in Figure 5.

Table 5. Interaction Energies and Structural Data for NMA–mTIP3P Complexes^a

conformation	model			HF				MP2			
	R/Å	θ/°	E	R/Å	θ/°	E	E _{scaled}	R/Å	θ/°	E	
I	OHO1	1.76	145	-7.70	1.97	143	-7.03	-8.15	1.91	139	-8.73
	OHO2	1.76	126	-7.26	1.97	122	-7.25	-8.41	1.92	117	-8.78
	NHO	1.92	174	-6.18	2.12	167	-5.14	-5.96	2.06	167	-6.82
II	OHO1	1.76	145	-7.71	1.96	144	-7.10	-8.23	1.91	139	-8.79
	OHO2	1.77	122	-7.26	1.97	117	-7.53	-8.73	1.93	110	-9.14
	NHO	1.93	172	-6.30	2.14	179	-5.48	-6.36	2.03	215	-7.55
III	OHO1	1.76	137	-8.07	1.98	135	-7.45	-8.64	1.94	131	-9.29
	OHO2	1.76	126	-7.27	1.97	122	-7.23	-8.39	1.92	117	-8.75
	NHO	1.92	174	-6.18	2.12	173	-5.26	-6.10	2.01	167	-6.94
IV	OHO1	1.76	137	-8.08	1.98	135	-7.53	-8.73	1.94	131	-9.36
	OHO2	1.76	122	-7.26	1.98	117	-7.49	-8.69	1.94	110	-9.09
	NHO	1.92	172	-6.32	2.13	183	-5.61	-6.50	2.02	215	-7.66

^a Energies are given in kcal/mol. Ab initio calculations used the 6-31G* basis set.

Table 6. Interaction Energies and Structural Data for NMA–TIP_nP Complexes^a

conformation	mTIP3P			TIP4P			TIP5P			
	R/Å	θ/°	E	R/Å	θ/°	E	R/Å	θ/°	E	
I	OHO1	1.76	145	-7.70	1.72	147	-7.94	1.79	145	-6.42
	OHO2	1.76	126	-7.26	1.73	129	-7.43	1.79	126	-6.05
	NHO	1.92	174	-6.18	1.95	171	-5.53	1.86	150	-6.67
II	OHO1	1.76	145	-7.71	1.72	147	-7.95	1.78	146	-6.42
	OHO2	1.77	122	-7.26	1.73	126	-7.42	1.80	121	-6.07
	NHO	1.93	172	-6.30	1.96	169	-5.66	1.87	150	-6.86
III	OHO1	1.76	137	-8.07	1.72	139	-8.28	1.79	137	-6.78
	OHO2	1.76	126	-7.27	1.73	129	-7.44	1.79	126	-6.06
	NHO	1.92	174	-6.18	1.95	171	-5.53	1.86	150	-6.67
IV	OHO1	1.76	137	-8.08	1.72	139	-8.29	1.79	137	-6.78
	OHO2	1.76	122	-7.26	1.73	127	-7.43	1.80	121	-6.07
	NHO	1.92	172	-6.32	1.96	170	-5.67	1.87	150	-6.87

^a Some of the data from Table 5 are reproduced again here to aid in comparison. Energies are given in kcal/mol.

Table 7. Binding Energies in kcal/mol for NMA···Water Complexes Following Optimization of All Intermolecular Degrees of Freedom for NMA Conformation I^a

conformation	mTIP3P	TIP4P	TIP5P	HF	MP2
OHO1	-7.75	-7.94	-6.51	-7.12	-8.95
OHO2	-7.27	-7.44	-6.09	-7.41	-9.10
NHO	-6.18	-5.53	-6.92	-5.14	-6.82

^a Ab initio calculations used the 6-31G* basis set.

Although the binding energies for the TIP_nP models vary, it can be seen that the optimized structures are in fact very similar to each other. The exception is the N–H···O complex with the TIP5P model, for which the presence of lone-pair sites at the tetrahedral positions makes the H–O–H plane bend away from the N–H···O vector. Nevertheless, it is interesting to note that this bending is also apparent in the MP2/6-31G* structure (but not in the HF/6-31G* structure) although the N–H···O distance is underestimated for the TIP5P model, for the reasons discussed above. For the O–H···O complexes, all models are approximately equally distant from the ab initio structures.

Solvation Free Energy of NMA in TIP_nP Water. To assess the effect of the choice of water model on thermodynamic properties, the solvation free energy of NMA in

Table 8. Free Energies of Solvation (ΔA_{solv}) of NMA in Various Water Models in kcal/mol

water model	TP	TI	mean
mTIP3P	-11.81 ± 0.05	-11.60 ± 0.04	-11.71 ± 0.03
TIP4P	-10.22 ± 0.07	-9.94 ± 0.06	-10.08 ± 0.05
TIP5P	-10.54 ± 0.06	-10.27 ± 0.06	-10.41 ± 0.04
mTIP3P ^a	-10.4 ± 0.09	-11.3 ± 0.05	-10.85 ± 0.07
expt ^b			-10.1

^a Reference 41. ^b Reference 60.

mTIP3P, TIP4P, and TIP5P was calculated using both the thermodynamic perturbation and thermodynamic integration methods as described in the ‘Methods’ section. The resulting solvation free energies are given in Table 8.

In each case, the value obtained from TP is 0.2–0.3 kcal/mol larger in magnitude than the value from TI, and this difference gives an indication of the accuracy of the calculations. The average ΔA_{solv} for TIP4P and TIP5P are similar and lie within 0.3 kcal/mol of the experimental determination. The value for mTIP3P is 1.3 kcal/mol larger in magnitude than the TIP5P value and around 1.6 kcal/mol larger than found in experiment. The values obtained in this work for mTIP3P are slightly larger than those obtained in ref 41 but lie within 0.5 kcal/mol of the value previously

Table 9. Free Energies of Solvation (ΔA_{solv}) for Neutral Solutes in Various Water Models in kcal/mol^a

solute	mTIP3P	TIP4P	TIP5P	expt
ethane	-0.04 ± 0.03	$+0.93 \pm 0.03$	-0.09 ± 0.03	1.833 ^b
benzene	-5.09 ± 0.04	-3.35 ± 0.05	-4.37 ± 0.04	-0.767^c

^a The figures given are the mean of the values calculated with TP and TI, for which the individual values show similar trends to those observed for NMA. ^b ΔG_{solv} from ref 61. ^c ΔG_{solv} from ref 62.

obtained using TI.⁴¹ The difference in the values obtained with TI and TP in this study is of the order of 0.2 kcal/mol, compared to 0.9 kcal/mol in ref 41.

The solvation free energies obtained for TIP4P and TIP5P are closer to the experimental value than the value for mTIP3P. Once again, this suggests that the combination of the CHARMM force field with the TIP4P and TIP5P water models has the potential to give reasonable results.

We also calculated the free energies of solvation for two neutral and nonpolar solutes; ethane and benzene, which are related to the side chains of isoleucine and phenylalanine. The results are presented in Table 9. The calculated solvation free energies show only small variation with water model, as would be expected for neutral, nonpolar solutes. In both cases, the TIP4P model gives values closest to the experimentally determined value for ΔG_{solv} ; however, the values are not particularly satisfactory. The differences are partly due to comparing calculated ΔA values with experimental ΔG values but also indicate that solvation energies for small molecules are not reproduced particularly well^{30,32} unless the force field has been designed to reproduce solvation thermodynamics.^{27,28}

Given the well-known difficulties in calculating solvation free energies for charged molecules,^{55,56} we do not attempt to calculate any values for analogues of charged amino acids here.

Further insight can be obtained from the extensive study of solvation free energies in ref 32. In this study, the OPLS-AA force field⁵ was used to investigate the solvation free energies of (neutral) amino acid analogs in different water models. Although NMA was not explicitly studied, acetamide was chosen as the analog of asparagine. For the five water models investigated (TIP3P, SPC, TIP4P, SPC/E, and TIP4P-Ew), the solvation free energy of acetamide was found to vary over a range of 0.2 kcal/mol (between -8.32 (SPC/E) and -8.53 (SPC) kcal/mol). The values for TIP3P and TIP4P were identical to within the uncertainty of the simulations (-8.51 and -8.52 kcal/mol, respectively). This agreement may be due to the fact that the OPLS potential was originally designed to be compatible with the TIP3P, TIP4P, and SPC water models.⁵⁷ Although the parameters for acetamide differ between the OPLS-AA and CHARMM22 force fields, this nevertheless suggests that, for a given force field, the difference in solvation energy with different water models (at least among the models considered) is small. For the worst cases in the study of ref 32, *p*-cresol and 3-methylindole (analogs of tyrosine and tryptophan, respectively), the spread of predicted solvation free energies was found to be 0.66 and 0.86 kcal/mol. In most cases among the analogs, TIP3P (the original TIP3P model, not the CHARMM-modified

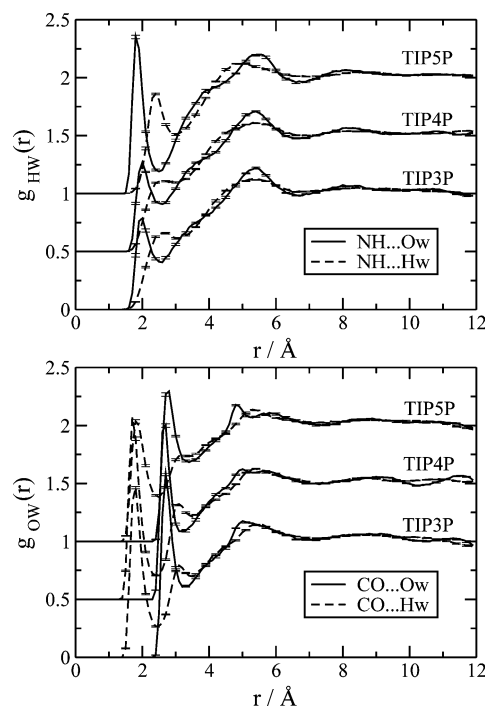


Figure 6. Solvent distribution functions $g_{\text{OW}}(r)$ and $g_{\text{HW}}(r)$, where H and O are NMA atoms and W is a water atom (H or O) for NMA in TIP n P water at 300 K. The functions for TIP4P and TIP5P have been displaced vertically for clarity. The error bars show the error on the mean determined by dividing the trajectory into ten equal sections and calculating the solvent distribution function for each section.

TIP3P model) was found to give the lowest limit in 10 out of 15 cases, suggesting that TIP3P often gives a lower (i.e., smaller if positive, more negative if negative) estimate of the solvation free energy, in agreement with the results from this work. The highest values were not dominated by one particular water model.

Structural Properties of TIP n P Solvation. In addition to considering the interaction of individual water molecules with NMA and the calculation of thermodynamic properties, it is also instructive to investigate the structure of the water around the NMA and, in particular, the solvent distribution functions around the N–H hydrogen bond donor and the C–O hydrogen bond acceptor moieties. The NH–W and CO–W distribution functions (where W is a water atom, either oxygen or hydrogen) were calculated from a 1.9 ns trajectory using stochastic boundary conditions (as described in the ‘Methods’ section) for each of the TIP water models. The results are shown in Figure 6.

The CO–W distribution functions are almost identical for the three models. This is because the H atom parameters are similar in all three cases. For the NH–W distribution functions, the mTIP3P and TIP4P profiles are similar, whereas the TIP5P profile is significantly different, due to the presence of lone pairs in the TIP5P model which become involved in the NH–O interaction. The NH–O peak in the TIP5P distribution function is sharper and larger than for mTIP3P and TIP4P, indicating a larger number of water molecules involved in hydrogen bonds at this site. The O atoms in the first hydration shell are also slightly closer to the NH donor, with a peak position of 1.9 Å for TIP5P

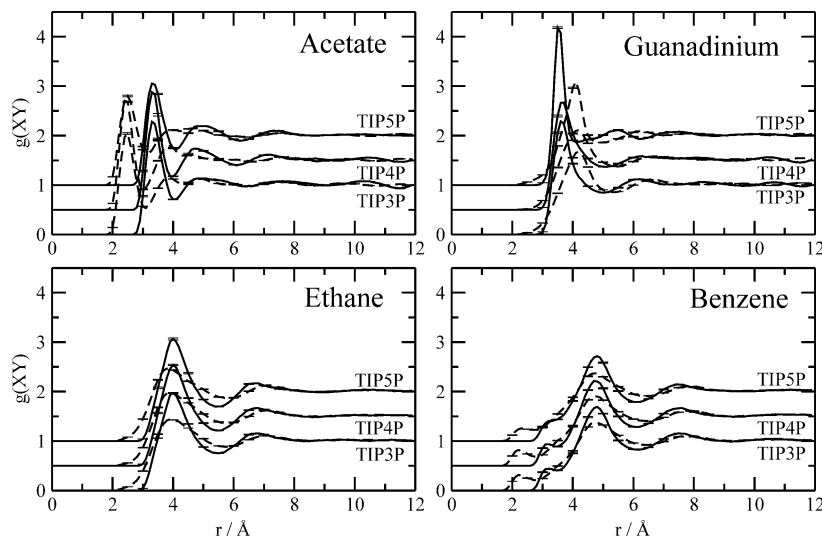


Figure 7. Solvent distribution functions $g_{XY}(r)$ for acetate, guanadinium, ethane, and benzene in TIP n P water at 300 K, where Y is a water atom (O – solid line, H – dashed line) and X is a reference atom or site (see text for details). The functions for TIP4P and TIP5P have been displaced vertically for clarity. The error bars show the error on the mean determined by dividing the trajectory into ten equal sections and calculating the solvent distribution function for each section.

compared to 2.1 Å for TIP4P and 2.0 Å for mTIP3P. Which of these descriptions is closer to reality is not yet known. A recent experimental neutron diffraction study has provided detailed structural information for solvated L-glutamic acid;⁵⁸ a similar study of NMA would provide the experimental data necessary to evaluate these distribution functions. Beyond the first solvation shell, the structure of the distribution functions are similar, although there is little useful information beyond the first peak and trough. The feature at around $r = 4.8$ Å in the CO–O distribution functions can be assigned to the first solvation shell of the NH group on the other side of the molecule.

In ref 32, no analogs of charged amino acids were considered. Such cases are likely to provide a tough challenge for the water models, since hydrogen-bonding between the solute and solvent will be stronger in these cases than for neutral species. To investigate this aspect, we calculated solvent distribution functions around two charged model compounds, acetate and guanadinium, which are closely related to the side chains of the amino acids aspartate and arginine. We also considered two neutral and nonpolar solutes, ethane and benzene, which are similar in nature to the side chains of isoleucine and phenylalanine.

The distribution functions for acetate (around the carbon atom of the CO₂ group), guanadinium, benzene, and ethane (around the geometric center of the molecule) are shown in Figure 7. For the neutral, nonpolar solutes, the profiles for all water models are almost identical. However for charged solutes, the TIP5P model behaves differently; for guanadinium, $g(XO)$ displays a very large first peak, corresponding to strong NH–O interactions. The opposite can be seen for acetate, where the solute–water hydrogen bonds are weaker for TIP5P than for the other models.

Dynamics of Crambin in TIP n P Water. One potential side effect of using an inappropriate water model could be instabilities in the protein structure or even unfolding. For this reason, simulations of the small protein crambin were performed using the TIP n P water models. To our knowledge,

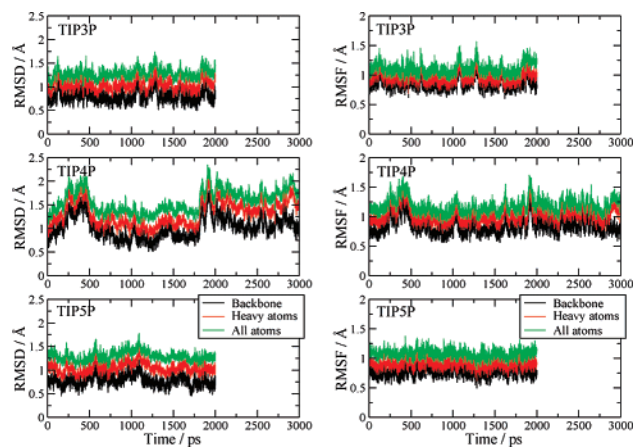


Figure 8. Left: Root-mean-square deviation (rmsd) of the protein structure from the crystal structure during the simulations. Right: Root-mean-square fluctuation (rmsf) of the protein about the mean structure over the simulations.

these are the first simulations of a biomolecular system performed using the TIP5P water model.

In Figure 8(left), the root-mean-square deviation (rmsd) is presented as a function of time from each of the simulations. The rmsd is taken with respect to the crystal structure.⁴⁶ The protein remains stable in the mTIP3P and TIP5P simulations, with average backbone rmsd values of 0.80 and 0.78 Å, respectively. The average backbone rmsd from the TIP4P simulation is slightly larger, with a value of 0.96 Å, due to the fluctuations observed at $t = 250$ – 500 ps and $t = 1750$ – 2000 ps. To check that this was not due to unfolding of the protein, the trajectory was continued for a further nanosecond, during which the rmsd was found to remain stable, with a mean value of 1.01 Å.

The root-mean-square fluctuation of the protein along the trajectories (determined with respect to the corresponding mean structure) is shown in Figure 8 (right) and also indicates that the structure is stable in all three simulations. Decomposition of the rmsf into contributions per residue reveals

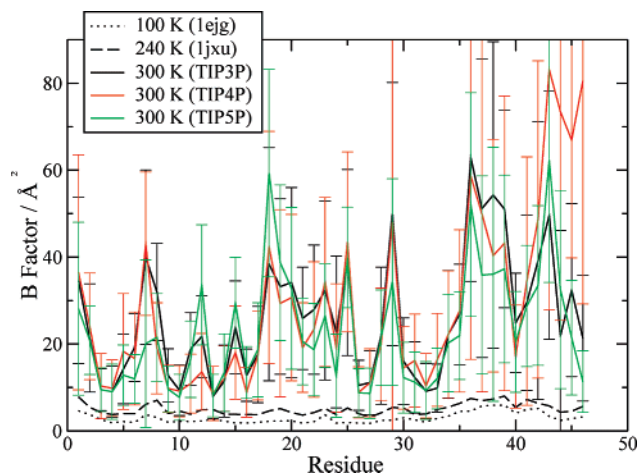


Figure 9. Residue fluctuations (plotted as thermal B -factors) from the simulations and from two crystal structures. The error bars represent ± 1 standard deviation about the mean.

that the fluctuations in the rmsd from the TIP4P trajectory are due to motion of sections of the β -sheet, between residues 42 and 46, as shown in Figure 9. Comparison with thermal B -factors from crystal diffraction data reveal common features (the same flexible and rigid regions) but are not in quantitative agreement. This is because the simulations were performed at 300 K, whereas the crystallographic data was measured at 100 K (1ejg) and 240 K (1jxu).

Discussion and Conclusions

In this paper we have investigated various features of the interaction between N-methylacetamide, other small solute molecules, and a small protein (all represented with the CHARMM22 force field), with three different water models, mTIP3P, TIP4P, and TIP5P, in order to assess whether the use of the CHARMM22 force field with water models not considered in the original parametrization leads to an imbalance in simulations.

Results obtained for NMA–water complexes with mTIP3P and TIP4P show very similar structural and energetic properties. In addition, the free energy of solvation for NMA in TIP4P appears to be closer to the experimental value than that calculated with mTIP3P. This is not surprising, since TIP4P provides a better description of the bulk water structure, while not significantly distorting the details of the NMA–water interactions. It therefore seems likely that TIP4P will give reasonable local structural and energetic results when used as a solvent model together with the CHARMM22 protein force field.

TIP5P, on the other hand, behaves differently to mTIP3P and TIP4P, mainly due to the presence of lone pair sites on the oxygen atom. These alter the details of the interactions, in particular when the water oxygen acts as a hydrogen acceptor. This can be seen most clearly in the $\text{NH}\cdots\text{O}$ distribution function for NMA and guanadinium. Whether this is due to the functional form of the TIP5P model itself (i.e., the presence of lone pair sites) or to the TIP5P parameters remains to be investigated.

Despite the differences observed for TIP5P in the details of the solvation structure, simulations of crambin show that

the protein remains stable in TIP5P water as well as in mTIP3P and TIP4P water. This gives a first indication that TIP5P may be used (with care) in biomolecular simulations using the CHARMM22 force field. This is significant, since TIP5P will provide a much better description than TIP3P in simulations at low temperature. Even at room temperature, the use of TIP5P can be expected to give a significant improvement over TIP3P, since the thermodynamic and kinetic properties of the TIP5P model are much closer to those of real water than those of TIP3P. The additional computational expense required for the five-point model is largely offset by the computing power now available.

Further comparison of simulation results with experimental data will be required in order to assess which of the water models gives the best description of the details of the protein–water interface. Furthermore, extensive work will be necessary examining a wide range of structural, dynamical, and thermodynamical properties of small and large biomolecules in solution before a clear picture emerges of the relative behavior of different standard water molecular mechanics models with any given macromolecular force field. However, the present results, although limited in the force fields tested and properties examined, do suggest that such research will be potentially fruitful.

Acknowledgment. D.R.N. is grateful to the Swiss National Science Foundation (SNF) for the award of an Advanced Research Fellowship.

References

- (1) Teeter, M. M. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 577–600.
- (2) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (3) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.
- (4) Hermans, J.; Berendsen, H. J. C.; van Gunsteren, W. F.; Postma, J. P. M. *Biopolymers* **1984**, *23*, 1513–1518.
- (5) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (6) Millot, C.; Stone, A. J. *Mol. Phys.* **1992**, *77*, 439–462.
- (7) Millot, C.; Soetens, J.-C.; Martins Costa, M. T. C.; Hodges, M. P.; Stone, A. J. *J. Phys. Chem. A* **1998**, *102*, 754–770.
- (8) Nutt, D. R.; Stone, A. J. *J. Chem. Phys.* **2002**, *117*, 800–807.
- (9) Hodges, M. P.; Stone, A. J.; Xantheas, S. S. *J. Phys. Chem. A* **1997**, *101*, 9163–9168.
- (10) Engkvist, O.; Stone, A. J. *J. Chem. Phys.* **1999**, *110*, 12089–12096.
- (11) Sadtchenko, V.; Ewing, G. E.; Nutt, D. R.; Stone, A. J. *Langmuir* **2002**, *18*, 4632–4636.
- (12) Nutt, D. R.; Stone, A. J. *Langmuir* **2004**, *20*, 8715–8720.
- (13) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (14) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910–8922.

- (15) Berendsen, H. J. C.; Postma, J. P. M.; von Gunsteren, W. F.; Hermans, J. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*; Pullman, B., Ed.; D. Reidel: Dordrecht, Holland, 1981.
- (16) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
- (17) Stillinger, F. H.; Rahman, A. *J. Chem. Phys.* **1974**, *60*, 1545.
- (18) Guillot, B. *J. Mol. Liq.* **2002**, *101*, 219–260.
- (19) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- (20) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578–1599.
- (21) Lee, M. S.; Feig, M.; Salsbury, F. R.; Brooks, C. L. *J. Comput. Chem.* **2003**, *24*, 1348–1356.
- (22) Koppole, S.; Smith, J. C.; Fischer, S. *J. Mol. Biol.* **2006**, *361*, 604–616.
- (23) Reiher, W. E. Thesis, Harvard University, 1985.
- (24) Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 1902–1921.
- (25) MacKerell, A. D., Jr. et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- (26) Mark, P.; Nilsson, L. *J. Phys. Chem. A* **2001**, *105*, 9954–9960.
- (27) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.
- (28) Geerke, D. P.; van Gunsteren, W. F. *Phys. Chem. Chem. Phys.* **2006**, *7*, 671–678.
- (29) Price, D. J.; Brooks, C. L., III *J. Comput. Chem.* **2002**, *23*, 1045–1057.
- (30) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740–5761.
- (31) Kim, B.; Young, T.; Harder, E.; Friesner, R. A.; Berne, B. *J. Phys. Chem. B* **2005**, *109*, 16529–16538.
- (32) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *2005*, 134508.
- (33) Hess, B.; van der Vegt, N. F. A. *J. Phys. Chem. B* **2006**, *110*, 17616–17626.
- (34) Vega, C.; Sanz, E.; Abascal, J. L. F. *J. Chem. Phys.* **2005**, *122*, 114507.
- (35) Clough, S. A.; Beers, Y.; Klein, G. P.; Rothman, L. S. *J. Chem. Phys.* **1973**, *59*, 2254–2259.
- (36) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665–9678.
- (37) Abascal, J. L. F.; Sanz, E.; García Fernández, R.; Vega, C. *J. Chem. Phys.* **2005**, *122*, 234511.
- (38) Paricaud, P.; Předota, M.; Chialvo, A. A.; Cummings, P. T. *J. Chem. Phys.* **2005**, *122*, 244511.
- (39) Amos, R. D. *CADPAC: The Cambridge Analytic Derivatives Package, Issue 6*; Technical Report; University of Cambridge: 1995. A suite of quantum chemistry programs developed by R. D. Amos with contributions from I. L. Alberts, J. S. Andrews, S. M. Colwell, N. C. Handy, D. Jayatilaka, P. J. Knowles, R. Kobayashi, K. E. Laidig, G. Laming, A. M. Lee, P. E. Maslen, C. W. Murray, J. E. Rice, E. D. Simandiras, A. J. Stone, M. D. Su, and D. J. Tozer.
- (40) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comp. Phys.* **1977**, *23*, 327–341.
- (41) Wan, S.; Stote, R. H.; Karplus, M. *J. Chem. Phys.* **2004**, *121*, 9539–9548.
- (42) Brooks, C. L., III; Karplus, M. *J. Chem. Phys.* **1983**, *79*, 6312–6325.
- (43) Price, D. J.; Brooks, C. L., III *J. Comput. Chem.* **2005**, *26*, 1529–1541.
- (44) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420–1426.
- (45) Straatsma, T. P.; Berendsen, H. J. C. *J. Chem. Phys.* **1988**, *89*, 5876–5886.
- (46) Jelsch, C.; Teeter, M. M.; Lamzin, V.; Pichon-Pesme, V.; H. B. R. Lecomte, C. *Proc. Natl. Acad. Sci.* **2000**, *97*, 3171–3176.
- (47) van der Spoel, D.; van Maaren, P. J. *J. Chem. Theory Comput.* **2006**, *2*, 1–11.
- (48) MacKerell, A. D. Jr.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 10559–10560.
- (49) Guo, H.; Karplus, M. *J. Phys. Chem.* **1992**, *96*, 7273–7287.
- (50) Baudry, J.; Smith, J. C. *J. Mol. Struct. (Theochem)* **1994**, *308*, 103–113.
- (51) Han, W.-G.; Suhai, S. *J. Phys. Chem.* **1996**, *100*, 3942–3949.
- (52) Buck, M.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 11000–11015.
- (53) Mennucci, B.; Martínez, J. M. *J. Phys. Chem. B* **2005**, *109*, 9818–9829.
- (54) Kitano, M.; Fukuyama, T.; Kuchitsu, K. *Bull. Chem. Soc. Jpn.* **1973**, *46*, 384–387.
- (55) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *110*, 16066–16081.
- (56) Kastenholz, M. A.; Hünenberger, P. H. *J. Chem. Phys.* **2006**, *124*, 224501.
- (57) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.
- (58) McLain, S. E.; Soper, A. K.; Watts, A. *J. Phys. Chem. B* **2006**, *110*, 21251–21258.
- (59) Fellers, R. S.; Leforestier, C.; Braly, L. B.; Brown, M. G.; Saykally, R. J. *Science* **1999**, *284*, 945–948.
- (60) Wolfenden, R. *Biochemistry* **1978**, *17*, 201–204.
- (61) Wilhelm, E.; Battino, R.; Wilcock, R. J. *Chem. Rev.* **1977**, *77*, 219–262.
- (62) Wauchope, R. D.; Haque, R. *Can. J. Chem.* **1972**, *50*, 133.

Atomic Charge Calculation of Metallobiomolecules in Terms of the ABEEM Method

Zhong-Zhi Yang^{*,†} and Bao-Qiu Cui^{†,‡}

College of Chemistry and Chemical Engineering, Liaoning Normal University, Dalian 116029, P. R. China, and Department of Chemistry, Jinzhou Teacher College, Jinzhou 121000, P. R. China

Received December 25, 2006

Abstract: Applying the atom-bond electronegativity equalization method (ABEEM) to metallo-biomolecules, the ABEEM parameters for transition metals (V, Cr, Mn, Fe, Co, Ni, Cu, and Zn) were calibrated through linear regression and least-squares optimization by choosing more than 300 training molecules. The quality of the ABEEM charge calculated in terms of the optimized electronegativity and hardness parameters for the training set is assessed by comparison with B3LYP/6-31G* charges. For a check, the ABEEM charges of some large metallobiomolecules have been calculated, and the obtained results correlate quite well to those calculated with the B3LYP/6-31G* method. The linear correlation coefficients R are all over 0.98. This shows that the ABEEM method can predict the charge distributions of large metallobiomolecules with high accuracy.

Introduction

Electron density determines all properties of a molecular system. Therefore, atomic charges of a molecule, as concentrated electron density distribution, are of great importance. On the one hand, atomic charges certainly show Coulomb interactions between different molecular sites. The atomic charges are useful indexes of molecular reactivity, particularly for electrophilic and/or nucleophilic reactions. Furthermore, atomic charges are indicators of molecular sites between which the hydrogen bonds may form for both intramolecular and intermolecular cases. On the other hand, atomic charges are also used in many packages, where they may be treated on an equal footing as important parameters, such as bond lengths, etc. The electronegativity equalization method based on DFT^{1–4} is such an approach that allows fast calculation of atomic charges in a large set of molecules. In this field, a noteworthy one is Mortier and Nalewajski's electronegativity equalization method (EEM), which has been used to predict atomic charges in molecules, electron population normal modes, etc.^{5–6} Besides, the charge equili-

bration method (Qeq) was developed by Rappé et al.⁷ Cioslowski et al. analyzed electron flow and electronegativity equalization by using charge-constrained calculations.⁸ York and Yang presented the chemical potential equalization principle to describe the redistribution of electrons upon perturbation by an applied field.⁹ De Proft et al. presented a nonempirical electronegativity equalization scheme.¹⁰ Ghosh put forward a semiempirical electronegativity equalization procedure to predict bond energies of diatomic molecules.¹¹ No et al. proposed a partial equalization of the orbital electronegativity method.¹² In order to explicitly treat the chemical bonds, Yang et al. developed an atom-bond electronegativity equalization method (ABEEM),^{13–22} which has been applied to predict charge distributions for large molecules and aqueous solutions.

Recently, Bultinck et al. have reformulated and validated the EEM approach with showing its necessity and amenability to fast calculation of atomic charges over a molecule.^{23,24} They also pointed out there was a need to extend those EEM approaches to comprise more elements in order to use them in more extensive large molecular systems. To our knowledge, there is little work that involves transition metals in those EEM methods. However, it has been estimated that a lot of proteins and enzymes purified to apparent homogeneity

* Corresponding author phone: +86-411-82159607; fax: +86-411-84258977; e-mail: zzyang@lnnu.edu.cn.

[†] Liaoning Normal University.

[‡] Jinzhou Teacher College.

require transition-metal ions as cofactors for biological function. These transition metals include V, Mn, Fe, Co, Ni, Cu, and Zn.²⁵ At present, two chromium-containing biomolecules in nature have been known.²⁶ All of these transition metals play an important role in a variety of biological systems. Thus to investigate metallobiomolecules is very interesting. The focus of the current paper is to further develop the ABEEM method so that it can be widely applied to metallobiomolecules. In this paper, the metallobiomolecules involving the first row of transition metals including V, Cr, Mn, Fe, Co, Ni, Cu, and Zn are investigated.

This article is organized as follows. First, a brief formalism of the ABEEM method is given. Second, a large set of the training molecules that contains the common organic groups as well as some transition metals is chosen and then a large amount of ab initio calculations on the training molecules in order to calibrate the ABEEM parameters is done. Third, the ABEEM parameters calibrated for transition metals, quality of the ABEEM atomic charges, and applicability of the ABEEM are discussed. Finally, a brief summary is given.

ABEEM Formalism

The electronegativity equalization principle formulated by Sanderson²⁷ states that when a molecule is formed, electronegativities of the constituent atoms become equal, yielding molecular, equalized electronegativity. Several formalisms mentioned above have been developed from this principle. In the ABEEM method, a molecule is divided into atom regions c , bond regions, and lone-pair electron regions t . By using a definition of electronegativity in light of DFT, we can express the effective electronegativity χ_c of any atom and χ_t of any bond or lone-pair electron as

$$\chi_c = \chi_c^* + 2\eta_c^* q_c + k \left[\sum_{d \neq c} \frac{q_d}{R_{c,d}} + \sum_t \frac{q_t}{R_{c,t}} \right] \quad (1)$$

$$\chi_t = \chi_t^* + 2\eta_t^* q_t + k \left[\sum_c \frac{q_c}{R_{t,c}} + \sum_{s \neq t} \frac{q_s}{R_{t,s}} \right] \quad (2)$$

In eq 1, χ_c^* and $2\eta_c^*$ are valence state electronegativity and valence state hardness of atom c , respectively; q_c and q_d are the partial charges of atom c and atom d , respectively; q_t is the partial charges of bond or lone-pair electron t ; $R_{c,d}$ represents the distance between atom c and atom d , and $R_{c,t}$ represents the distance between atom c and bond or lone-pair electron t ; and k is an overall correction coefficient in this formalism. As for the symbols in eq 2, the meanings are analogous to those of the symbols in eq 1. The electronegativity equalization principle demands $\chi_c = \chi_t = \bar{\chi}$, with $\bar{\chi}$ being the electronegativity of the molecule. When an arbitrary molecule is partitioned into m regions, one has $m+1$ unknown quantities (m charges of q and one value $\bar{\chi}$) in the m equations. These equations, along with the constraint equation on its total charge, can be solved to give $\bar{\chi}$ and the charge distribution q in the molecule if all parameters χ^* and $2\eta^*$ are known.

Calibration of the ABEEM Parameters for Metallobiomolecules

Choice of Training Set. Here we are mainly concerned with metallobiomolecules, so lots of molecules including transition metals were chosen as a training set. The initial structures of all model metallobiomolecules that include transition metals were taken directly from PDB and the Cambridge Crystallographic Data Center (CCDC). In most cases of metallobiomolecules, metal ions are coordinated with nitrogen, oxygen, and/or sulfur atoms of biological ligands, thus amino, imidazolyl, carbonyl, carboxylate, phenolate, alkoxide, thiolate, thioether, porphyrin, and others are chosen as ligands. Indeed, the ligands also play an important role in the calibration process, which will be discussed in detail in the Results and Discussion section. The calibration set consists of 387 metallobiomolecules, holding 30 molecules containing a V atom, 20 molecules containing a Cr atom, 50 molecules containing a Mn atom, 80 molecules containing an Fe atom, 60 molecules containing a Co atom, 47 molecules containing a Ni atom, 56 molecules containing a Cu atom, and 64 molecules containing a Zn atom. The frame structures of some metallobiomolecules used in the training set are available in part (I) of the Supporting Information. These molecules should ensure the calibration process and the chemical relevance.

Quantum Chemical Calculations. The organic ligand molecules like amido acid were optimized by the DFT (B3LYP/6-31G*) method. The initial structures of metallobiomolecules were taken from PDB and CCDC. Hydrogen atoms were added with the GaussView program of the Gaussian 03 package. In order to keep up the experimental structure, the heavy atoms of the crystal structures were not optimized, and the hydrogen atoms were locally optimized at the B3LYP/6-31G* level, using the Gaussian 03 program.²⁸ At the same time, the B3LYP/6-31G* method was used to calculate the charge distribution via Mulliken population analysis in this study for both the calibration training set and the metallobiomolecules for check.

Calibration of Parameters. The ABEEM parameters (χ^* and $2\eta^*$) related to C, H, O, and N atoms are mainly based on the previous studies by Yang et al.^{13–16,20,21} The metal–ligand bond is dependent on the detailed nature of the valence state or orbitals of the ligands as well as the effective nuclear charge, coordination numbers, and geometry of the metal ion. In order to deal simply with the metal–ligand bond charges, we assume that there is only one bond charge between a ligand and a metal ion. The method for the bond charge allocation between the metal ion and the coordinated atom of the ligand is the same as the σ bond allocation between them. The detail of this method can be found in refs 13–16. Mulliken charges of the molecules in the training set were calculated by the B3LYP/6-31G* method and then were brought into eqs 1 and 2 in order to determine the ABEEM parameters by the least-square-root algorithm.

Results and Discussion

Calibrated Parameters. The calibration of the ABEEM parameters for transition metals and their bonds proves to be a highly cumbersome task. Each additional element and

Table 1. Defined Atom and Bond Types and the Optimized Values of Parameters of Valence Electronegativity χ^* and Valence Hardness $2\eta^*$ in the ABEEM^a

code ^b	atom and bond type	description ^c	χ^*	$2\eta^*$
2303	V ³⁺	V ³⁺ coordinates O, N and/or S atom in ligands	8.76	1.88
2305	V ⁵⁺	V ⁵⁺ coordinates O and/or S atom in ligands	13.81	2.15
2307	V ⁵⁺	V ⁵⁺ coordinates N atom in ligands	13.81	2.11
2403	Cr ³⁺	Cr ³⁺ coordinates O and/or S atom in ligands	9.33	2.03
2407	Cr ³⁺	Cr ³⁺ coordinates N atom in ligands	9.33	2.20
2406	Cr ⁶⁺	Cr ⁶⁺ coordinates O, N and/or S atom in ligands	15.30	1.93
2502	Mn ²⁺	Mn ²⁺ coordinates O and/or N atom in ligands	7.81	3.08
2515	Mn ²⁺	Mn ²⁺ coordinates O atom in phosphoric ligands	7.96	2.84
2503	Mn ³⁺	Mn ³⁺ coordinates O and/or N atom in ligands	10.57	3.02
2602	Fe ²⁺	Fe ²⁺ coordinates O, N and/or S atom in ligands	7.86	3.05
2603	Fe ³⁺	Fe ³⁺ coordinates O, N and/or S atom in ligands	11.50	3.50
2612	Fe ³⁺	Fe ³⁺ coordinates O atom in carbonic ligands	12.30	3.65
2616	Fe ³⁺	Fe ³⁺ coordinates S atom in Cys ligands	12.30	3.65
2702	Co ²⁺	Co ²⁺ coordinates O, N and/or S atom in ligands	8.06	3.01
2716	Co ²⁺	Co ²⁺ coordinates S atom in Cys ligands	8.11	2.72
2703	Co ³⁺	Co ³⁺ coordinates N atom in ligands	13.15	4.05
2802	Ni ²⁺	Ni ²⁺ coordinates O, N and/or S atom in ligands	8.40	3.14
2901	Cu ⁺	Cu ⁺ coordinates O and/or N atom in ligands	5.66	2.91
2902	Cu ²⁺	Cu ²⁺ coordinates O and/or N atom in ligands	8.42	3.45
2916	Cu ²⁺	Cu ²⁺ coordinates S atom in Cys ligands	8.46	3.13
3002	Zn ²⁺	Zn ²⁺ coordinates O and/or N atom in ligands	8.60	3.54
3016	Zn ²⁺	Zn ²⁺ coordinate S atom in Cys ligands	8.66	3.20
8125	M–O	V (Cr, Mn, Fe, Co, Ni, Cu, and Zn)-O single bond	4.31	25.49
7125	M–N	V (Cr, Mn, Fe, Co, Ni, Cu, and Zn)-N single bond	3.81	16.94
1625	M–S	V (Cr, Mn, Fe, Co, Ni, Cu, and Zn)-S single bond	5.21	35.03

^a The unit of χ^* is Pauling unit; the unit of $2\eta^*$ is Pauling/electron. ^b "Code" denotes the label defined in the the ABEEM program to identify the atom or bond type. ^c In this description, taking V⁵⁺ as an example, "V⁵⁺ coordinating by N atom in ligands" stands for V⁵⁺ coordinating by N atom not including O and/or S atoms in ligands. If V⁵⁺ ion coordinates by N and S and/or O atoms, then the atom type can be adopted by the atom type which coordinates by S and/or O atoms. As for the other metal atom types, the meanings are analogous to those mentioned above.

different atom type would require new ABEEM parameters. This is also due to the sensitivity of the fitness function for the parameters. The new defined atom and bond types as well as the calibrated valence state electronegativity and hardness of the ABEEM parameters related to the transition metals are listed in Table 1. The unlisted ABEEM parameters and their atom type codes related to this paper are available in parts (II) and (III) of the Supporting Information. Besides, from our experience of calibration, it is known that the geometrical parameters like bond stretching, angle bending, and others, in practice, only have very small effects on the charges, and they mainly determine the geometries.

In the calibration process, there are two main factors that influence the ABEEM parameters. On one hand, the chemical surrounding should be taken into account. It is very important for metal atoms to be coordinated by different atoms. For example, Cu and Zn atoms coordinating by an S atom in cysteine have different properties from coordinating by other atoms, such as N and O. In this paper, atom types can be also classified mainly according to these ligands atoms that form different geometric environments, like tetrahedral, square planer, octahedral, etc. For instance, the Fe(II) atom is six-coordinate with three His residues, two Asp residues, and a hydroxyl in the hemerythrin compound. The iron ion coordinates with trigonal-bipyramidal geometry by three His residues, an Asp residue, and solvent molecules in an Fe-containing superoxide dismutase. The iron ion coordinates with a square plane by nitrogen atoms of porphyrin in iron-

Table 2. Absolute Electronegativity (χ) and Hardness (η) Parameters for Some Metal Ions^a

metal ions	Cu ⁺	Mn ²⁺	Fe ²⁺	Ni ²⁺	Cu ²⁺	Zn ²⁺	Fe ³⁺
χ	14.0	24.4	23.4	26.7	28.6	28.8	43.7
η	6.3	9.3	7.3	8.5	8.3	10.8	13.1

^a All values are in eV.

containing protoporphyrin. In our calibration, hemerythrin, superoxide dismutase, and protoporphyrin have been chosen as the model molecules. Obviously, a different training set has a somewhat different effect on the parameters. But the effect is very limited if the training set contains a sufficient number of training molecules and atom types. On the other hand, how to obtain atomic charges to calibrate electronegativity and hardness is also expected to play an important role, because charge distributions are strongly dependent on the choice of basis sets. However, the atomic charges from a different choice of basis sets have a similar trend and correlate to each other nearly linearly. The B3LYP/6-31G* method is used for calibrations in this paper. The HF/STO-3G method was used in some previous calibrations for largely reducing the huge computational work.^{5,13–16}

In order to compare the valence state electronegativity values of the transition metals, Table 2 lists the so-called absolute electronegativity and hardness that were calculated from the experimental ionization potentials and electron affinities by Pearson^{2,29} for some metal ions. A correlation of the absolute electronegativity and hardness with our

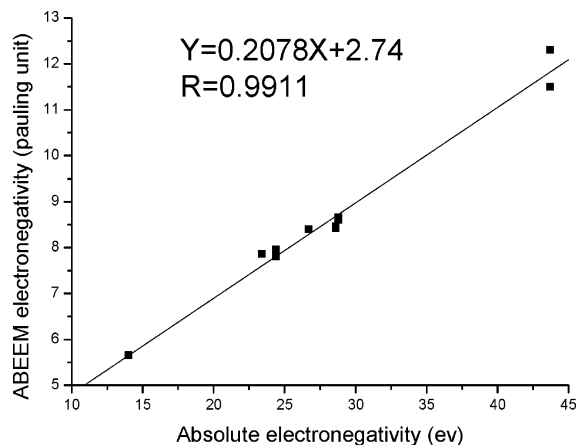


Figure 1. The correlation diagram between the absolute electronegativity parameters and the ABEEM parameters χ^* for some metal ions.

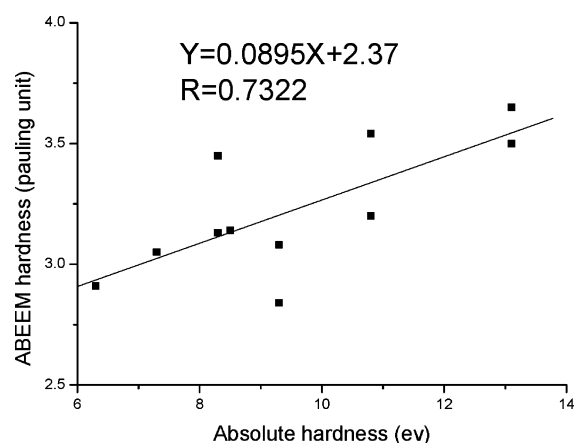


Figure 2. The correlation diagram between the absolute hardness parameters and the ABEEM parameters $2\eta^*$ for some metal ions.

parameters is shown in Figures 1 and 2, respectively. It is noted that the ABEEM parameter values from the current calibration have the same trend with those from the experimental quantities. The same trend and the good correlation also show our ABEEM parameters are reasonable.

Quality of the ABEEM Atomic Charges. The quality of the ABEEM atomic charges is assessed by comparison

with the DFT charges for the training molecules. Figure 3 gives the ABEEM atomic charge distributions versus the B3LYP/6-31G* atomic charge distributions for the training molecules that contain Mn ion. Also included are the correlation constants for the best fitting linear function between both types of the charges in these figures. In fact, the ABEEM charge distributions versus the B3LYP/6-31G* charge distributions for the Mn ion in Figure 3(b) are only a magnification of a small piece of that of all atoms, including H, C, N, O, and so on, in the Mn training molecules in Figure 3(a). In the same way, a comparison of ABEEM and DFT charge distributions about Fe training molecules is shown in Figure 4. And the correlation diagrams for other ions are available in part (IV) of the Supporting Information.

The linear correlation coefficients R containing all atoms over 0.99 are in Figures 3 and 4. It is obvious that the ABEEM charges can well reproduce the DFT charges. But there are also some deviations for Mn and Fe ions between the ABEEM charges and the DFT charges. These deviations may be greatly reduced considering the detailed atom types and the coordination numbers (geometric environments) as well as the multiplicities in different metallobiomolecules. It is well-known that Fe is a very important element in these biomolecules like protoporphyrin, superoxide dismutase, hemerythrin, and so on. For reducing the number of the parameters, however, only one atom type for Fe(II) is assumed in this paper. So the ABEEM charges show some deviations from DFT charges in some compounds. For example, the ABEEM charge of Fe(II) is 1.23 e , and the DFT charge is 1.08 e in the hemerythrin compound (PDB code 1HMD). But, if only one metallobiomolecule is considered, then the trend of the ABEEM charges is in fair accordance with the trend of the DFT charges. In order to obtain better fitting and to make this kind of deviation decrease for metallobiomolecules, adding more atom types that can reflect more complicated geometric environments may be required.

Anyway, in this kind of correlation, the metal ions are in various molecules and their positive charges are quite large. The fact that the ABEEM adequately predicts atomic charges for the transition metals, using the same parameters and fewer

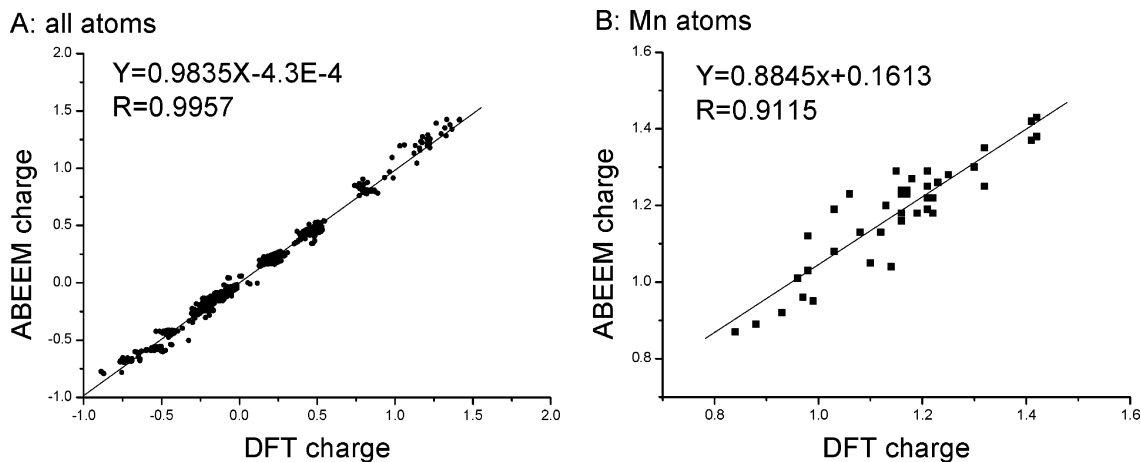


Figure 3. Comparison of the ABEEM and the DFT charge distributions for Mn training molecules.

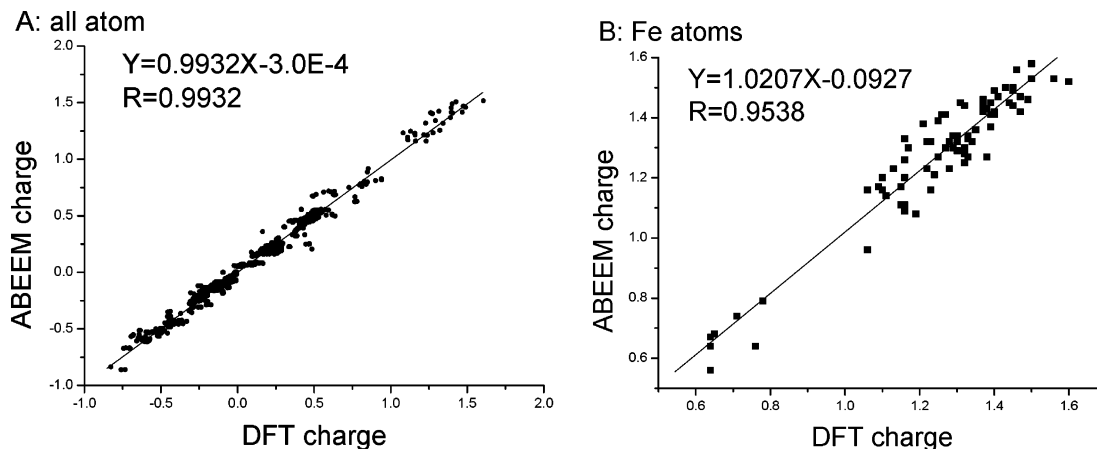


Figure 4. Comparison of the ABEEM and the DFT charge distributions for Fe training molecules.

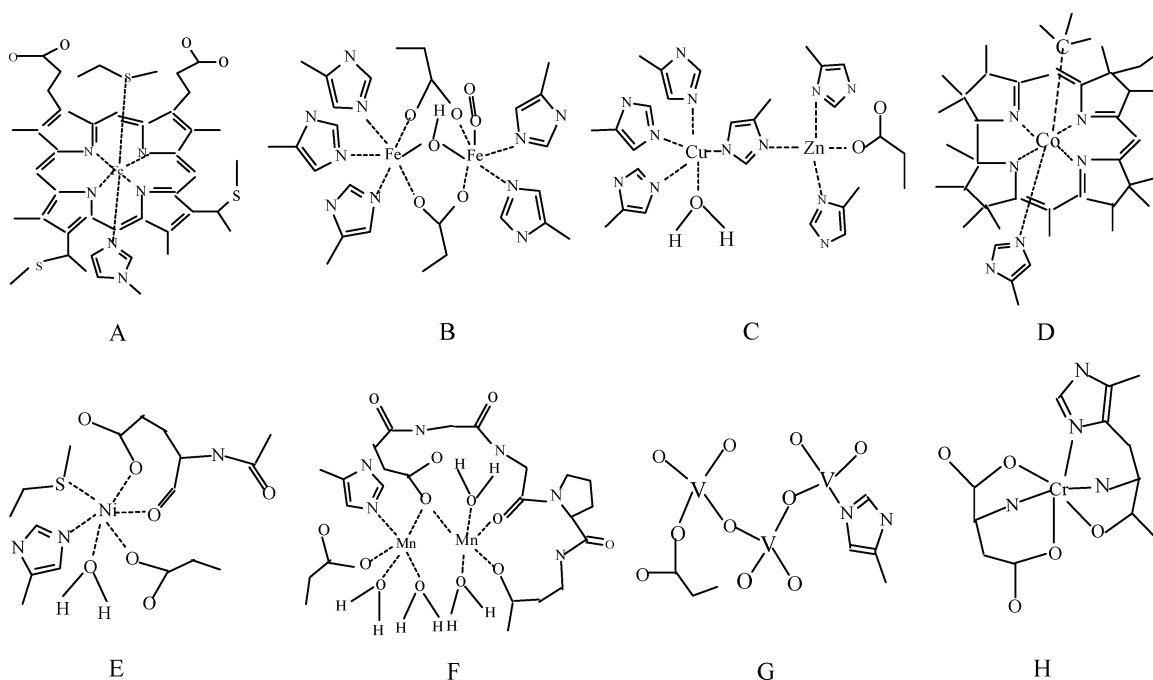


Figure 5. The frame structures of eight metallobiomolecules: A = $C_{43}H_{52}FeN_6O_4S_3$, B = $C_{26}H_{14}Fe_2N_{10}O_7$, C = $C_{27}H_{42}CuN_{12}O_3$ -Zn, D = $C_{40}H_{62}CoN_6$, E = $C_{19}H_{36}N_4NiO_7S$, F = $C_{26}H_{48}Mn_2N_7O_{15}$, G = $C_7H_{11}N_2O_{10}V_3$, H = $C_{10}H_{13}CrN_4O_6$.

Table 3. Correlation Equations of the ABEEM Charge Distribution Y versus the DFT (B3LYP/6-31G*) Charge Distribution X^a

molecules (PDB code ^b and total charge)	$Y = AX + B$	R	S	U
$C_{43}H_{52}FeN_6O_4S_3$ (1CXC, -2)	$Y = 0.9982X - 1.2E-05$	0.9915	0.0270	0.1086
$C_{26}H_{14}Fe_2N_{10}O_7$ (1HMO, +1)	$Y = 0.9739X + 2.3E-04$	0.9948	0.0262	0.1396
$C_{27}H_{42}CuN_{12}O_3Zn$ (2SOD, +2)	$Y = 0.9741X - 1.1E-03$	0.9932	0.0353	0.1220
$C_{40}H_{62}CoN_6$ (1BMT, 0)	$Y = 1.0049X - 8.8E-09$	0.9862	0.0322	0.1334
$C_{19}H_{36}N_4NiO_7S$ (2TDX, 0)	$Y = 0.9980X + 2.0E-06$	0.9909	0.0354	0.1447
$C_{26}H_{48}Mn_2N_7O_{15}$ (1DQ6, +2)	$Y = 1.0106X - 3.2E-04$	0.9928	0.0373	0.1735
$C_7H_{11}N_2O_{10}V_3$ (1H2F, -2)	$Y = 0.9728X - 7.6E-04$	0.9930	0.0435	0.1481
$C_{10}H_{13}CrN_4O_6$ (-2)	$Y = 1.0153X + 5.6E-06$	0.9951	0.0309	0.1466

^a $Y = AX + B$. R being the correlation coefficient, S being the standard error, and U being the maximum error. ^b The framework structures of $C_{10}H_{13}CrN_4O_6$ can be found in ref 30.

atom types, shows the calibrated parameters are applicable within a large range of metallobiomolecules.

Applicability of the ABEEM. In order to check these calibrated parameters, we applied the optimal atomic electronegativity and hardness parameters to calculate charge distributions of some large metallobiomolecules that do not

belong to the calibration set. These metallobiomolecules include haloperoxidase, hemoglobin, hemocyanin, hemerythrin, vitamin B₁₂, carboxypeptidase, carbonic anhydrase, and so on. Most of their framework structures are also directly taken from PDB. Eight of these metallobiomolecule framework structures are shown in Figure 5. These metal-

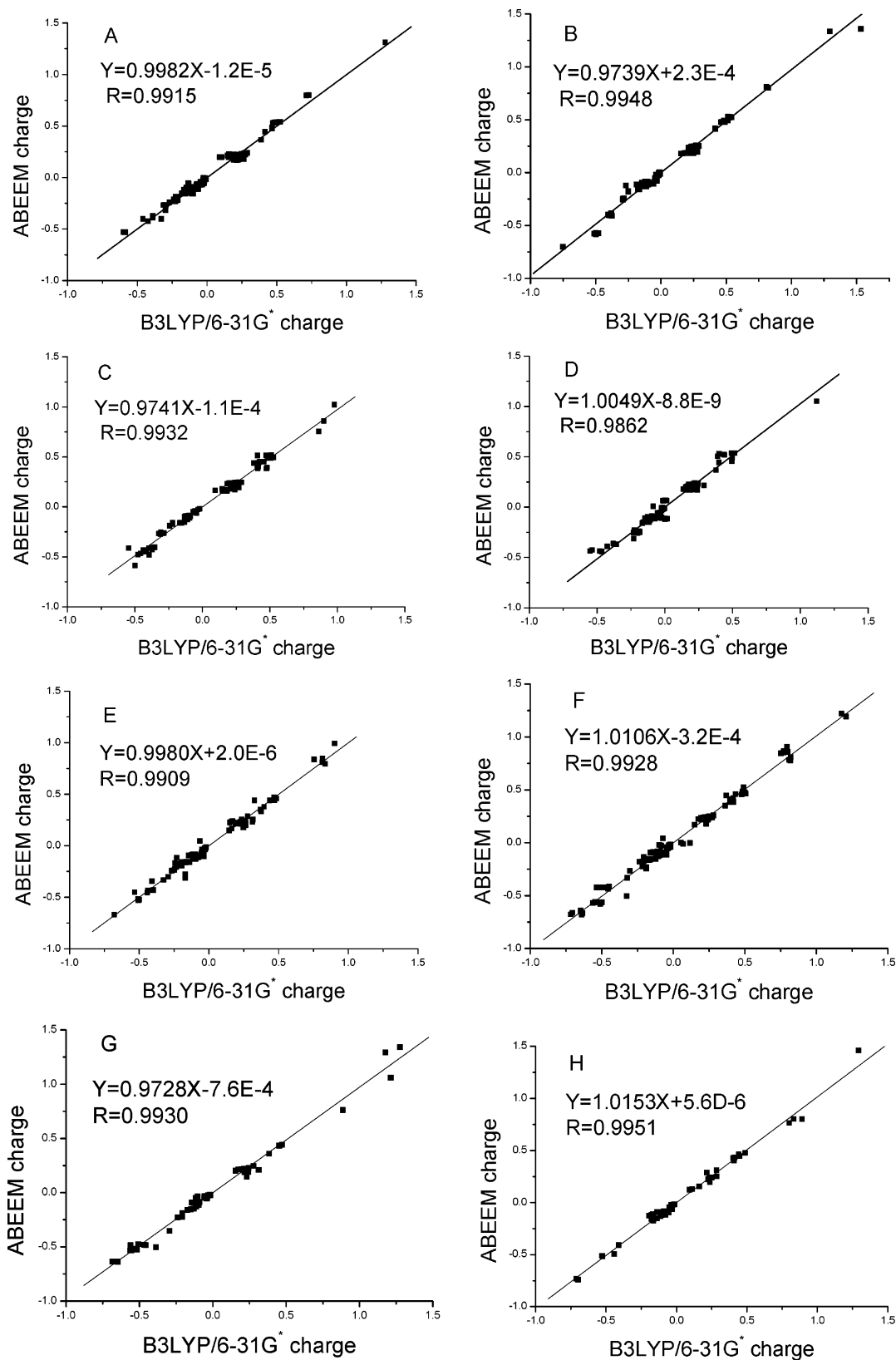


Figure 6. Correlation of the ABEEM and the B3LYP/6-31G* charge distributions for metallobiomolecules: A = $C_{43}H_{52}FeN_6O_4S_3$, B = $C_{26}H_{14}Fe_2N_{10}O_7$, C = $C_{27}H_{42}CuN_{12}O_3Zn$, D = $C_{40}H_{62}CoN_6$, E = $C_{19}H_{36}N_4NiO_7S$, F = $C_{26}H_{48}Mn_2N_7O_{15}$, G = $C_7H_{11}N_2O_{10}V_3$, and H = $C_{10}H_{13}CrN_4O_6$.

lobiomolecules are quite complicated in their structures. The correlation equations about the ABEEM charge distribution versus the DFT (B3LYP/6-31G*) charge distribution for the eight metallobiomolecules are listed in Table 3. The linear

correlation coefficients R are all over 0.98, the standard errors S are smaller than 0.045, and the maximum errors U are smaller than 0.18 that mainly relate to metal ions that have the largest charges. At the same time, Figure 6 gives the

schematic diagrams of the ABEEM charges versus the DFT (B3LYP/6-31G*) charges of these metallobiomolecules. Remarkably, these diagrams show that the ABEEM method can well reproduce the DFT charges. This means that the calibrated ABEEM parameters involving the transition metals of the fourth row in the periodic table of the elements are applicable to more metallobiomolecules.

Atomic charges are very important indicators. Fast calculation of atomic charges for a large molecule or for a large number of molecules could be very useful. For a comparison, speed of the ABEEM calculation of atomic charges is very fast, about 2000 times faster than the usual Mulliken population analysis in the usual SCF MO procedure. Thus, charge calculations with high accuracy and high speed for large metallobiomolecules become an outstanding advantage of the ABEEM method. Once the configuration of a molecular system is determined, its atomic charge distributions will be calculated quickly.

Summary

We have extended the ABEEM method to involve the transition metals. The ABEEM parameters, the valence state electronegativity and hardness, of the transition metals such as V, Cr, Mn, Fe, Co, Ni, Cu, and Zn were calibrated. Although only a few transition-metal atom types are involved, the quality of the ABEEM charges of transition metals and other atoms in the training molecules is good. For the investigated metallobiomolecules, the charge distributions obtained by the ABEEM method are in fair correlation with those obtained from the B3LYP/6-31G* method. The linear correlation coefficients R of the charge distributions for more than 300 training molecules and the above-mentioned eight large metallobiomolecules are all over 0.98. The present study shows the calibrated parameters are reasonable and applicable, and the ABEEM method can predict and calculate charge distribution of metallobiomolecules with high accuracy and high speed. At present, more applications of the ABEEM method in this respect are in progress.

Acknowledgment. This research has been aided by a major grant from the National Natural Science Foundation of China (No. 20633050).

Supporting Information Available: Frame structures of some model molecules (I) (Figure S1); some optimized values of parameters of χ^* and $2\eta^*$ in the ABEEM (II) (Table S1); atom type codes in ligands (III) (Figure S2); and comparisons of the ABEEM and the DFT charge distributions for V, Cr, Co, Ni, Cu, and Zn training molecules (IV) (Figure S3). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Parr, R. G.; Donnelly, R. A.; Levy, M.; Palke, W. E. Electronegativity: The density functional viewpoint. *J. Chem. Phys.* **1978**, *68*, 3801.
- (2) Parr, R. G.; Pearson, R. G. Absolute Hardness: Companion Parameter to Absolute Electronegativity. *J. Am. Chem. Soc.* **1983**, *105*, 7512.
- (3) Parr, R. G.; Yang, W. *Density Functional Theory of Atoms and Molecules*; Oxford University Press and Clarendon Press: New York and Oxford, 1989.
- (4) Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual Density Functional Theory. *Chem. Rev.* **2003**, *103*, 1793.
- (5) Mortier, W. J.; Ghosh, S. K.; Shankar, S. Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules. *J. Am. Chem. Soc.* **1986**, *108*, 4315.
- (6) Nalewajski, R. F. Electrostatic effects in interactions between hard (soft) acids and bases. *J. Am. Chem. Soc.* **1984**, *106*, 944.
- (7) Rappé, A. K.; Goddard, W. A., III. Charge equilibration for molecular dynamic simulations. *J. Phys. Chem.* **1991**, *95*, 3358.
- (8) Cioslowski, J.; Martinov, M. Electronegativity Equalization in Polyene Carbon Chains. *J. Phys. Chem.* **1996**, *100*, 6156.
- (9) York, D. M.; Yang, W. A chemical potential equalization method for molecular simulations. *J. Chem. Phys.* **1996**, *104*, 159.
- (10) De Proft, F.; Langenaeker, W.; Geerlings, P. A non-empirical electronegativity equalization scheme. Theory and applications using isolated atom properties. *J. Mol. Struct.* **1995**, *339*, 45.
- (11) Ghosh, S. K. Electronegativity, hardness, and a semiempirical density functional theory of chemical binding. *Int. J. Quantum Chem.* **1994**, *49*, 239.
- (12) No, K. T.; Grant, J. A.; Scheraga, H. A. Determination of Net Atomic Charges Using a Modified Partial Equalization of Orbital Electronegativity Method. 1. Application to Neutral Molecules as Models for Polypeptides. *J. Phys. Chem.* **1990**, *94*, 4732.
- (13) Yang, Z. Z.; Wang, C. S. Atom-Bond Electronegativity Equalization Method. 1. Calculation of the Charge Distribution in Large Molecules. *J. Phys. Chem. A* **1997**, *101*, 6315.
- (14) Yang, Z. Z.; Wang, C. S. Atom-bond electronegativity equalization method and its applications based on density functional theory. *J. Theory Comput. Chem.* **2003**, *2*, 273.
- (15) Cong, Y.; Yang, Z. Z. General atom-bond electronegativity equalization method and its application in prediction of charge distributions in polypeptide. *Chem. Phys. Lett.* **2000**, *316*, 324.
- (16) Yang, Z. Z.; Xiao, H. Y. Calculation of Charges Distribution in iron(II) complex by Using ABEEM Model. *Chem. J. Chin. Univ.* **2005**, *26*, 117.
- (17) Yang, Z. Z.; Wu, Y.; Zhao, D. X. Atom-bond electronegativity equalization method fused into molecular mechanics. I. A seven-site fluctuating charge and flexible body water potential function for water clusters. *J. Chem. Phys.* **2004**, *120*, 2541.
- (18) Li, X.; Yang, Z. Z. Hydration of Li⁺-ion in atom-bond electronegativity equalization method—7P water: A molecular dynamics simulation study. *J. Chem. Phys.* **2005**, *122*, 084514.
- (19) Yang, Z. Z.; Li, X. Molecular-dynamics simulations of alkaline-earth metal cations in water by atom-bond electronegativity equalization method fused into molecular mechanics. *J. Chem. Phys.* **2005**, *123*, 094507.
- (20) Zhang, Q.; Yang, Z. Z. An investigation of alkane conformations based on the ABEEM/MM model. *Chem. Phys. Lett.* **2005**, *403*, 242.

- (21) Yang, Z. Z.; Zhang, Q. Study of Peptide Conformation in Terms of the ABEEM/MM Method. *J. Comput. Chem.* **2006**, *27*, 1.
- (22) Yang, Z. Z.; Qian, P. A study of *N*-methylacetamide in water clusters: Based on atom-bond electronegativity equalization method fused into molecular mechanics. *J. Chem. Phys.* **2006**, *125*, 064311.
- (23) Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Waroquier, M.; Tollenaere, J. P. The Electronegativity Equalization Method I: Parametrization and Validation for Atomic Charge Calculations. *J. Phys. Chem. A* **2002**, *106*, 7887.
- (24) Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Van, Alsenoy, C.; Tollenaere, J. P. The Electronegativity Equalization Method II: Applicability of Different Atomic Charge Schemes. *J. Phys. Chem. A* **2002**, *106*, 7895.
- (25) Holm, R. H.; Kennepohl, P.; Solomon, E. I. Structural and Functional Aspects of Metal Sites in Biology. *Chem. Rev.* **1996**, *96*, 2239.
- (26) Vincent, J. B. Elucidating a Biological Role for Chromium at a Molecular Level. *Acc. Chem. Res.* **2000**, *33*, 503.
- (27) Sanderson, R. T. An Interpretation of Bond lengths and a Classification of Bonds. *Science* **1951**, *114*, 670.
- (28) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, Jr., T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision D.01*; Gaussian, Inc.: Wallingford, CT, 2004.
- (29) Pearson, R. G. Chemical hardness and bond dissociation energies. *J. Am. Chem. Soc.* **1988**, *110*, 7684.
- (30) Madafoglio, B. K.; Manning, T. M.; Murdoch, C. M.; Tulip, W.; Cooper, M. K.; Hambley, T. W.; Freeman, H. C. Three Chromium(III) Complexes with Mixed Amino Acid Ligands: (L-Cysteinato)-(L-histidinato) chromium(III) 3''5-Hydrate, (L-Aspartato)(L-histidinato) chromium(III) 1-5-Hydrate and Bis(DL-histidinato) chromium(III) Chloride 4''2-Hydrate. *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.* **1990**, *C46*, 554.

CT600379N

Molecular Dynamics of Organophosphorous Hydrolases Bound to the Nerve Agent Soman

Thereza A. Soares,[†] Mohamed A. Osman,[‡] and T. P. Straatsma^{*†}

Pacific Northwest National Laboratory, 902 Battelle Blvd., P.O. Box 999 MSIN K7-90, Richland, Washington 99352, and School of Electrical Engineering and Computer Science, Washington State University, Pullman, Washington 99164

Received January 25, 2007

Abstract: The organophosphorous hydrolase (OPH) from *Pseudomonas diminuta* is capable of degrading extremely toxic organophosphorous compounds with a high catalytic turnover and broad substrate specificity. Although the natural substrate for OPH is unknown, its triple-mutant H254G/H257W/L303T exhibits a 3 order of magnitude increase in catalytic efficiency and modified stereospecificity toward the most toxic *SpSc* enantiomer of soman. Molecular dynamics simulations and binding free-energy calculations have been undertaken for the wild-type and triple-mutant H254G/H257W/L303T enzymes bound to the *SpSc*-soman enantiomer. Comparison of the simulations indicates that substrate binding induces conformational changes of the loops near the active site. The coordination of the zinc cations in the active site of OPH differs between the free enzyme and the complexes. This suggests that the active site of OPH can accommodate several catalytically active coordination geometries, consistent with the fact that the enzymatic activity of the wild-type OPH can be enhanced by alterations to the metal content of the enzyme. It is also argued that the enhanced efficiency of the triple mutant is determined by enzyme-transition-state complementarity. These results provide a qualitative, molecular-level explanation for the 3 order of magnitude increase in catalytic efficiency of the triple-mutant toward *SpSc*-soman.

Introduction

Organophosphates are extremely toxic chemicals produced by the reaction of alcohols and phosphoric acid. Their biological effect is the inactivation of acetylcholinesterase, which results in the accumulation of the neurotransmitter acetylcholine in the synaptic cleft. In excess, acetylcholine overactivates the postsynaptic receptors and decreases the rate of signal transmission of neurons. Organophosphates are exclusively synthetic compounds that were first developed as insecticides in the 1930s. However, their toxic properties were rapidly identified and further developed as nerve agents during World War II.¹ Soman, also known by its NATO designation GD (O-pinacolyl methylphosphonofluoridate) was the third of the so-called G-series nerve agents to be synthesized [along with GA (tabun), GB (sarin), and GF (cyclosarin)]. The median lethal dose, LCt50, is 70 mg min m⁻³ in humans, and its sole application is as a military weapon.^{2–4}

The bacterial enzyme organophosphorous hydrolase (OPH) has been shown to catalyze the cleavage of P–O, P–F, and P–S bonds in a variety of organophosphate triesters and related phosphonates with a high catalytic turnover and broad substrate specificity.^{5–11} Although its natural substrate is unknown, OPH exhibits a turnover of 10⁴ s⁻¹ for the best substrates, while the corresponding values for k_{cat}/K_M approach the diffusion limit of 10⁸ M⁻¹ s⁻¹.^{12,5} In addition, OPH exhibits stereoselectivity for the hydrolysis of chiral organophosphate triesters, and mutant forms have been engineered with enhanced catalytic activity toward the most toxic stereoisomers of analogs of sarin, soman, and VX.^{7,8,10,11} Among them, the triple-mutant H254G/H257W/L303T exhibits a 3 order of magnitude increase in catalytic efficiency and modified stereospecificity toward the most toxic *SpSc* enantiomer of soman.^{13,14} As a result, catalytically enhanced OPH mutants have a potential application as biosensors for this class of nerve agents, demonstrated by experimental studies in which the immobilization of OPH in nanopores led to enhanced stability and catalytic reaction rates compared to the free enzyme in solution.¹⁵

* Corresponding author. E-mail: tps@pnl.gov.

[†] Pacific Northwest National Laboratory.

[‡] Washington State University.

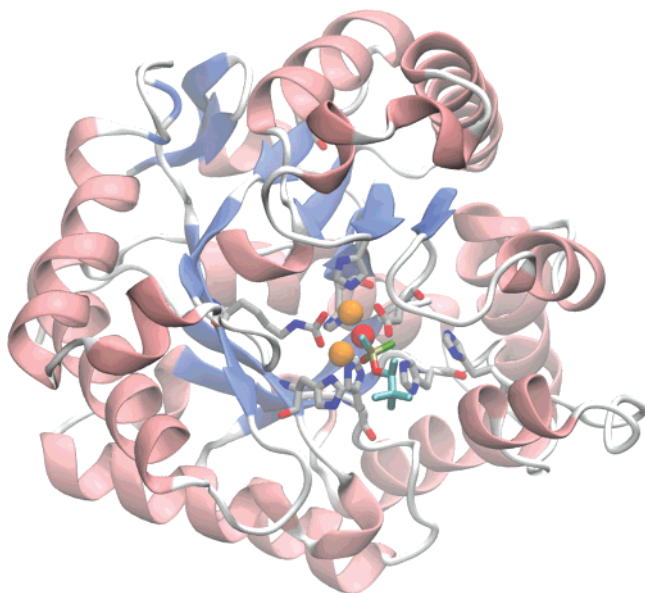


Figure 1. Cartoon representation of the OPH monomer. Active site residues (carbon atoms in gray) and *SpSc*-soman (carbon atoms in cyan) are represented as sticks. Water oxygen atoms (red) and Zn^{2+} cations (orange) are represented in CPK.

X-ray structures have been obtained for the wild-type and mutant forms of OPH complexes with different substrate analogs.^{16–19} These structures reveal a homodimeric (α/β)₈-barrel containing an active site with two divalent metal ions coordinated to the protein through interactions with four histidine and one aspartate residue. The two metal ions are bridged by a water molecule and a carbomoylated lysine residue (Figure 1). Although Zn^{2+} is the native metal ion, substantial activity is observed after substitution by Co^{2+} , Cd^{2+} , Mn^{2+} , or Ni^{2+} .^{5,20} The kinetic constants, k_{cat} and $k_{\text{cat}}/K_{\text{M}}$, are dependent upon the identity of the specific metal cations within the active site. A single-step mechanism has been proposed where the bridging solvent molecule is activated for an in-line $\text{S}_{\text{N}}2$ -nucleophilic attack via complexation to the binuclear metal center and a hydrogen-bonding interaction with residue Asp301.^{8,21}

Enzymatic catalysis is an inherently dynamic process²² in which the binding and release of ligands are often accompanied by conformational changes that may involve large-scale structural rearrangements or local fluctuations in atomic positions.^{23–26} In particular for OPH, kinetic studies have shown that the rate-limiting step of enzymatic hydrolysis appears to involve a conformational change or diffusion-controlled dissociation.⁶ Previous molecular dynamics (MD) simulations suggest that OPH undergoes a substantial conformational change, inducing the opening of a gateway in a pocket where the location of the substrate-leaving group is expected.²⁷ During this rearrangement, Tyr309 was proposed to assist the leaving group, as it exits from the hydrophobic pocket. However, mutational studies of the mutant Y309F failed to find any significant difference in the magnitude of either k_{cat} or $k_{\text{cat}}/K_{\text{M}}$ when compared to those of the wild-type enzyme for the hydrolysis of either paraoxon or a sarin analog.²¹

In order to characterize the contribution of local and global motions to the enhanced activity of the OPH triple-mutant

H254G/H257W/L303T toward *SpSc*-soman, molecular dynamics simulations were performed for OPH wild-type and the complexes of the wild-type/triple-mutant and the substrate *SpSc*-soman. These simulations indicate that substrate binding induces conformational changes of the loops near the active site and that the coordination geometry of the zinc cations in the active site of the enzyme differs between the free enzyme and the complexes. In addition, binding energies of association between wild-type/triple-mutant and the substrate *SpSc*-soman were calculated from a thermodynamic cycle based on continuum electrostatics and a surface-area-dependent nonpolar term. The comparison of calculated and experimentally derived binding energies provides a rationale for the enhanced activity of the triple-mutant enzyme.

Methods

Molecular Systems. MD simulations in explicit solvent were performed for the wild-type (OPH_{wtc}) and triple-mutant (OPH_{tmc}) monomers of the enzyme organophosphorous hydrolase bound to soman and for the unbound wild-type homodimer (OPH_{wt}). Each monomer contains a structurally and catalytically independent active site,^{16–19} and therefore the structural dynamics of the monomers and dimers are expected to be comparable. Initial coordinates were taken from crystallographic structures with PDB entry codes 1EZ2 and 1P6C at 1.9 and 2.0 Å resolution, respectively.¹³ The cocrystallized diisopropylmethyl phosphonate analog present in both structures was replaced by the *SpSc*-soman enantiomer. Amino acid protonation states were assigned accordingly to a pH of 7.0, and hydrogen atoms were generated using the Prepare module of NWChem.²⁸ The zinc ions were treated using a nonbonded model with a formal charge of +2.²⁹ The geometries of the soman and carbamylated Lys169 were fully optimized by using density functional theory, the B3LYP functional, and the DZVP basis set.³⁰ Partial atomic charges for soman and the carbamylated Lys169 were calculated at the Hartree–Fock level with the 6-31G* basis set and restrained electrostatic potential procedure³¹ on the geometry-optimized structures. These charges were used in combination with the AMBER95 force field parameters.³² The structures were solvated in a cubic box with dimensions of 7.4, 7.6, and 10.0 nm³ for OPH_{wtc} , OPH_{tmc} , and OPH_{wt} , respectively. Water molecules within 0.28 nm of any atom in the solute were removed. Periodic boundary conditions and the SPC/E water model³³ were used to describe the solvent molecular interactions.

Molecular Dynamics Simulations. MD simulations were carried out for the NPT ensemble with a time step of 1 fs during the equilibration and 2 fs during the production runs. The temperatures of the solute and solvent were controlled by separately coupling them to a Berendsen thermostat³⁴ with a relaxation time of 0.1 ps. The pressure was maintained at 1.025×10^5 Pa by means of isotropic coordinate scaling with a relaxation time $t = 0.4$ ps. A time step of 2 fs was used to integrate the equations of motion on the basis of the leapfrog algorithm.³⁵ The bond lengths between hydrogen and heavy atoms were constrained by using the SHAKE algorithm³⁶ with a tolerance of 10^{-3} nm. A short-range cutoff of 1.0 nm was used for all nonbonded interactions, and long-range electrostatic interactions were treated by the smooth particle mesh Ewald (PME) method.³⁷ The equilibration procedure consisted of thermalization of the solvent, with

the solute atoms fixed, for 20 ps at 298.15 K, followed by minimization of all solute atoms, keeping the solvent coordinates fixed, and then simulation of the complete system by raising the temperature from 0 to 298.15 K in 20 ps increments of 50 K each for MD simulation. Data production was carried out for 5 ns, and configurations of the trajectory were recorded every 0.2 ps. Within the modest simulation times of 5 ns, several structural properties, including backbone root-mean-square-deviation (RMSD), have reached convergence. All simulations were performed with the NWChem program,²⁸ and the analyses of molecular trajectories were carried out using data-intensive trajectory analysis capabilities of the DIANA module.³⁸

Electrostatic Calculations. The free energy of ligand–protein association may be approximated in the form³⁹

$$\Delta G_{\text{bind}} = \Delta G_{\text{elec}} + \Delta G_{\text{np}} - T\Delta S \quad (1)$$

where ΔG_{elec} is the electrostatic contribution, ΔG_{np} is the nonpolar or hydrophobic term, and $-T\Delta S$ describes the change in entropy (conformational, translational, and rotational) upon complexation. The change in flexibility of the enzyme and substrate upon binding is assumed to be similar in the wild type and triple mutant. In this case, the entropic contribution will be the same for the two systems and cancel in the calculation of the relative free energy of binding. The electrostatic term ΔG_{elec} was calculated by solving the linearized Poisson–Boltzmann equation with the Adaptive Poisson–Boltzmann Solver package.⁴⁰ The calculations were performed at 298.15 K and an ionic strength of 0 M. Dielectric constants of 2 and 78 were assigned to the solute and solvent, respectively. The dielectric boundary between the solute and solvent was based on the molecular surface definition calculated with a probe sphere radius of 0.14 nm. For the boundary conditions, the focusing method was applied⁴¹ with multigrid points $65 \times 65 \times 65$, coarse grid lengths $0.33 \times 0.33 \times 0.33 \text{ nm}^3$, and fine grid dimensions $0.16 \times 0.16 \times 0.16 \text{ nm}^3$. The nonpolar term ΔG_{np} is estimated on the basis of the amount of the solvent-accessible surface area buried upon binding

$$\Delta G_{\text{np}} = \gamma\Delta A \quad (2)$$

where ΔA is the change in solvent-accessible surface area upon binding and γ is the apolar constant $0.105 \text{ kJ mol}^{-1} \text{ \AA}^{-2}$. This empirical coefficient (effective microscopic interfacial tension) is calibrated to reproduce experimental transfer free energies of alkane molecules from the liquid alkane phase into water.^{42,43} In order to estimate the relative strength of binding of *SpSc*-soman to the wild-type and triple-mutant forms of OPH, we have applied the above equation to 20 structures sampled at regular time intervals from each of the two MD simulations.

Results and Discussion

Structural Stability and Flexibility of OPH. Atom-positional RMSD and root-mean-square fluctuation (RMSF) were calculated from the MD simulations for OPH_{wtc}, OPH_{tmc}, and OPH_{wt} with respect to the X-ray structures 1EZ2 (wild type) and 1P6C (triple mutant), respectively (Figures 2 and 3). Monomeric and dimeric wild-type ensembles exhibit comparable RMSD values, which converge to 0.1

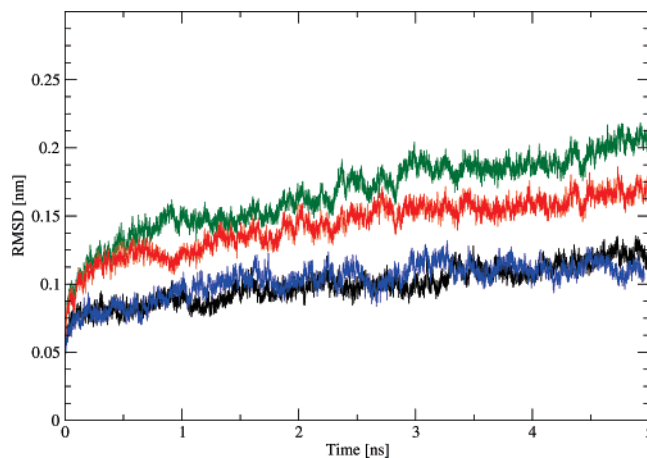


Figure 2. Root-mean-square deviation of the α -carbon atoms of OPH_{wtc} (black), OPH_{tmc} (green), OPH_{tmc} without the loop regions corresponding to residues 260–276 and 305–315 (red), and OPH_{wt} (blue) with respect to the X-ray structures 1EZ2 and 1P6C as described in the text.

nm after a time period of 3 ns. The OPH_{tmc} ensemble exhibits RMSD values of about 0.2 nm. The loop regions corresponding to residues 260–276 and 305–315 are responsible for the slow convergence of the RMSD along the OPH_{tmc} simulation. Convergence is obtained when these residues are excluded from the RMSD calculations (Figure 2). The residues in both loops are involved in large-scale atomic displacement as discussed below. Likewise, atom-position RMSF for the monomeric and dimeric wild-type simulations are very similar (Figure 3). One exception is the region corresponding to residues 60–70, which displays larger flexibility in the OPH_{wtc} and OPH_{tmc} simulations compared to that in OPH_{wt}. The other is the region formed by residues 130–140, which is more flexible in the OPH_{tmc} ensembles than in the X-ray structure 1P6C. These two regions are part of a loop/short-helix/loop motif located away from the active site, at the protein interface between the two monomers. The apparent reason for their increased flexibility is the exposure to the solvent in the monomer simulations. However, the effect of these fluctuations on the overall structure of the monomers is negligible, as shown by the similarity of the secondary structure pattern between the monomeric and dimeric wild-type ensembles of OPH for which only small and localized differences can be observed (Figure 4). Examples of such differences are the β -strand conformation around residues Val198–His201 in OPH_{wtc} and OPH_{tmc} but not in OPH_{wt} and Ile168–Ala171 in OPH_{wt} and OPH_{tmc} but not in OPH_{wtc}. It can thus be assumed that the MD ensemble for the OPH monomers is representative of the ensemble corresponding to their respective dimeric forms.

The RMSF calculated for the MD simulations can be correlated with the isotropic temperature factors (B-factors) from the crystallographic coordinates (Figure 3). The sizes of the B-factors are representative of the amount of disorder present in the crystal on the time scale of the diffraction experiment, which ranges from seconds to days. The divergence between RMSF and B-factors around residues 205 and 240 in the wild-type simulations is likely due to differences in the experiment and simulation time scales or differences between the crystal and solution environments or force-field limitations. Because the solution and crystal

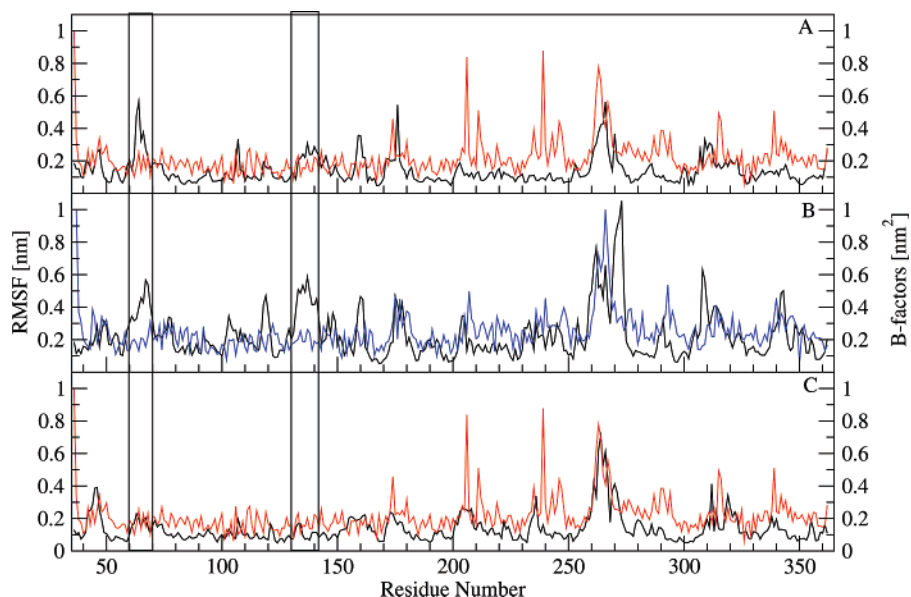


Figure 3. Root-mean-square fluctuations of C_{α} atoms of (A) OPH_{wtc} , (B) OPH_{tmc} , and (C) OPH_{wt} with respect to the X-ray structures 1EZ2 and 1P6C. Isotropic B-factors from the X-ray structures 1EZ2 (red) and 1P6C (blue). The regions highlighted correspond to residues 60–70 and 130–140 discussed in the text.

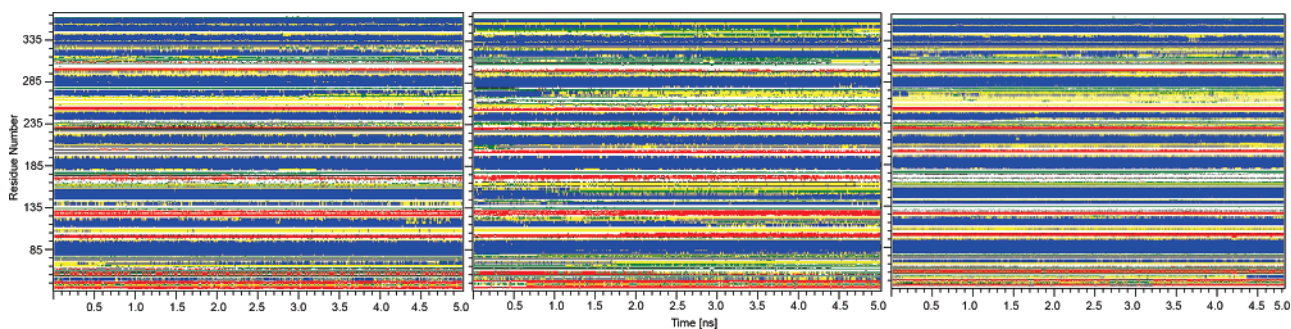


Figure 4. Representation of secondary structure pattern along time for OPH_{wtc} (left), OPH_{tmc} (middle), and OPH_{wt} (right). Color patterns are blue for α -helix, gray for 3-helix, red for β -sheet, black for β -bridge, green for bend, yellow for turn, and white for coil.

environments are different, the comparison of RMSF data from molecular simulations with experimental B-factors should be carried out with caution but can be illustrative to identify common flexible components. In the three simulations, residues 260–270 (260–275 for OPH_{tmc}) exhibit the largest atomic fluctuations. They form a loop/ α -helix motif that, together with seven short loops, delimits the entrance of the active site. In the OPH_{tmc} ensemble, these loops exhibit a discernible atom fluctuation, which is absent in both wild-type ensembles (Figure 3). These results indicate that the mutations H254G, H257W, and L303T alter the internal dynamics of the mutant, leading to an increased flexibility of residues in the entrance of the active site of OPH_{tmc} .

The essential dynamics analysis method was applied to the MD trajectories to separate low-frequency motions that typically determine the kinetics of enzymatic activity from the much larger number of remaining high-frequency motions. This technique is based on a principal component analysis of the positional displacement from an average structure. The separation in low- and high-frequency components is made through a change of coordinate system from Cartesian to these principal components or eigenvectors of the covariance matrix. This method, in principle, allows for the extrapolation of motions in the direction of selected

eigenvectors. In this study, however, this analysis is used to describe the low-frequency, persistent motions observed in the molecular trajectories. The eigenvalue percentages and amplitudes corresponding to the 10 eigenvectors with highest eigenvalues calculated from the OPH_{wtc} , OPH_{tmc} , and OPH_{wt} simulations, respectively, are presented in Table 1. Only a few modes are required to account for the large-amplitude mobility observed in the three ensembles. Most of the large-scale atomic displacement in the OPH_{wtc} and OPH_{tmc} simulations is contained in the first eigenvectors, with eigenvalues of 0.46 and 0.65 nm^2 , respectively. In the OPH_{wt} simulation, the first eigenvector has an eigenvalue of less than 0.23 nm^2 and accounts for 17.9% of the total atom displacement observed in this ensemble.

The contributions of C_{α} atoms to the first and second eigenvectors are presented in Figure 5. They represent the relative displacement of each residue due to the motion described by a given eigenvector. The residues that contribute most to the motions along the first and second eigenvectors in the simulations are confined to the loops in the entrance of the active site. Two regions of the OPH_{wtc} and OPH_{tmc} ensembles exhibit the largest atomic displacements along the first eigenvector: residues 175–180/260–276 and 260–276/305–315, respectively (Figure 5). In the OPH_{wt} ensemble,

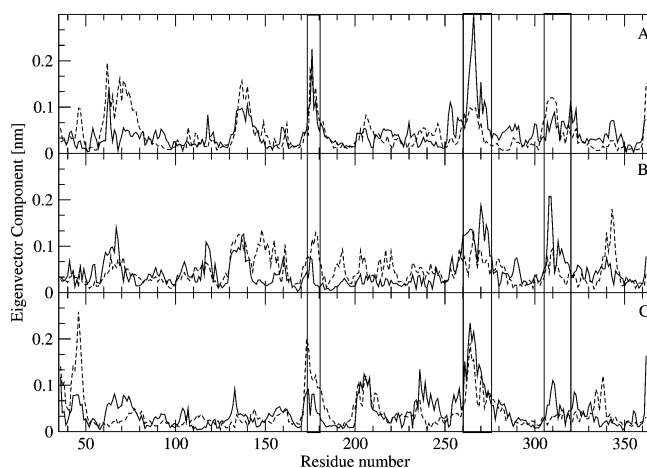


Figure 5. Eigenvector components for atomic displacement along the first (black lines) and second (dashed lines) eigenvectors for the MD-generated ensembles of (A) OPH_{wtc} , (B) OPH_{tmc} , and (C) OPH_{wt} . Residues 175–180 (loop L1), 260–276 (loop L2), and 305–315 (loop L3) are highlighted.

Table 1. Magnitudes of the Eigenvalues Calculated from the Covariance Matrix of α -Carbon Coordinates Corresponding to the MD Simulations OPH_{wtc} , OPH_{tmc} , and OPH_{wt}

eigenvalues	Eigenvalue Magnitudes and Amplitudes		
	magnitude [%]/amplitude [nm ²]		
	OPH_{wtc}	OP_{tmc}	OPH_{wt}
1	30.85/0.460	30.30/0.650	17.91/0.226
2	6.73/0.100	11.16/0.240	12.58/0.159
3	4.61/0.069	5.18/0.110	5.22/0.066
4	3.79/0.057	3.97/0.085	3.32/0.042
5	2.98/0.045	3.15/0.067	2.94/0.037
6	2.43/0.036	2.52/0.056	2.41/0.034
7	1.91/0.029	2.10/0.045	2.04/0.026
8	1.63/0.024	1.82/0.039	1.84/0.023
9	1.53/0.023	1.58/0.034	1.49/0.09
10	1.36/0.020	1.45/0.031	1.45/0.018

where the substrate soman is absent from the active site, only residues 260–276 contribute significantly to the motion along the first eigenvector. These residues, together with residues 43–50/175–180, also present the largest atom displacement in the motion described by the second eigenvector (eigenvalue of ca. 12.6%) in the OPH_{wt} simulation (Table 1). The loop regions formed by residues 175–180, 260–276, and 305–315 will be referred here as L1, L2, and L3, respectively.

Loops L1 and L2 are located opposite each other across the active site entrance, whereas loops L2 and L3 are located side-by-side. In the motion described by the first eigenvector in the OPH_{wtc} simulation, loops L1 and L2 move in opposite directions from the entrance of the active site (see Figure 6A). In the OPH_{tmc} simulation, a similar motion is observed for loops L2 and L3 also with respect to the active site entrance (Figure 6B). In both ensembles, these displacements lead to a widening of the active site entrance. A qualitative estimate of the width of the opening is given by the distance between residues in loops L1, L2, and L3 and residues located across the active site. The C_{α} atom distances between L2 residue Ala270 and Phe132, a residue immediately

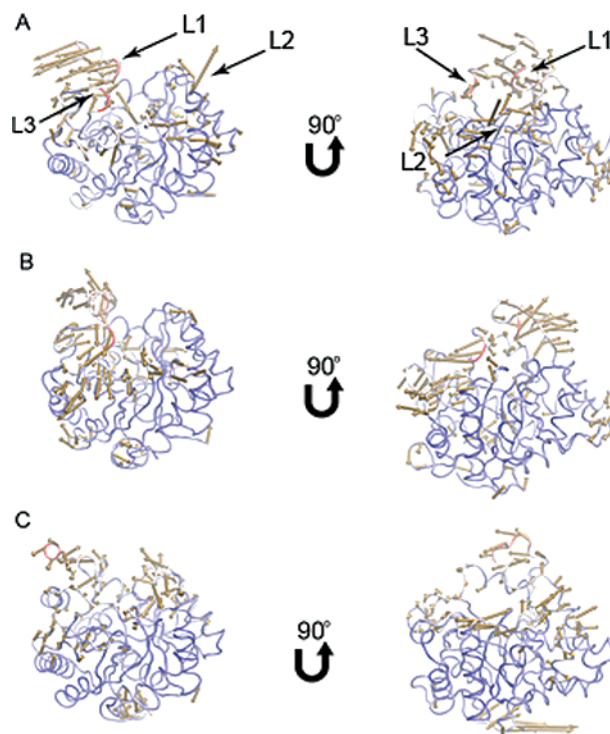


Figure 6. Porcupine plots representing the highest amplitude motions along the first eigenvector for (A) OPH_{wtc} , (B) OPH_{tmc} , and (C) OPH_{wt} . In C, one monomer of the dimer simulation is shown. Only eigenvector components larger than 0.05 ps are shown for clarity. C_{α} trace colored by atom contribution to the first eigenvector. The color gradient is from red (highest displacement) to blue (lowest displacement). The loop regions L1, L2, and L3 correspond to residues 175–180, 260–276, and 305–315, respectively.

opposite from loop L2 in the direction of the first eigenvector, fluctuates between 1.68 and 2.50 nm. The C_{α} atom distances between loop L3 residue Ser308 and residue Pro178 opposite this loop fluctuates between 2.36 and 2.56 nm. Similarly for the OPH_{tmc} ensemble, the C_{α} atom distances between loop L2 residue Ala270 and residue Phe132 is found to be between 1.78 and 2.50 nm, while this distance for loop L3 residue Tyr309 and residue Ser205 fluctuates between 1.72 and 2.32 nm. For OPH_{tmc} , this movement is accompanied by the displacement of Tyr309 toward the active site entrance where its aromatic ring makes hydrophobic contacts with the aliphatic chain of the substrate, reducing solvent access. In the OPH_{wt} ensemble, loop L2 and, to a lesser extent, its neighboring loops display much lower amplitude and less ordered collective displacements along the first eigenvector (Figure 6C). In this ensemble, the distances between the C_{α} atoms of loop L2 residue Ala270 and residue Phe132 fluctuate between 1.68 and 1.79 nm and those of loop L3 residue Ser308 and residue Pro178 fluctuate between 2.35 and 2.43 nm. These collective motions do not result in any significant conformational rearrangement of residues or displacement of Tyr309 toward the active site entrance of OPH_{wt} , and thus these changes appear to be triggered upon substrate binding.

Binding of *SpSc*-Soman to OPH_{wtc} and OPH_{tmc} . The atom-positional RMSD for the *SpSc*-soman bound to OPH_{wtc} and OPH_{tmc} as a function of time is presented in Figure 7.

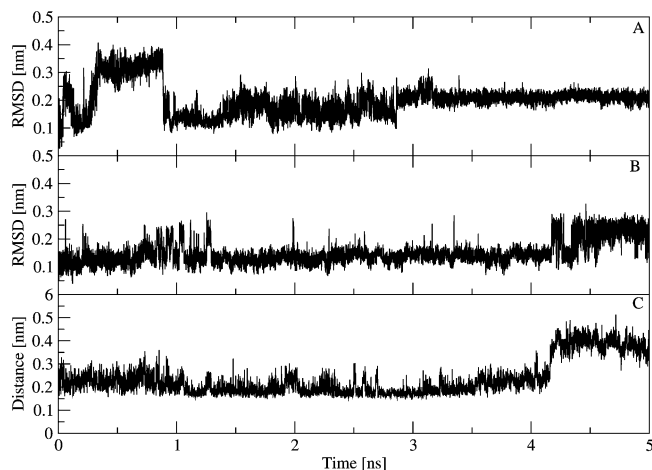


Figure 7. Root-mean-square deviation of the heavy atoms for soman. (A) *SpSc*-soman bound to the wild-type OPH. (B) *SpSc*-soman bound to the triple-mutant OPH. (C) Distance between the phosphoester oxygen atom of *SpSc*-soman and one of the water molecules coordinated to the β -metal in the triple mutant.

In the OPH_{wtc} complex, the substrate undergoes a positional deviation of 0.35 nm from the initial conformation, which subsequently decreases to 0.1–0.15 nm and converges to a plateau around 0.2 nm (Figure 7A). In the OPH_{tmc} complex,

soman exhibits a less fluctuating RMSD pattern around 0.1–0.15 nm during the first 4 ns and then reaches a second plateau around 0.2–0.25 nm (Figure 7B). It also exhibits an average positional fluctuation of 0.21 nm in the OPH_{tmc} complex compared to the corresponding average of 0.25 nm in the OPH_{wtc} complex. This behavior is correlated with a hydrogen bond between the phosphoester oxygen atom of soman and a β -metal-coordinated water molecule that appears to impose a conformational and spatial restraint on the substrate (Figure 7C). This interaction was not observed for the OPH_{wtc} simulation.

The time-dependent behavior of interactions between the nucleophilic water and catalytically important residues was also monitored as shown in Figure 8. The water molecule initially bridging the two cations exhibited an occupancy of 100% in the wild-type ensembles OPH_{wtc} and OPH_{wt} and interacts tightly with the β -metal and with the carboxyl group of Asp301 (Figure 8). In the OPH_{tmc} simulation, this water molecule formed a third interaction with the carbonyl group of Asp253. This hydrogen bond persists for the time simulated, whereas the hydrogen bond between the water molecule and the Asp301 is disrupted after 4 ns. The hydrogen bond between the bridging water molecule and Asp301 is replaced by a new hydrogen bond between the water and His55 imidazole group. One of the water molecules

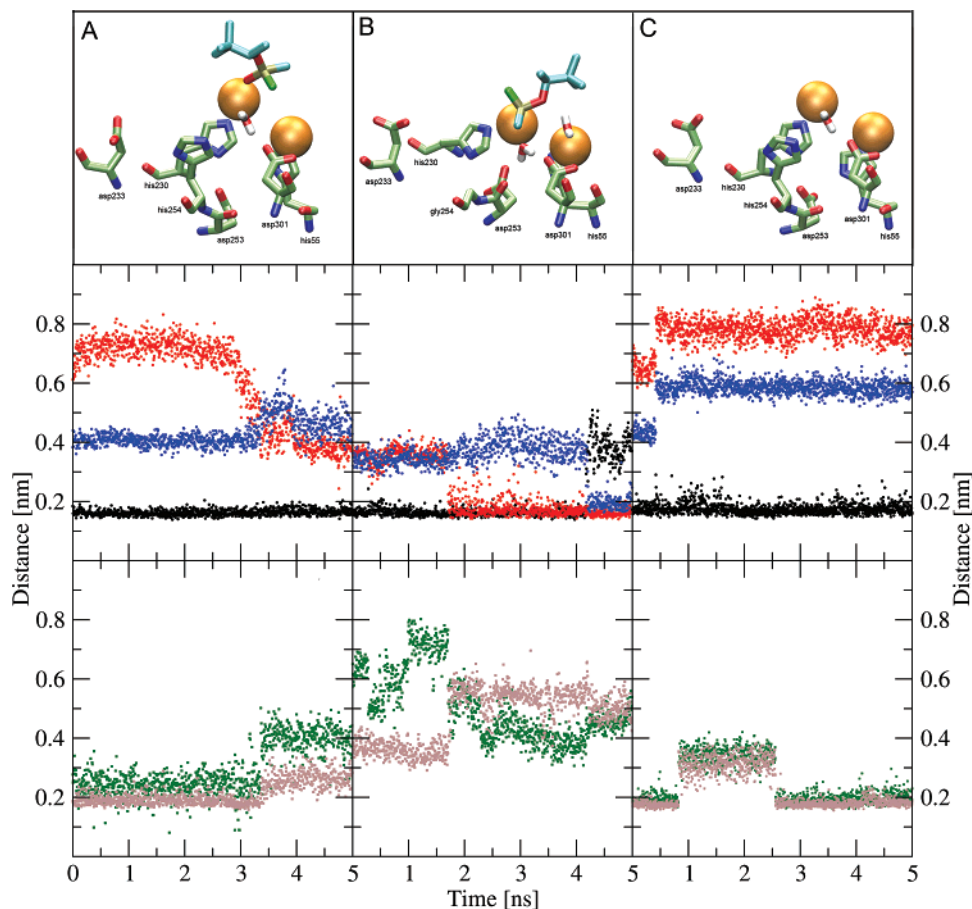


Figure 8. Active site snapshots and distances between catalytic residues and the catalytic water molecule as a function of time. Black circles, Asp301-O δ and water hydrogen atoms; red circles, Asp253-O δ and water hydrogen atoms; blue circles, his55-N ϵ and water hydrogen atoms; green circles, Asp253-O δ and His55-N ϵ atoms; and brown circles, Asp253-O δ and His230-N ϵ atoms. (A) OPH_{wtc}, (B) OPH_{tmc}, and (C) OPH_{wt}. Distances are averaged over 4 ps.

Table 2. Calculated and Experimental^{7,13} Binding Energy Differences for the Complexes Formed by *SpSc*-soman and the Wild-Type and Triple-Mutant H254G/H257W/L303T Forms of OPH

con-formers	Binding Energies ^a [kcal mol ⁻¹]						$\Delta\Delta G_{\text{calc}}$	$\Delta\Delta G_{\text{exp}}^b$
	OPH _{wtc}			OPH _{tmc}				
	ΔG_{elec}	ΔG_{nb}	ΔG_{calc}	ΔG_{elec}	ΔG_{nb}	ΔG_{calc}		
conf.-1	2.71	-11.61	-8.90	3.24	-14.00	-10.76	-3.2	-3.8
conf.-2	4.30	-12.19	-7.89	2.80	-14.55	-11.75		
conf.-3	2.14	-12.20	-10.06	2.86	-14.14	-11.28		
conf.-4	2.70	-11.40	-8.70	3.46	-14.00	-10.54		
conf.-5	4.60	-11.77	-7.17	3.42	-13.94	-10.52		
conf.-6	3.35	-13.30	-9.95	2.99	-13.92	-10.93		
conf.-7	3.01	-12.62	-9.61	0.82	-13.86	-13.04		
conf.-8	2.52	-13.16	-10.64	2.37	-13.99	-11.62		
conf.-9	4.22	-12.60	-8.38	3.74	-13.95	-10.21		
conf.-10	3.19	-13.44	-10.25	4.39	-14.38	-9.99		
conf.-11	3.68	-10.71	-7.03	3.07	-14.07	-11.00		
conf.-12	3.84	-10.54	-6.70	0.27	-14.23	-13.96		
conf.-13	3.85	-10.49	-6.64	2.50	-13.88	-11.38		
conf.-14	4.23	-10.70	-6.47	1.53	-13.71	-12.18		
conf.-15	4.78	-10.73	-5.95	1.14	-14.10	-12.96		
conf.-16	4.87	-12.47	-7.60	1.82	-13.70	-11.88		
conf.-17	3.83	-12.14	-8.31	2.60	-13.63	-11.03		
conf.-18	4.09	-12.15	-8.05	2.12	-13.31	-11.19		
conf.-19	3.62	-12.23	-8.61	1.36	-13.33	-11.97		
conf.-20	5.26	-12.35	-7.09	4.18	-13.58	-9.40		
averages	3.74	-11.94	-8.20	2.53	-13.91	-11.38		

^a $\Delta\Delta G_{\text{calc}}$ is the calculated binding energy difference; ΔG_{elec} and ΔG_{nb} are electrostatics and nonpolar contributions to ΔG_{calc} . ^b $\Delta\Delta G_{\text{exp}}$ is the experimental binding energy difference calculated from $k_{\text{cat}}/K_M^{7,13}$ at 298.15 K and according to $\Delta\Delta G_{\text{exp}} = RT \ln[(k_{\text{cat}}/K_M)_{\text{OPH}_{\text{tmc}}}/(k_{\text{cat}}/K_M)_{\text{OPH}_{\text{wtc}}}]$.²⁵

coordinated to the β -zinc forms a hydrogen bond with Asp301 and moves to the site earlier occupied by the bridging water molecule. The hydrogen bond between residue Asp253 and the water molecule is possible due to the H254G mutation, which eliminates any potential sterical hindrance by the residue His254 side chain. This interaction is absent in both wild-type simulations regardless of the presence of the substrate soman.

Binding energies between the *SpSc*-soman and the wild-type and triple-mutant forms of OPH were calculated by solving the linearized Poisson–Boltzmann equation in conjunction with a solvent-accessible surface area term for apolar interactions. The calculated binding free energies together with the nonpolar (ΔG_{nb}) and electrostatic (ΔG_{elec}) contributions are presented in Table 2. This model has been shown to overestimate experimental values in part because it does not account for the loss of translational and rotational entropy associated with the binding process. These contributions have been estimated to be 29–63 kJ mol⁻¹ for the binding of small ligands to proteins⁴⁴ and 2–4 kJ mol⁻¹ per frozen (protein or ligand) internal rotational degree of freedom.^{45,46} Therefore, comparisons between calculated and experimental values are presented in terms of differences between the binding energies of the complexes OPH_{wtc} and OPH_{tmc}. A comparison between $\Delta\Delta G_{\text{calc}}$ and $\Delta\Delta G_{\text{exp}}$, where $\Delta\Delta G = \Delta G_{\text{tmc}} - \Delta G_{\text{wtc}}$, shows good agreement between calculated and experimental values since the approximations inherent to the continuum model cancel out. Calculated values indicate that the binding of soman to either the wild-type or the triple-mutant OPH is driven by nonpolar interactions, whereas electrostatics interactions have only a small contribution to the overall binding energy. While this

is true for the thermodynamics of binding of soman to both enzymes, the difference in free energy of binding to the two enzymes, which determines the difference in specificity, has van der Waals and electrostatics contributions of similar magnitude.

The specificity constants k_{cat}/K_M corresponding to the binding of *SpSc*-soman to the wild-type and triple-mutant enzymes have been determined experimentally.^{7,13} They correspond to values of ca. $1.6 \pm 0.1 \times 10^1$ and 10^4 M⁻¹ s⁻¹, respectively. The specificity constant is a useful quantity to estimate relative binding energies $\Delta\Delta G$ between complexes composed of the same substrate and different enzymes (a parent and a mutant of the parent enzyme) because it accounts for both the activation energy and the binding energy.²⁵ The $\Delta\Delta G_{\text{exp}}$ estimated from experimental values of k_{cat}/K_M is -15.9 kJ mol⁻¹ more favorable to the binding of *SpSc*-soman to the triple-mutant OPH. Yet, *SpSc*-soman binds to wild-type and triple-mutant enzymes with similar K_M values of 2.5 ± 0.5 and 3.1 nM, respectively.^{7,47} The $\Delta\Delta G_{\text{exp}}$ estimated from the experimental values of K_M yields a binding energy difference of $+0.88$ kJ mol⁻¹ less favorable for the complex between *SpSc*-soman and the triple mutant. Therefore, it is the favorable interaction resulting from the enzyme-transition state complementarity—not from the enzyme-substrate complementarity—that makes the triple mutant more efficient than the wild type.

A catalytic mechanism has been proposed for the hydrolysis activity of OPH.^{8,21} In this mechanism, the solvent molecule bridging the two metal ions is activated for an in-line S_N2-nucleophilic attack via complexation to the binuclear metal center and a hydrogen-bonding interaction with residue Asp301.^{8,21} The substrate binds to the active site in an orientation that polarizes the P–O bond of phosphate ester through an interaction of the phosphoryl oxygen with the β -cation. The existence of a bridging water molecule between the two metal ions hydrogen-bonded with Asp301 is observed in both simulations and is consistent with the proposed mechanism. In the context of such a model, we hypothesize that the two hydrogen bonds involving soman and the nucleophilic water, respectively, can enhance the activity of the triple mutant through the stabilization of the transition-state conformation of the substrate. This would happen by means of three effects or their combination: (i) The interactions would force the substrate conformation closer to the transition state structure. (ii) By spatially and conformationally restraining the substrate and the catalytic water with respect to each other, these hydrogen bonds could facilitate the hydrolysis reaction by positioning the substrate and the water molecule in an optimal conformation for the nucleophilic attack. (iii) The negative charge on the trigonal bipyramidal phosphoester of the transition state would be neutralized by the hydrogen bond between the phosphoester oxygen atom of soman and a β -metal coordinated water.

Coordination Number of Zinc Ions in the Active Site.

The active site of OPH contains two Zn²⁺ cations required for full catalytic activity of the enzyme.⁵ In the available crystallographic structures of OPH, the α -metal ion is coordinated to residues His55, His57, and Asp301, while the β -metal ion is coordinated to residues His201 and His230.

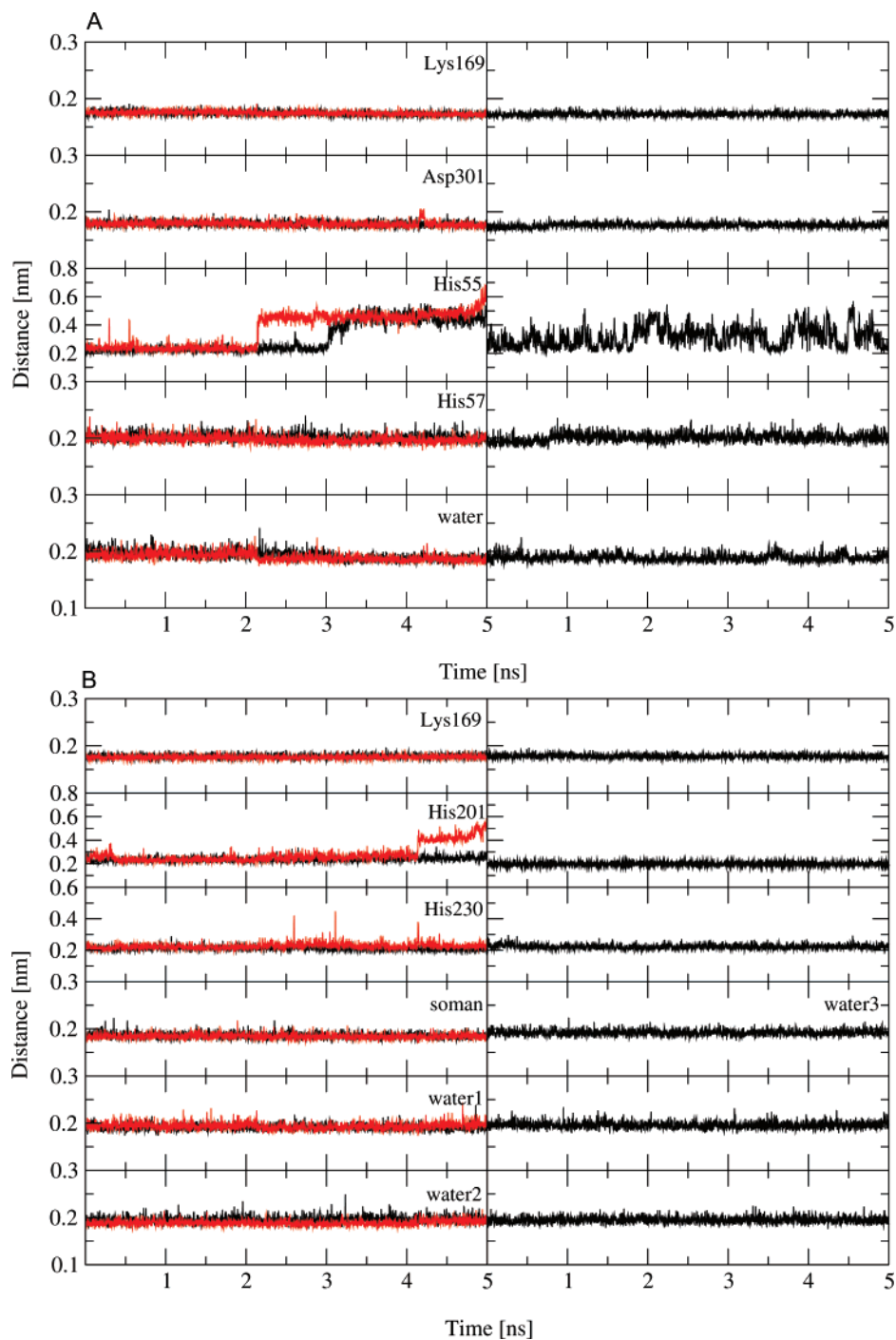


Figure 9. Distances between the Zn^{2+} cations and coordination ligands in the active site of OPH. Left columns: OPH_{wtc} in black and OPH_{tmc} in red. Right columns: OPH_{wt} in black. (A) α -Metal: coordination ligands are the carbamylated residues Lys169, Asp301-O δ , His55-N ϵ , His57-N δ , and one water molecule. (B) β -Metal: coordination ligands are the carbamylated Lys169, His201-N ϵ , His230-N ϵ , soman phosphoryl oxygen atom (only in the OPH_{wtc} in black and OPH_{tmc} simulations on the right), and water molecules.

In addition to these ligands, the two metal ions are bridged via a carboxylated lysine residue and a water molecule, which is thought to be the nucleophile for the hydrolytic attack on the phosphorus atom of the substrate.⁴⁸ The more buried α -metal has a ligand coordination number of five. The β -metal is more solvent-exposed and may acquire additional water ligands. The distances between divalent cations and coordination groups along time for the MD-derived ensembles of OPH are presented in Figure 9A and B.

During the initial 2 ns of simulation, the α -metal maintains a penta-coordinated geometry in the three ensembles (Figure 9). This cation is coordinated to residues His55, His57, and Asp301; the carboxylated Lys169; and a solvent molecule. In the simulations containing the substrate soman, residue His55 moves away from the Zn^{2+} cation, which then adopts a tetra-coordinated geometry. In the OPH_{wt} ensemble, the interaction between residue His55 and the α -metal fluctuates around the range 0.2–0.4 nm. The β -metal is coordinated

to different groups in the three simulations. In the presence of soman, the β -metal binds to the carboxylated residue Lys169, the substrate phosphoryl oxygen, residues His201 and His230, and two water molecules—among which is the nucleophile for the reaction catalyzed by OPH (Figure 9). In the absence of the substrate, a third water molecule replaces the substrate phosphoryl oxygen in the coordination of the metal. The β -metal exhibits predominantly a hexa-coordinated geometry in the MD simulations, but its interaction with His201 is disrupted after 4 ns of simulation in the OPH_{tmc} ensemble, leading to a penta-coordinated geometry.

The coordination geometry observed within the simulation times in this study of the α - and β -metals in the MD-derived ensembles can be summarized as follows: (i) in the absence of the substrate, the α -metal is predominantly penta-coordinated; (ii) in the presence of the substrate, the α -metal shifts from penta- to tetra-coordinated after ca. 2–3 ns; (iii) the β -metal is hexa-coordinated in the two wild-type simulations, regardless of the presence of the substrate; (iv) the β -metal transitions from hexa- to penta-coordinated in the OPH_{tmc} ensemble after 4 ns of simulation.

Previously, MD simulations of OPH bound to the substrates sarin and paraoxon were carried out by Koça et al.⁴⁹ to examine the coordination number of the Zn²⁺ cations in the protein active site. As in the present work, a nonbonded model²⁹ was employed to treat the Zn²⁺ metals with a charge of +2 in conjunction with the AMBER force field³² and the PME method.³⁷ It was found that such a model could not reproduce the penta-coordinated geometry of the divalent cations observed in the X-ray structures of OPH¹⁸ and obtained by means of ab initio optimization of a model of the active site of OPH composed of the side-chains of a few selected residues.⁴⁹ In these simulations, the less buried β -zinc had a hexa-coordinated geometry, whereas the coordination number of the more buried α -zinc was dependent on the presence of the substrate/inhibitor. In the presence of the substrate, the α -metal was penta-coordinated. Without a substrate, a water molecule was found to move in and bind to the α -metal that became hexa-coordinated. Given that a charge of +1.5 led to a penta-coordinated geometry of both zinc ions, the hexa-coordinated geometry was dismissed as an artifact of the force field used.⁴⁹

The present simulations differ from the study by Koça et al.⁴⁹ in two aspects: the α -zinc does not exhibit hexa-coordination regardless of the presence of the substrate, and the β -zinc is predominantly hexa-coordinated in the three ensembles, although it adopts a penta-coordinated geometry in the OPH_{tmc} ensemble after 4 ns. Because the present simulations are at least 5 times longer than the simulations carried out by Koça et al.,⁴⁹ it is possible that the discrepancies between the two studies are due to the different time lengths. However, on the basis of the published X-ray structures of OPH, the claim by Koça et al.⁴⁹ that both Zn²⁺ metals are penta-coordinated appears incorrect. It has been shown that both Zn²⁺ ions in the active site of OPH can be replaced with cadmium or manganese ions without a loss of enzymatic activity.^{5,20} Further examination of the OPH structures currently available in the PDB database shows the more solvent-exposed β -metal with coordination spheres

which are octahedral (in the complexes with Cd²⁺/Cd²⁺, Mn²⁺/Mn²⁺, and Zn²⁺/Cd²⁺),⁵⁰ tetrahedral (in the complex Zn²⁺/Zn²⁺),⁵¹ and trigonal bipyramidal (also in the complex Zn²⁺/Zn²⁺).⁵⁰ In the X-ray structure of the triple mutant, the substitution of His254 with a glycine creates a cavity that allows for the introduction of a third metal binding site.¹³ This third ion, which is absent from the X-ray structure containing a substrate analog, is coordinated in a tetrahedral arrangement by residues Asp253 and His230 and two water molecules, resulting in a distortion of the binuclear center.¹³

The picture emerging from these structural data is that the OPH active site can accommodate several catalytically active coordination geometries. Each of these geometries is possible through the coordination of the metals to additional water ligands or substrate chemical groups. This is consistent with the fact that the enzymatic activity of the wild-type OPH can be enhanced by alterations to the metal content of the enzyme,^{5,52} suggesting that the changes in catalytic metal geometries are associated with the alteration of substrate specificities.^{53,54} Zn²⁺ ions can adopt different numbers of ligands in their first coordination shell, either in solution or as part of metalloproteins,^{55–57} and the energy difference between complexes with different coordination numbers is relatively low.^{58,59} Indeed, much of the ubiquity of Zn²⁺ in functionally diverse proteins has been attributed to its intrinsically flexible coordination geometry.^{55–57,60,61} Furthermore, the solution structure of a protein often differs from its crystal structure due to effects of crystal packing and much lower hydration conditions. In solution, proteins possess a conformational flexibility that includes a wide range of hydration states not seen in the crystal. In addition, although the electron density at the zinc positions in proteins is usually well-defined, the zinc ligation geometry in the X-ray structure may still be influenced by the restraints and parameters used during the refinement procedure. These factors can affect the conformation and position of ligands (side chains and water molecules) around the divalent center, resulting in changes of the metal coordination geometry.

Conclusion

Organophosphorous hydrolase is unique among other organophosphate-degrading enzymes because it can hydrolyze phosphofluoridates, such as soman and sarin, and phosphothioates, such as VX, which constitute the major chemical warfare deterrents stockpiled by the United States and the former Soviet Union. Practical applications of OPH for the detection and detoxification of nerve agents and various environmental pollutants will require enzymes with enhanced structural stability and improved catalytic efficiency. A detailed analysis of the dynamic behavior of wild-type OPH and triple-mutant H254G/H257W/L303T bound to the substrate *SpSc*-soman has been undertaken and compared against the wild-type in the absence of a substrate in an effort to understand how these motions contribute to the enhanced specificity of the mutant. The analyses have shown that, upon substrate binding, OPH undergoes conformational changes that result in the widening of the active site entrance. The conformational changes are mostly limited to the loops in the entrance of the active site and exhibit larger amplitudes

in the triple mutant. These structural rearrangements corroborate previous kinetic studies, indicating that a conformational change might be the rate-limiting step for OPH hydrolysis activity.⁶ It has also been shown that the active site of OPH can accommodate several catalytically active coordination geometries, each of these geometries being possible through the coordination of the metals to additional water ligands or substrate chemical groups. This is consistent with the fact that the enzymatic activity of the wild-type OPH can be enhanced by alterations to the metal content of the enzyme,^{5,52} suggesting that the changes in catalytic metal geometries are associated with the alteration of substrate specificities.^{53,54} Furthermore, the complex between the triple mutant and *SpSc*-soman is stabilized by hydrogen bonds between the phosphoester oxygen atom of soman and a β -metal coordinated water and between the carbonyl group of Asp253 and the catalytic water. This latter interaction is possible only due to the H254G mutation, which eliminates any potential sterical hindrance by the His254 side chain. By spatially and conformationally restraining the substrate and the catalytic water with respect to each other, these hydrogen bonds are expected to facilitate the hydrolysis reaction by positioning the substrate and the water molecule in an optimal conformation for the nucleophilic attack. On the basis of binding energy calculations, we have argued that the enhanced efficiency of the triple mutant is determined by enzyme-transition-state complementarity and not by enzyme-substrate complementarity. Possibly, the hydrogen bonds occurring in the triple-mutant complex stabilize a conformation of the substrate closer to the transition state structure, enhancing the triple mutant specificity.

Acknowledgment. This research was supported by the D.O.E. Office of Advanced Scientific Computing Research. The authors acknowledge the William R. Wiley Environmental Molecular Sciences Laboratory for the computational resources required for this work. T.A.S. and T.P.S. acknowledge Dr. Roberto D. Lins for fruitful discussions. M.A.O. acknowledges Dr. Erich Vorpapel for help with QM calculations. Pacific Northwest National Laboratory is operated for the Department of Energy by Battelle.

References

- Gunderson, C. H.; Lehmann, C. R.; Sidell, F. R.; Jabbari, B. *Neurology* **1992**, *43*, 946–950.
- Brown, M. A.; Brix, K. *J. Appl. Toxicol.* **1998**, *6*, 393–408.
- Lallement, G.; Clarencon, D.; Masqueliez, C.; Baubichon, D.; Galonnier, M.; Burckhart, M.; Peoc'h, M. F.; Mestries, J. C. *Arch. Toxicol.* **1998**, *2*, 84–92.
- Polhuijs, M.; Langenberg, J. P.; Benschop, H. P. *Toxicol. Appl. Pharmacol.* **1997**, *1*, 156–161.
- Omburo, G. A.; Kuo, J. M.; Mullins, L. S.; Raushel, F. M. *J. Biol. Chem.* **1992**, *267*, 13278–13283.
- Caldwell, S. R.; Newcomb, J. R.; Schlechtand, K. A.; Raushel, F. M. *Biochemistry* **1991**, *30*, 7438–7444.
- Li, W. A.; Lum, K. T.; Chen-Goodspeed, M.; Sogorb, M. A.; Raushel, F. M. *Bioorg. Med. Chem.* **2001**, *9*, 2083–2091.
- Li, W. A.; Aubert, S. D.; Raushel, F. M. *J. Am. Chem. Soc.* **2003**, *125*, 7526–7527.
- Chen-Goodspeed, M.; Sogorb, M. A.; Wu, F. Y.; Hong, S. B.; Raushel, F. M. *Biochemistry* **2001**, *40*, 1325–1331.
- Chen-Goodspeed, M.; Sogorb, M. A.; Wu, F. Y.; Raushel, F. M. *Biochemistry* **2001**, *40*, 1332–1339.
- Gopal, S.; Rastogi, V.; Ashman, W.; Mulbry, W. *Biochem. Biophys. Res. Commun.* **2000**, *279*, 516–517.
- Dumas, D. P.; Wild, J. R.; Raushel, F. M. *Experientia* **1990**, *46*, 729–731.
- Hill, C. M.; Li, W. S.; Thoden, J. B.; Holden, M. H.; Raushel, F. M. *J. Am. Chem. Soc.* **2004**, *125*, 8990–8991.
- Benschop, H. P.; Konings, C.; van Genderen, J.; de Jong, L. P. A. *Toxicol. Appl. Pharmacol.* **1984**, *72*, 61–74.
- Lei, C.; Shin, Y.; Liu, J.; Ackerman, E. J. *J. Am. Chem. Soc.* **2002**, *124*, 11242–11243.
- Benning, M. M.; Kuo, J. M.; Raushel, F. M.; Holden, H. *Biochemistry* **1994**, *33*, 15001–15007.
- Benning, M. M.; Kuo, J. M.; Raushel, F. M.; Holden, H. *Biochemistry* **1995**, *34*, 7973–7978.
- Vanhooke, J. L.; Benning, M. M.; Raushel, F. M.; Holden, H. M. *Biochemistry* **1996**, *35*, 6020–6025.
- Grimsley, J. K.; Calamini, B.; Wild, J. R.; Mesecar, A. D. *Arch. Biochem. Biophys.* **2005**, *442*, 169–179.
- Rochu, D. N.; Viguié, F. R.; Crouzier, D.; Froment, M. T.; Masson, P. *Biochem. J.* **2004**, *380*, 627–633.
- Aubert, S. D.; Li, Y. C.; Raushel, F. M. *Biochemistry* **1994**, *43*, 5707–5715.
- Boehr, D. D.; Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2006**, *106*, 3055–3079.
- Boehr, D. D.; McElheny, D.; Dyson, H. J.; Wright, P. E. *Science* **2006**, *13*, 1638–1642.
- Schnell, J. R.; Dyson, H. J.; Wright, P. E. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 119–140.
- Fersht, A. Enzyme and Substrate Complementarity and Binding Energy. In *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*; W. H. Freeman: New York, 1998.
- Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598–1603.
- Koça, J.; Zhan, C. G.; Rittenhouse, R. C.; Ornstein, R. J. *Am. Chem. Soc.* **2003**, *123*, 817–826.
- Straatsma, T. P. *NWChem, a Computational Chemistry Package for Parallel Computers*, version 4.5; Pacific Northwest National Laboratory: Richland, WA, 2003.
- Stote, R. H.; Karplus, M. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 12–31.
- Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.* **1992**, *70*, 560–571.
- Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.
- Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

- (34) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (35) Hockney, R. W. The Potential Calculation and Some Applications. In *Methods in Computational Physics*; Alder, B., Fernbach, S., Rotenberg, M., Eds.; Academic Press: New York, 1970; Vol. 9.
- (36) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (37) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (38) Straatsma, T. P. *Data Intensive Computing for Complex Biological Systems*; Technical Report 2007, Pacific Northwest National Laboratory: Richland, WA, 2007.
- (39) Froloff, N.; Windemuth, A.; Honig, B. H. *Protein Sci.* **1997**, *6*, 1293–1301.
- (40) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037–10041.
- (41) Gilson, M. K.; Sharp, K. A.; Honig, B. H. *J. Comput. Chem.* **1988**, *9*, 327–335.
- (42) Simonson, T.; Brünger, A. T. *J. Phys. Chem.* **1994**, *98*, 4683–4694.
- (43) Honig, B. H.; Yang, A. S. *Adv. Protein Chem.* **1995**, *46*, 27–58.
- (44) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047–1069.
- (45) Doig, A. J.; Sternberg, M. J. E. *Protein Sci.* **1995**, *4*, 2247–2251.
- (46) Janin, J. *Proteins: Struct., Funct., Genet.* **1995**, *21*, 30–39.
- (47) Raushel, F. M. Personal communication, 2006.
- (48) Samples, C. R.; Howard, T.; Raushel, F. M.; DeRose, V. J. *Biochemistry* **2005**, *44*, 11005–11013.
- (49) Koça, J.; Zhan, C. G.; Rittenhouse, R. C.; Ornstein, R. L. *J. Comput. Chem.* **2003**, *24*, 368–378.
- (50) Benning, M. M.; Shim, H.; Raushel, F. M.; Holden, H. M. *Biochemistry* **1998**, *40*, 2712–2722.
- (51) Benning, M. M.; Hong, S. B.; Raushel, F. M.; Holden, H. M. *J. Biol. Chem.* **2000**, *275*, 30556–30560.
- (52) Hong, S. B.; Raushel, F. M. *Biochemistry* **1996**, *35*, 10904–10912.
- (53) Grimsley, J. K.; Scholtz, J. M.; Pacehand, C. N.; Wild, J. R. *Biochemistry* **1997**, *36*, 14366–14374.
- (54) Di-Sioudi, B. D.; Miller, C. E.; Lai, K.; Grimsley, J. K.; Wild, J. R. *Chem. Biol. Interact.* **1999**, *14*, 211–223.
- (55) Alberts, I. L.; Nadassy, K.; Wodak, S. J. *Protein Sci.* **1998**, *7*, 1700–1716.
- (56) Dudev, T.; Lim, C. *Chem. Rev.* **2003**, *103*, 773–787.
- (57) Wang, M. D. J.; Dudev, T.; Lim, C. *J. Phys. Chem. B* **2006**, *110*, 1889–1895.
- (58) Bock, C. W.; Katz, A. K.; Glusker, J. P. *J. Am. Chem. Soc.* **1995**, *117*, 3754–3765.
- (59) Dudev, T.; Lim, C. *J. Am. Chem. Soc.* **2000**, *122*, 11146–11153.
- (60) Lipscomb, W. N.; Strater, N. *Chem. Rev.* **1996**, *96*, 2375–2433.
- (61) Christianson, D. W.; Cox, J. D. *Annu. Rev. Biochem.* **1999**, *68*, 33–57.

CT700024H

JCTC Journal of Chemical Theory and Computation

Homoleptic Carbonyls of the Second-Row Transition Metals: Evaluation of Hartree–Fock and Density Functional Theory Methods[†]

Xuejun Feng,^{*,‡} Jiande Gu,[§] Yaoming Xie,^{||} R. Bruce King,^{*,||} and Henry F. Schaefer III^{||}

School of Chemical and Material Engineering, Southern Yangtze University, Wuxi 214122, P. R. China, Drug Design & Discovery Center, Shanghai Institute of Materia Medica, CAS, Shanghai 201203 P. R. China, and Center for Computational Chemistry and Department of Chemistry, University of Georgia, Athens, Georgia 30602

Received January 29, 2007

Abstract: The homoleptic mono- and multinuclear carbonyls for Mo, Tc, Ru, and Rh, namely, Mo(CO)₆, Ru(CO)₅, Tc₂(CO)₁₀, Ru₃(CO)₁₂, Rh₄(CO)₁₂, and Rh₆(CO)₁₆, are investigated theoretically by the Hartree–Fock method and three density functional theory (DFT) methods, i.e., BP86, B3LYP, and MPW1PW91, along with the SDD ECP basis sets. The results predicted by all the methods are basically in agreement with each other. The MPW1PW91 and BP86 methods predict geometric parameters and vibrational spectra, respectively, closest to the experimental values. For Ru₃(CO)₁₂ the relative energies of the *D*_{3h} isomer with only terminal CO groups and the *C*_{2v} isomer with two bridging CO groups are within 3 kcal/mol of each other with the lower energy isomer depending upon the computational method used. For Rh₄(CO)₁₂ the global minimum is predicted to have *C*_{3v} symmetry, with three bridging and nine terminal carbonyls, in accord with experiment. The Rh₆(CO)₁₆ structure has *T*_d symmetry and satisfies the Wade–Mingos rules for an octahedral cluster. Using the MPW1PW91 method the Rh–Rh distances in Rh₄(CO)₁₂ are found to be 2.692 Å and 2.750 Å and those in Rh₆(CO)₁₆ to be 2.785 Å.

1. Introduction

Well characterized isolable homoleptic carbonyl derivatives of the second-row transition metals include Mo(CO)₆,^{1–3} Tc₂(CO)₁₀,^{4,5} Ru(CO)₅,⁶ and Ru₃(CO)₁₂.^{7,8} In addition, multinuclear homoleptic rhodium carbonyls were observed as early as 1943.⁹ Subsequent X-ray diffraction studies^{10–12} have shown these rhodium carbonyls to be tetranuclear Rh₄(CO)₁₂ and hexanuclear Rh₆(CO)₁₆. The geometric parameters for Rh₄(CO)₁₂ and Rh₆(CO)₁₆ have then been determined by subsequent experimental work.^{13–20} Comparison of the structures of second-row transition-metal carbonyls with

those of the corresponding first-row transition-metal carbonyls is of interest since in some cases the structures are different. For example, the structure of Ru₃(CO)₁₂ has all terminal CO groups with an equilateral Ru₃ triangle,^{7,8} whereas the structure of the corresponding isoelectronic Fe₃(CO)₁₂ has ten terminal CO groups and two bridging CO groups with an isosceles Fe₃ triangle.²¹

Density functional theory (DFT) certainly appears to be a powerful and effective computational tool to study organotransition-metal chemistry.^{22–39} In this connection we have used the B3LYP and BP86 methods along with the all-electron DZP basis sets to study a series of first-row transition-metal carbonyl derivatives.^{40–47} Our results also show that the BP86 method may be somewhat more reliable than the B3LYP method for those organometallic systems.

There are fewer theoretical studies on compounds containing the second-row transition metals. All electron computations on second-row transition-metal derivatives are expected

[†] This paper is dedicated to the memory of F. Albert Cotton (1930–2007), a pioneer in structural metal carbonyl chemistry as well as many other areas of inorganic chemistry.

* Corresponding author e-mail: fxj@sytu.edu.cn and rbking@chem.uga.edu.

[‡] Southern Yangtze University.

[§] Shanghai Institute of Materia Medica.

^{||} University of Georgia.

to be much more expensive in terms of computing resources than those on corresponding first-row transition-metal derivatives. Effective core potential (ECP) and related basis sets^{48–51} provide a simple but efficient approach for reducing the computational effort while considering relativistic effects, especially for the second- (and third-) row transition metals. ECP methods have been tested on the second-row metal carbonyls Mo(CO)₆ and Ru(CO)₅ in 1996⁵² and were subsequently used to study the molecular structures of Ru₃(CO)₁₂ isomers⁵³ and most recently⁵⁴ the infrared spectra of rhodium carbonyl clusters. In the present paper we use ECP basis sets to explore the performance of different DFT methods, including a new generation DFT method as well as a Hartree–Fock method on the experimentally known second-row transition-metal carbonyls. We anticipate this work to provide a basis for more extensive future DFT studies on second-row transition-metal carbonyls and related organometallic compounds.

2. Theoretical Methods

A Hartree–Fock method and three different density functional theory (DFT) methods were used in the present study. The Hartree–Fock self-consistent-field (SCF) method was chosen by us here despite its lack of treatment for the electron correlation effect because it was found by Cotton and co-workers⁵⁵ to give satisfactorily optimized geometry and other properties for compounds containing the second- and third-row transition-metal atoms, such as the palladium(III) derivative Pd₂(hpp)₄Cl₂.

The density functional theory (DFT) methods used here include the B3LYP method, which is the hybrid DFT/Hartree–Fock method using Becke's three-parameter functional (B3) with the Lee–Yang–Parr (LYP) correlation functional.^{56,57} The second DFT method is the BP86 method, which uses Becke's 1988 exchange functional (B) with Perdew's 1986 gradient corrected correlation functional method (P86).⁵⁸ The third DFT method is a new generation functional MPW1PW91, which is a combination of the modified Perdew–Wang exchange functional with Perdew–Wang's 91 gradient-correlation functional.⁵⁹ This MPW1PW91 functional has been shown to be better than the first generation functional for the heavy transition-metal compounds.⁶⁰

The Stuttgart/Dresden double- ζ (SDD) ECP basis sets^{61,62} were used for the Mo, Tc, Ru, and Rh heavy atoms. In these basis sets the 28 core electrons in the transition-metal atoms are replaced by an effective core potential (ECP), and the valence basis sets are contracted from (8s7p6d) primitive sets to (6s5p3d). The effective core approximation includes relativistic contributions which become significant for the heavy transition-metal atoms. For the C and O atoms, the all electron DZP basis sets are used. They are Huzinaga–Dunning's contracted double- ζ contraction sets^{63,64} plus a set of spherical harmonic d polarization functions with the orbital exponents $\alpha_d(\text{C}) = 0.75$ and $\alpha_d(\text{O}) = 0.85$. The DZP basis sets for C and O atoms may be designated as (9s5p1d/4s2p1d). For Rh₆(CO)₁₆, there are 696 contracted Gaussian functions. All of the computations were carried out with the Gaussian 03 program⁶⁵ in which the fine grid (99 590) is

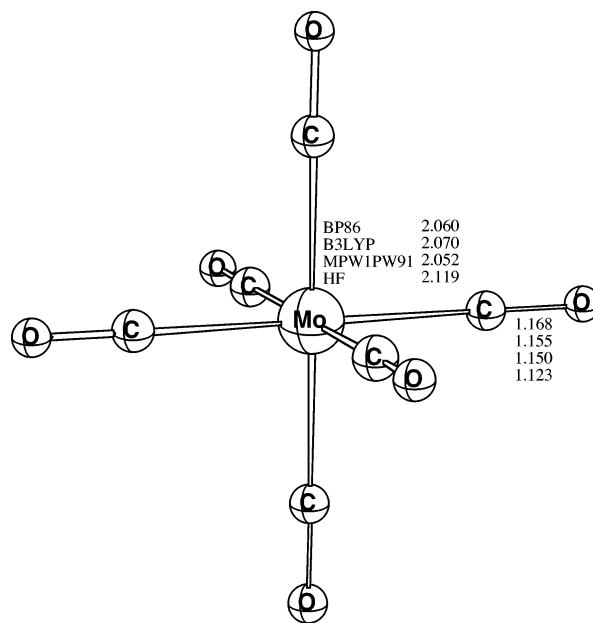


Figure 1. The optimized structure of Mo(CO)₆ (*O_h* symmetry). The distances are given in Å.

Table 1. Comparison of Experimental and Theoretical Geometrical Parameters of Mo(CO)₆ (*O_h*)

	BP86	B3LYP	MPW1PW91	HF	exp. ⁵	exp. ³
Mo–C	2.060	2.070	2.052	2.119	2.063(3)	2.059(3)
C–O	1.168	1.155	1.150	1.123	1.145(2)	1.125(5)
mean absolute errors	0.017	0.014	0.012	0.035		

chosen for evaluating integrals numerically. In order to assess the performance of the f functions for the transition-metal atoms, we have also run the Hartree–Fock method with the SDD basis sets plus a set of polarization f functions, and the results are compared with the Hartree–Fock results without the f functions.

The geometries of all structures were fully optimized using the HF method and the three DFT methods independently. The harmonic vibrational frequencies were also obtained at the same levels. The corresponding infrared intensities were evaluated analytically as well.

3. Results and Discussion

3.1. Mo(CO)₆. The infrared vibrational spectrum for molybdenum hexacarbonyl Mo(CO)₆ was first reported in 1955 and in 1962.¹ Its geometric parameters have been measured by electron diffraction² and X-ray diffraction.³

The optimized structure of Mo(CO)₆ is the expected octahedron (Figure 1). The related geometric parameters as well as the available experimental bond distances are listed in Table 1. There are only two independent bond distances, namely the equivalent Mo–C bonds and the equivalent C–O bond. The theoretical bond distances with different methods are in reasonable agreement with each other. However, the MPW1PW91 results are the closest to the experimental values.

3.2. Tc₂(CO)₁₀. Although the element technetium does not exist in nature, its carbonyl derivative ditechneum deca-

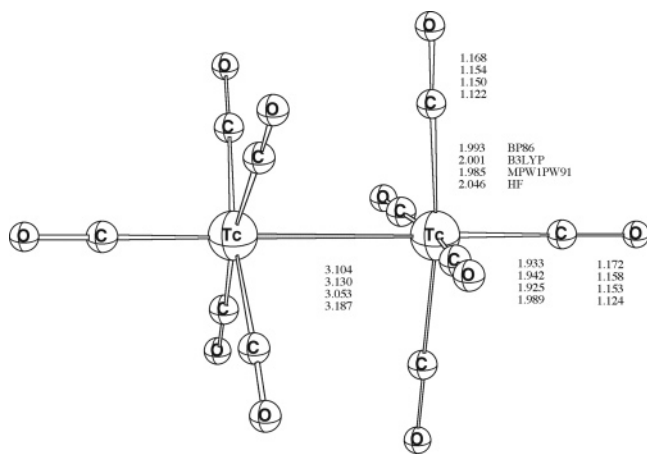


Figure 2. The optimized structure of $\text{Tc}_2(\text{CO})_{10}$ (D_{4d} symmetry). The distances are given in Å.

Table 2. Comparison of Experimental and Theoretical Geometrical Parameters of $\text{Tc}_2(\text{CO})_{10}$ (D_{4d})

	BP86	B3LYP	MPW1PW91	HF	exp. ⁵
Tc–Tc	3.104	3.130	3.053	3.187	3.036(6)
Tc–C (eq)	1.993	2.001	1.985	2.046	2.000(13)
C–O (eq)	1.168	1.154	1.150	1.122	1.122(12)
Tc–C (api)	1.933	1.942	1.925	1.989	1.899(11)
C–O (api)	1.172	1.158	1.153	1.124	1.205(13)
mean	0.038	0.043	0.028	0.074	
absolute errors					

carbonyl $\text{Tc}_2(\text{CO})_{10}$ was first prepared as early as in 1961, and its crystal unit was reported in 1962.⁴ Subsequently the geometrical parameters of $\text{Tc}_2(\text{CO})_{10}$ were measured.⁵

The optimized structure of $\text{Tc}_2(\text{CO})_{10}$ has D_{4d} symmetry (Figure 2). The theoretical geometrical parameters as well as the experimental bond distances are listed in Table 2. The theoretical axial Tc–Tc distances fall in the range from 3.05 to 3.19 Å, among which the HF method predicts it too long (3.187 Å) and the MPW1PW91 method predicts it to be the shortest (3.053 Å) and closest to the experimental value (3.036 Å).⁵ The MPW1PW91 method gives Tc–C and C–O bond distances closest to the experimental values.

3.3. Ru(CO)₅ and Ru₃(CO)₁₂. Ruthenium is in the same group as iron, and the carbonyls of ruthenium are expected to be similar to those of iron. Indeed, ruthenium pentacarbonyl $\text{Ru}(\text{CO})_5$ and triruthenium dodecacarbonyl $\text{Ru}_3(\text{CO})_{12}$ have been found to be stable species.^{6,7} However, the structure of $\text{Ru}_3(\text{CO})_{12}$ is different from that of $\text{Fe}_3(\text{CO})_{12}$.

The optimized structure of $\text{Ru}(\text{CO})_5$, like that of $\text{Fe}(\text{CO})_5$, has D_{3h} symmetry (Figure 3). The theoretical geometrical parameters as well as the experimental bond distances are listed in Table 3. The theoretical Ru–C distances, whether axial or equatorial, fall in the range from 1.94 to 2.02 Å. For most bond distances the HF method predicts the longest and the MPW1PW91 method predicts the shortest with the MPW1PW91 method giving values closes to the experimental values.

Possible triangular structures for $\text{Ru}_3(\text{CO})_{12}$ include the experimentally known^{7,8} D_{3h} structure with all terminal CO groups (Figure 4) or a C_{2v} structure (Figure 5) similar to the known structure²¹ of $\text{Fe}_3(\text{CO})_{12}$ with two CO groups bridging

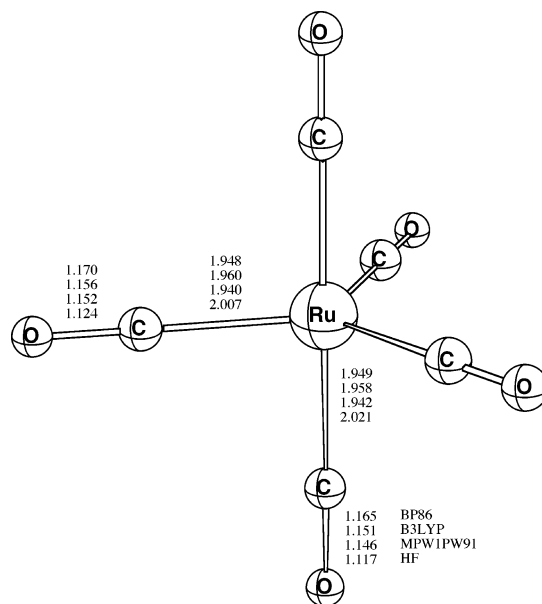


Figure 3. The optimized structure of $\text{Ru}(\text{CO})_5$ (D_{3h} symmetry). The distances are given in Å.

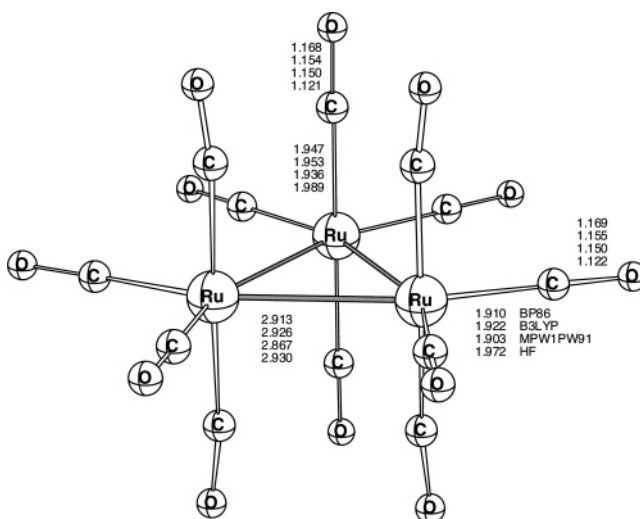


Figure 4. The optimized unbridged structure of $\text{Ru}_3(\text{CO})_{12}$ (D_{3h} symmetry). The distances are given in Å.

Table 3. Comparison of Experimental and Theoretical Geometrical Parameters of $\text{Ru}(\text{CO})_5$ (D_{3h})

	BP86	B3LYP	MPW1PW91	HF	exp. ⁶
Ru–C(ax)	1.949	1.958	1.942	2.021	1.941(13)
C–O (ax)	1.165	1.151	1.146	1.117	1.126(2)
Ru–C(eq)	1.948	1.960	1.940	2.007	1.961(9)
C–O (eq)	1.170	1.156	1.152	1.124	1.127(2)
mean	0.026	0.018	0.017	0.034	
absolute errors					

one of the edges of the triangle. The relative energetics of the two types of structures (Table 4) depends on the computational method used. Only the Hartree–Fock method indicates the known D_{3h} isomer of $\text{Ru}_3(\text{CO})_{12}$ to be much lower in energy (18.3 kcal/mol) than the C_{2v} isomer. The DFT methods show the two isomers of $\text{Ru}_3(\text{CO})_{12}$ to have energies within ≤ 3 kcal/mol of each other. The BP86 and

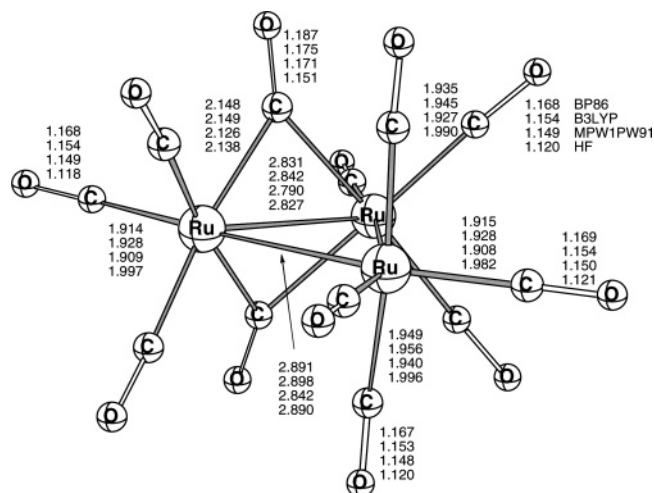


Figure 5. The optimized dibridged structure of $\text{Ru}_3(\text{CO})_{12}$ (C_{2v} symmetry). The distances are given in Å.

Table 4. Total Energies (hartrees) and Relative Energies (kcal/mol) for the Two $\text{Ru}_3(\text{CO})_{12}$ Isomers

	BP86	B3LYP	MPW1PW91	HF
$\text{Ru}_3(\text{CO})_{12}$ (D_{3h})	-1645.72062 (0.0)	-1645.33769 (0.0)	-1645.02067 (0.0)	-1635.54622 (0.0)
$\text{Ru}_3(\text{CO})_{12}$ (C_{2v})	-1645.72394 (-2.1)	-1645.33334 (2.7)	-1645.02122 (-0.3)	-1635.51702 (18.3)

Table 5. Comparison of Experimental and Theoretical Geometrical Parameters of the Unbridged Isomer of $\text{Ru}_3(\text{CO})_{12}$ (D_{3h})

	BP86 ^a	BP86 ^b	B3LYP	MPW1PW91	HF	exp. ⁷
Ru–Ru	2.913	2.912	2.926	2.867	2.930	2.854(4)
Ru–C(eq)	1.910	1.916	1.922	1.903	1.972	1.921(5)
C–O (eq)	1.169		1.155	1.150	1.122	1.127(2)
Ru–C(ax)	1.947	1.957	1.953	1.936	1.989	1.942(4)
C–O (ax)	1.168		1.154	1.150	1.121	1.133(2)
mean absolute errors	0.030		0.027	0.016	0.038	

^a This work. ^b Calculations in ref 53.

MPW1PW91 methods predict lower energies for the C_{2v} isomer by 2.1 and 0.3 kcal/mol, respectively, whereas the B3LYP method predicts a lower energy for the D_{3h} isomer by 2.7 kcal/mol. These calculations taken together suggest that the energies of both isomers of $\text{Ru}_3(\text{CO})_{12}$ are so similar that one isomer is readily converted to the other isomer. These similar energies of the D_{3h} and C_{2v} isomers of $\text{Ru}_3(\text{CO})_{12}$ are consistent with previous calculations⁶⁶ as well as the fluxional properties⁶⁷ found experimentally by NMR methods for $\text{Ru}_3(\text{CO})_{12}$.

For the experimentally known D_{3h} isomer of $\text{Ru}_3(\text{CO})_{12}$ (Figure 4) the HF method predicts too long Ru–Ru (2.93 Å) and Ru–C (1.972 Å and 1.989 Å) distances (Table 5). The three DFT methods predict similar results, among which again the MPW1PW91 method predicts the shortest bond distances (e.g., 2.867 Å for the Ru–Ru distances) but closest to the experimental results.⁷

Table 6 provides information on the geometrical parameters computed by various methods for the doubly bridged isomer of $\text{Ru}_3(\text{CO})_{12}$ (Figure 5). The single Ru–Ru edge of

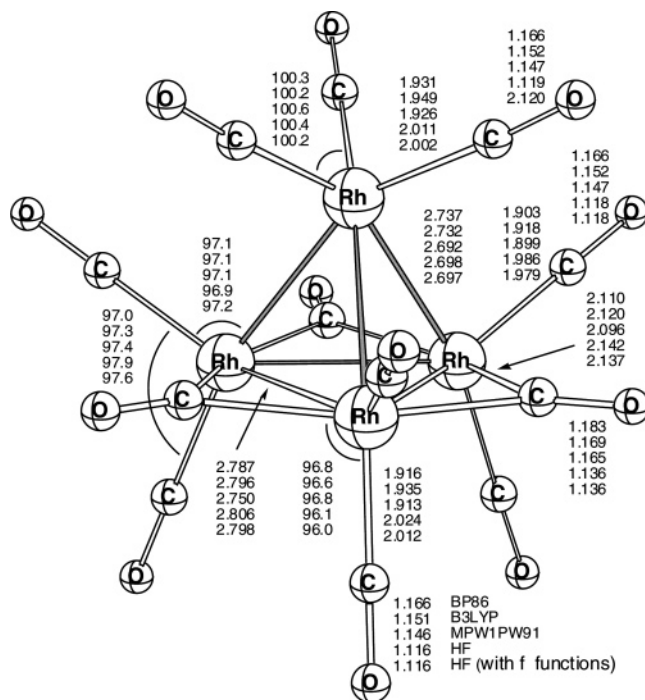


Figure 6. The optimized geometry of $\text{Rh}_4(\text{CO})_{12}$ (C_{3v} symmetry). The distances are given in Å.

Table 6. Theoretical Geometrical Parameters of the Doubly Bridged Isomer of $\text{Ru}_3(\text{CO})_{12}$ (C_{2v})

	BP86	B3LYP	MPW1PW91	HF
Ru–Ru (unbridged)	2.891	2.898	2.842	2.890
Ru–Ru (bridged)	2.831	2.842	2.790	2.827
Ru–C (to bridging CO)	2.148	2.149	2.126	2.138
Ru–C (to terminal CO)	1.914	1.928	1.909	1.997
	1.935	1.945	1.927	1.990
	1.949	1.956	1.940	1.996
	1.915	1.928	1.908	1.982

the Ru_3 isosceles triangle bridged by the two CO groups is 0.05–0.07 Å shorter than the two equivalent unbridged Ru–Ru edges in accord with numerous previous observations on CO bridged versus unbridged metal–metal bonds. The Ru–C distances to the bridging CO groups are ~ 0.2 Å longer than those to the terminal CO groups also in accord with previous experience in binuclear and trinuclear metal carbonyl derivatives.

3.4. $\text{Rh}_4(\text{CO})_{12}$. Possible tetrahedral structures for $\text{Rh}_4(\text{CO})_{12}$ can have either all terminal carbonyls (as found experimentally for $\text{Ir}_4(\text{CO})_{12}$) or three bridging carbonyls around the base of the tetrahedron and nine terminal carbonyls, as found experimentally for $\text{Co}_4(\text{CO})_{12}$ and $\text{Rh}_4(\text{CO})_{12}$. The relative energetics of these two structures for $\text{Rh}_4(\text{CO})_{12}$ are of interest.

The global minimum of $\text{Rh}_4(\text{CO})_{12}$ optimized by all the methods is a C_{3v} structure with three edge-bridging carbonyl groups and nine terminal carbonyl groups, namely, $\text{Rh}_4(\mu\text{-CO})_3(\text{CO})_9$ (Figure 6). Our theoretical geometrical parameters predicted by various methods are listed in Table 7. The experimental results^{11,68} and the previous theoretical results¹⁷ are also listed for comparison. The experimental (X-ray diffraction) structural parameters, which are slightly distorted

Table 7. Comparison of Experimental and Theoretical Geometric Parameters of the Global Minimum $\text{Rh}_4(\text{CO})_{12}$ (C_{3v})

	BP86	B3LYP	MPW1PW91	HF ^a	expt ^b	expt ^c	LDA ¹⁷	GGA ¹⁷
Rh(apex)–Rh(bas.)	2.737	2.732	2.692	2.698 (2.697)	2.715 (2.693–2.746)	2.673 (2.656–2.690)	2.683	2.996
Rh(bas.)–Rh(bas)	2.787	2.796	2.750	2.806 (2.798)	2.749 (2.710–2.804)	2.724 (2.693–2.767)	2.695	2.992
Rh(bas.)–C(bridg.)	2.110	2.120	2.096	2.142 (2.137)	2.00 (1.87–2.24)	2.101 (2.065–2.137)	2.065	2.120
Rh(bas.)–C(term.)	1.903	1.918	1.899	1.986 (1.978)	1.94 (1.69–2.16)	1.917 (1.878–1.953)	1.879	1.906
Rh(apex)–C	1.931	1.948	1.926	2.011 (2.002)	1.99 (1.88–2.11)	1.943 (1.929–1.962)	1.905	1.933
mean absolute errors	0.043	0.039	0.026	0.054				

^a The results predicted using larger basis sets (a set of f functions for Rh is added) are in parentheses. ^b Reference 11. Bond distances are averaged out for the equivalent Rh–Rh bonds with the range indicated in parentheses. ^c Reference 68. Bond distances are averaged out for the equivalent Rh–Rh bonds with the range indicated in parentheses.

Table 8. Total Energies (hartrees) and Relative Energies (kcal/mol) for the Two $\text{Rh}_4(\text{CO})_{12}$ Isomers

	BP86	B3LYP	MPW1PW91	HF	HF ^a
$\text{Rh}_4(\text{CO})_{12}$ (C_{3v})	–1803.39125 (0.0)	–1802.90099 (0.0)	–1802.59607 (0.0)	–1791.96822 (0.0)	–1791.99017 (0.0)
$\text{Rh}_4(\text{CO})_{12}$ (T_d)	–1803.34852 (26.8)	–1802.85890 (26.4)	–1802.54890 (29.6)	–1791.91958 (30.5)	–1791.94736 (26.9)

^a A set of f functions for Rh are added to DZP basis set.

in the crystal structure, are averaged out in Table 7 to represent the ideal C_{3v} symmetry.

Among the different methods, the predicted bond distances differ slightly (Table 7). For the Hartree–Fock results adding f functions to the Rh basis set has little effect (<0.01 Å). For most of the bonds, the HF method predicts the longest and the MPW1PW91 method predicts the shortest. The MPW1PW91 method predicts bond distances the closest to the most recent X-ray diffraction experimental values.⁶⁸

Our theoretical distances of the basal Rh–Rh bonds with CO bridges range from 2.750 to 2.806 Å, whereas those for the basal-apical Rh–Rh bonds without CO bridges range from 2.682 to 2.737 Å. The basal-apical Rh–Rh bond distances have been found to be consistently longer than those for the basal Rh–Rh bonds by 0.05 Å (DFT methods) or 0.1 Å (the HF method). Using the MPW1PW91 functional, the basal Rh–Rh bond distances were predicted to be 2.692 Å, which is only 0.02 Å different from the experimental value.⁶⁸ For the basal-apical Rh–Rh bond distances the MPW1PW91 predicted value (2.750 Å) is also closest to experiment (Table 7). The basal Rh–C distances to the bridging CO groups (2.096–2.142 Å) are significantly longer than those to the terminal CO groups (1.903–1.986 Å) indicating a lower Rh–C bond order for the bridging CO groups relative to the terminal CO groups. The apical Rh–C bond lengths are in the range of 1.926–2.011 Å. The MPW1PW91 predicted Rh–C bond distances are in good agreement with experiment⁶⁸ within 0.018 Å.

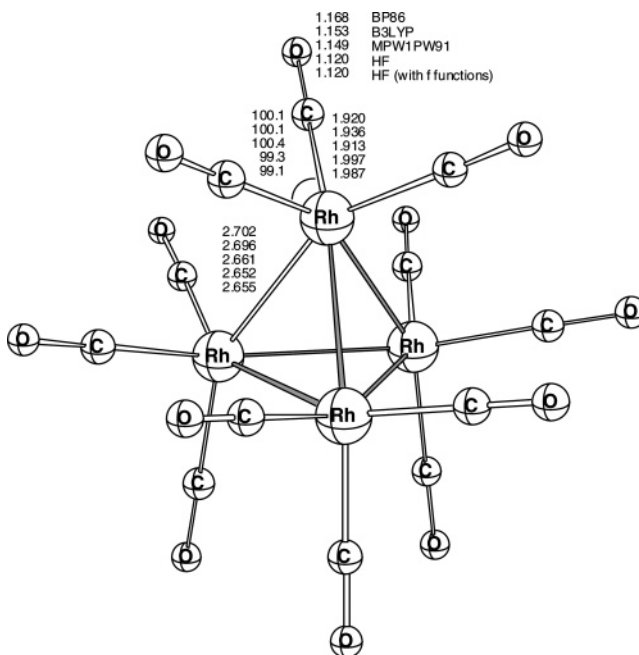
Besancon and co-workers¹⁷ reported density functional calculations on the bridged (C_{3v}) form of $\text{Rh}_4(\text{CO})_{12}$ with two different functionals (the LDA and GGA methods with Slater-type basis functions). Their LDA method predicted all bond distances similar to our MPW1PW91 results, while their GGA method predicted almost the same Rh–C bond lengths as our B3LYP and BP86 results but rather long Rh–Rh distances (>2.99 Å).

The T_d structure for $\text{Rh}_4(\text{CO})_{12}$ having exclusively terminal CO groups (Figure 7) was optimized using four methods. The energy of the T_d structure was found to be higher than that of the C_{3v} structure by 26.4–30.5 kcal/mol depending upon the method used (Table 8). Our optimized symmetry parameters and theoretical predicted values as well as the

Table 9. Comparison of Theoretical Geometrical Parameters of $\text{Rh}_4(\text{CO})_{12}$ (T_d)

	BP86	B3LYP	MPW1PW91	HF ^a	LDA ¹⁷
Rh–Rh	2.702	2.696	2.661	2.652 (2.655)	2.601
Rh–C	1.920	1.936	1.913	1.997 (1.987)	1.879
C–O	1.168	1.153	1.149	1.120 (1.120)	1.149

^a The results predicted by larger basis sets (a set of f functions for Rh is added) are given in parentheses.

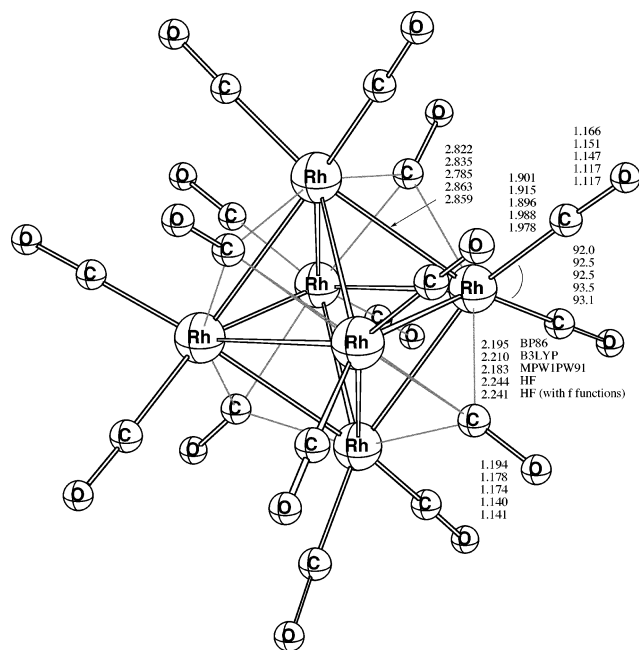
**Figure 7.** The optimized geometry of $\text{Rh}_4(\text{CO})_{12}$ (T_d symmetry). The distances are given in Å.

previous theoretical results¹⁷ are summarized in Table 9. The geometrical parameters obtained from the different methods agree reasonably with each other. Again, the effect of the f functions added to Rh atoms is negligible. Our theoretical predicted Rh–Rh bond lengths for the T_d structure of $\text{Rh}_4(\text{CO})_{12}$ are in the range from 2.661 to 2.702 Å, while the Rh–C bond lengths are from 1.913 to 1.997 Å. The Rh–Rh distances in the T_d $\text{Rh}_4(\text{CO})_{12}$ structure are found to be shorter than the unbridged Rh–Rh distances in the C_{3v} structure of $\text{Rh}_4(\text{CO})_{12}$ by 0.031–0.036 Å. Using the MPW1PW91 functional, the Rh–Rh bond distances and the

Table 10. Comparison of Experimental and Theoretical Geometrical Parameters of $\text{Rh}_6(\text{CO})_{16}$ (T_d)

	BP86	B3LYP	MPW1PW91	HF	exp. ¹²	exp. ⁷³
Rh–Rh	2.822	2.835	2.785	2.863 (2.859) ^a	2.776 (2.762–2.790)	2.750 ^a
Rh–C	1.901	1.915	1.896	1.988 (1.978) ^a	1.864 (1.849–1.879)	1.915 (1.900–1.930)
C–O	1.166	1.151	1.147	1.117 (1.117) ^a	1.155 (1.140–1.170)	1.126 (1.113–1.139)
Rh–C	2.195	2.210	2.183	2.244 (2.241) ^a	2.168 (2.156–2.180)	2.186 (2.182–2.190)
C–O	1.194	1.178	1.174	1.140 (1.141) ^a	1.201 (1.179–1.223)	1.153 (1.147–1.159)
mean absolute errors	0.026	0.028	0.009	0.037		

^a In parentheses are the results predicted by larger basis sets (a set of f functions for Rh is added).

**Figure 8.** The optimized structure of $\text{Rh}_6(\text{CO})_{16}$ (T_d symmetry). The distances are given in Å.

Rh–C bond lengths are predicted to be 2.661 Å and 1.913 Å, respectively. No experimental results are available for comparison. The previous theoretical results¹⁷ for the bond distances predicted by the LDA method are close to but slightly shorter than our MPW1PW91 predictions.

3.5. $\text{Rh}_6(\text{CO})_{16}$. The molecule $\text{Rh}_6(\text{CO})_{16}$ was the first metal carbonyl cluster to be discovered with more than four metal atoms. It is of historical interest related to chemical bonding theories for metal cluster bonding leading to the development of the Wade-Mingos rules.^{69–72} Optimization of the structure of $\text{Rh}_6(\text{CO})_{16}$ gave an idealized T_d structure with four equivalent face-bridging carbonyl groups and 12 equiv terminal carbonyl groups (Figure 8). The 12 equiv Rh–Rh distances in $\text{Rh}_6(\text{CO})_{16}$ (Table 10) are in the range 2.785–2.863 Å predicted by the different methods. These Rh–Rh distances in $\text{Rh}_6(\text{CO})_{16}$ are significantly longer than the Rh–Rh distances in $T_d \text{Rh}_4(\text{CO})_{12}$ by 0.120–0.211 Å, suggesting a lower effective Rh–Rh bond order in $\text{Rh}_6(\text{CO})_{16}$ than in $\text{Rh}_4(\text{CO})_{12}$. The MPW1PW91 method predicts the Rh–Rh bond distances to be 2.785 Å, which are the closest to the experimental values⁷³ (Table 10). The Rh–C distances for the terminal carbonyls fall in the range 1.896–1.988 Å, whereas the C–O bonds are in the range 1.117–1.166 Å. The Rh–C bond distances for the bridging carbonyls fall in the range 2.183–2.244 Å, whereas the C–O bond lengths

Table 11. Comparison of the Experimental and Theoretical Infrared $\nu(\text{CO})$ Frequencies for the Metal Carbonyls Studied in This Work

compound	method	infrared frequencies, ^a cm^{-1}
$\text{Mo}(\text{CO})_6$	Exp ⁷⁵	82m, 367s, 596vs, 2003vvs
(O_h)	BP86	84(1), 394(34), 602(96), 1979(1790)
	B3LYP	88(1), 381(51), 611(112), 2057(2080)
	MPW1PW91	90(1), 401(50), 629(127), 2099(2140)
	HF	100(2), 342(97), 627(181), 2298(2720)
$\text{Tc}_2(\text{CO})_{10}$	Exp ⁷⁶	1984s, 2017vs, 2065s
(D_{4d})	BP86	1978(1100), 2002(2340), 2047(1060)
	B3LYP	2052(1160), 2082(2710), 2120(1380)
	MPW1PW91	2096(1280), 2125(2770), 2166(1300)
	HF	2292(601), 2329(3560), 2329(3310)
$\text{Ru}(\text{CO})_5$	Exp ⁷⁷	2002s, 2039s
(D_{3h})	BP86	1985(1140), 2017(1290)
	B3LYP	2063(1310), 2100(1470)
	MPW1PW91	2106(1360), 2144(1510)
	HF	2301(2790), 2372(1710)
$\text{Ru}_3(\text{CO})_{12}$	Exp ⁷⁸	2061s, 2034s, 2015s, 1997m
(D_{3h})	BP86	2042(1790), 2015(2530), 1995(560), 1992(19)
	B3LYP	2118(2340), 2098(2990), 2074(460), 2071(33)
	MPW1PW91	2163(2380), 2140(3050), 2117(488), 2115(10)
	HF	2351(4220), 2348(4330), 2320(34), 2307(230)
$\text{Rh}_4(\text{CO})_9^-$ ($\mu\text{-CO}$) ₃	Exp ⁵⁴	1887m, 2044m, 2046m, 2071s, 2076s
(C_{3v})	BP86	1879(689), 1994(51), 2009(283), 2038(1410), 2039(1935)
	B3LYP	1953(867), 2080(31), 2095(362), 2123(1690), 2124(2510)
	MPW1PW91	1990(902), 2124(49), 2140(300), 2169(1720), 2169(2570)
	HF	2124(49), 2157(1620), 2348(509), 2372(1640), 2392(4390)
$\text{Rh}_6(\text{CO})_8^-$ ($\mu\text{-CO}$) ₄	Exp ⁵⁴	1819m, 2045w, 2075s
(O_h)	BP86	1801(667), 2014(96), 2049(2400)
	B3LYP	1884(834), 2100(56), 2134(2860)
	MPW1PW91	1921(864), 2142(55), 2177(2860)
	HF	2122(1420), 2364(129), 2395(3730)

fall in the range from 1.140 to 1.194 Å. Again the MPW1PW91 predictions are in good agreement with the experimental results.

3.6. Infrared Frequencies. Table 11 compares the vibrational frequencies found experimentally for the second-row metal carbonyl derivatives with the infrared-active harmonic vibrational frequencies calculated using the four different methods discussed in this paper. For $\text{Mo}(\text{CO})_6$, where more experimental data are available, all of the infrared active frequencies are given, whereas for the other metal carbonyl derivatives only the $\nu(\text{CO})$ frequencies are given.

The data in Table 11 clearly indicate that the BP86 functional is by far the best for predicting infrared frequencies, as was found earlier for the first-row transition-metal derivatives including $\text{Co}_4(\text{CO})_{12}$ and $\text{Co}_6(\text{CO})_{16}$ analogous to the rhodium carbonyl derivatives studied in this paper.^{74–78} The MPW1PW91 method, which is the most effective for predicting molecular geometries, is not even as effective for predicting $\nu(\text{CO})$ frequencies as the B3LYP method. The Hartree–Fock method predicts $\nu(\text{CO})$ frequencies so far from the experimental values as to be of limited value.

4. Concluding Remarks

We report here a systematic comparison between theory and experiment for the six homoleptic second-row transition-metal carbonyls whose structures have been determined by X-ray crystallography. A total of 26 bond distances have been predicted with each of four theoretical methods. The average errors for the different methods are 0.030 Å (BP86), 0.028 Å (B3LYP), 0.018 Å (MPW1PW91), and 0.050 Å (Hartree–Fock).

It is clear that the new generation DFT method MPW1PW91 is superior to earlier established methods for predicting the structures of these homoleptic second-row transition-metal carbonyls. However, the BP86 method, which is effective for predicting the infrared spectra of first-row transition-metal carbonyl derivatives, is found to be more effective than not only the MPW1PW91 but also the B3LYP method for predicting the infrared spectra of second-row transition-metal derivatives.

Acknowledgment. We appreciate the support of the China Scholarship Council (CSC: No. 2005A46003). National Science Foundation (U.S.A.) Grants No. CHE-0209857 and CHE-0451445 are similarly acknowledged.

Supporting Information Available: Harmonic vibrational frequencies (in cm^{-1}) and their IR intensities (in km/mol) for $\text{Mo}(\text{CO})_6$, $\text{Tc}_2(\text{CO})_{10}$, $\text{Ru}(\text{CO})_5$, two isomers of $\text{Ru}_3(\text{CO})_{12}$ and $\text{Ru}_4(\text{CO})_{12}$, and $\text{Rh}_6(\text{CO})_{16}$ (Table S1–S8). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) (a) Hawkins, N. J.; Matraw, H. C.; Sabol, W. W.; Carpenter, D. R. *J. Chem. Phys.* **1955**, *23*, 2422. (b) Jones, L. H. *J. Chem. Phys.* **1962**, *36*, 2375.
- (2) Arnesen, S.; Seip, H. M. *Acta Chem. Scand.* **1966**, *20*, 2711.
- (3) Mak, T. C. W. *Z. Kristallogr.* **1984**, *166*, 277.
- (4) (a) Hileman, J. C.; Huggins, D. K.; Kaesz, H. D. *J. Am. Chem. Soc.* **1961**, *83*, 2953. (b) Wallach, D. *Acta Crystallogr.* **1982**, *15*, 1058.
- (5) Bailey, M. F.; Dahl, L. F. *Inorg. Chem.* **1965**, *4*, 1140.
- (6) Huang, J.; Hedberg, K.; Davis, H. B.; Pomeroy, R. K. *Inorg. Chem.* **1990**, *29*, 3923.
- (7) Churchill, M. R.; Hollander, F. J.; Hutchinson, J. P. *Inorg. Chem.* **1977**, *16*, 2655.
- (8) Delley, B.; Manning, M. C.; Ellis, D. E.; Berkowitz, J. *Inorg. Chem.* **1982**, *21*, 2247.
- (9) Hieber, W.; Lagally, H. *Z. Anorg. Allg. Chem.* **1943**, *251*, 96.
- (10) Wei, C. H.; Wilkes, G. R.; Dahl, L. F. *J. Am. Chem. Soc.* **1967**, *89*, 4792.
- (11) Wei, C. H. *Inorg. Chem.* **1969**, *8*, 2384.
- (12) Corey, E. R.; Dahl, L. F.; Beck, W. *J. Am. Chem. Soc.* **1963**, *85*, 1202.
- (13) Heaton, B. T.; Jacob, C.; Podkorytov, I. S.; Tunik, S. P. *Inorg. Chim. Acta* **2006**, *359*, 3557.
- (14) Heaton, B. T.; Sabounchei, J.; Kernaghan, S.; Nakayama, H.; Eguchi, T.; Takeda, S.; Nakamura, N.; Chihara, A. *Bull. Chem. Soc. Jpn.* **1990**, *63*, 3019.
- (15) Evans, J.; Johnson, B. F. G.; Lewis, J.; Norton, J. R.; Cotton, F. A. *J. Chem. Soc., Chem. Commun.* **1973**, 807.
- (16) Evans, J.; Johnson, B. F. G.; Lewis, J.; Matheson, T. W.; Norton, J. R. *J. Chem. Soc., Dalton Trans.* **1978**, 626.
- (17) Besancon, K.; Laurency, G.; Lumini, T.; Roulet, R.; Bruyndonckx, R.; Daul, C. *Inorg. Chem.* **1968**, *37*, 5634.
- (18) Farrar, D. H.; Grachova, E. V.; Lough, A.; Patirana, C.; Poë, A. J.; Tunik, S. P. *J. Chem. Soc., Dalton Trans.* **2001**, 2015.
- (19) Walter, T. H.; Reven, L.; Oldfield, E. *J. Phys. Chem.* **1989**, *93*, 1320.
- (20) Salzmann, R.; Kaupp, M.; McMahon, M. T.; Oldfield, E. *J. Am. Chem. Soc.* **1998**, *120*, 4771.
- (21) Wei, C. H.; Dahl, L. F. *J. Am. Chem. Soc.* **1966**, *88*, 1821.
- (22) Ehlers, A. W.; Frenking, G. *J. Am. Chem. Soc.* **1994**, *116*, 1514.
- (23) Delly, B.; Wrinn, M.; Lüthi, H. P. *J. Chem. Phys.* **1994**, *100*, 5785.
- (24) Li, J.; Schreckenbach, G.; Ziegler, T. *J. Am. Chem. Soc.* **1995**, *117*, 486.
- (25) Jonas, V.; Thiel, W. *J. Phys. Chem.* **1995**, *102*, 8474.
- (26) Kaup, M.; Malkin, V. G.; Maklina, O. L.; Salahub, D. R. *Chem. Eur. J.* **1996**, *2*, 24.
- (27) Barckholtz, T. A.; Bursten, B. E. *J. Am. Chem. Soc.* **1998**, *120*, 1926.
- (28) Jemmis, E. D.; Giju, K. T. *J. Am. Chem. Soc.* **1998**, *120*, 6952.
- (29) Niu, S.; Hall, M. B. *Chem. Rev.* **2000**, *100*, 353.
- (30) Cotton, F. A.; Gruhn, N. E.; Gu, J.; Huang, P.; Lichtenberger, D. L.; Murillo, C. A.; Van Dorn, L. O.; Wilkinson, C. C. *Science* **2002**, *298*, 1971.
- (31) Macchi, P.; Sironi, A. *Coord. Chem. Rev.* **2003**, *100*, 353.
- (32) Siegbahn, P. E. M. *J. Am. Chem. Soc.*, **2005**, *127*, 17303.
- (33) Ziegler, T.; Autschbach, J. *Chem. Rev.* **2005**, *105*, 2695.
- (34) Mota, A. J.; Dedieu, A.; Bour, C.; Suffert, J. *J. Am. Chem. Soc.* **2005**, *127*, 7171.
- (35) Bühl, M.; Kabrede, H. *J. Chem. Theory Comput.* **2006**, *2*, 1282.
- (36) Brynda, M.; Gagliardi, L.; Wimark, P. O.; Power, P. P.; Roos, B. O. *Angew. Chem., Int. Ed.* **2006**, *45*, 3804.
- (37) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *124*, 224105.
- (38) Harvey, J. N. *Ann. Rep. Prog. Chem. Sect. C* **2006**, *102*, 203.

- (39) Strickland, N. S.; Harvey, J. N. *J. Phys. Chem. B* **2007**, *111*, 841.
- (40) Ignatyev, I.; King, R. B.; Schaefer, H. F., III *J. Am. Chem. Soc.* **2000**, *122*, 1989.
- (41) Xie, Y.; King, R. B.; Schaefer, H. F., III *J. Am. Chem. Soc.* **2000**, *122*, 8746.
- (42) Li, Q.; Liu, Y.; Xie, Y.; King, R. B.; Schaefer, H. F., III *Inorg. Chem.* **2001**, *40*, 5842.
- (43) King, R. B.; Schaefer, H. F., III *Pure Appl. Chem.* **2001**, *73*, 1059.
- (44) Kenny, J.; King, R. B.; Schaefer, H. F., III *Inorg. Chem.* **2001**, *40*, 900.
- (45) Richardson, N. R.; Xie, Y.; King, R. B.; Schaefer, H. F., III *J. Phys. Chem. A* **2001**, *105*, 11134.
- (46) Li, S.; Richardson, N. R.; Xie, Y.; King, R. B.; Schaefer, H. F., III *Faraday Discuss.* **2003**, *124*, 315.
- (47) Li, S.; Richardson, N. R.; Xie, Y.; King, R. B.; Schaefer, H. F., III *J. Phys. Chem. A* **2003**, *107*, 10118.
- (48) Fuentealba, P.; Preuss, H.; Stoll, H.; v. Szentpaly, L. *Chem. Phys. Lett.* **1989**, *89*, 418.
- (49) v. Szentpaly, L.; Fuentealba, P.; Preuss, H.; Stoll, H. *Chem. Phys. Lett.* **1982**, *93*, 555.
- (50) Fuentealba, P.; Stoll, H.; v. Szentpaly, L.; Schwerdtfeger, P.; Preuss, H. *J. Phys. B* **1983**, *16*, 1323.
- (51) Stoll, H.; Fuentealba, P.; Schwerdtfeger, P.; Flad, J.; v. Szentpaly, L.; Preuss, H. *J. Chem. Phys.* **1984**, *81*, 2732.
- (52) Van Wüllen, C. *Int. J. Quantum Chem.* **1996**, *58*, 147.
- (53) Hunstock, E.; Mealli, C.; Calhorda, M.; Reinhold, J. *Inorg. Chem.* **1999**, *38*, 5053.
- (54) Allian, A. D.; Wang, Y.; Saeys, M.; Kuramshina, G. M.; Garland, M. *Vib. Spectrosc.* **2006**, *41*, 101.
- (55) Cotton, F. A.; Gu, J.; Murillo, C. A.; Timmons, D. J. *J. Am. Chem. Soc.* **1998**, *120*, 13280.
- (56) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (57) Lee, C.; Yang, W. W.; Parr, R. G. *Phys. Rev.* **1988**, *B37*, 785.
- (58) (a) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098. (b) Perdew, J. P. *Phys. Rev.* **1986**, *B33*, 8822.
- (59) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.
- (60) Zhao, S.; Wang, W.; Li, Z.; Liu, Z. P.; Fan, K.; Xie, Y.; Schaefer, H. F., III *J. Chem. Phys.* **2006**, *124*, 184102.
- (61) Dolg, M.; Stoll, H.; Preuss, H. *Theor. Chim. Acta* **1993**, *85*, 441.
- (62) Bergner, A.; Dolg, M.; Kuechle, W.; Stoll, H.; Preuss, H. *Mol. Phys.* **1993**, *80*, 1431.
- (63) Dunning, T. H. *J. Chem. Phys.* **1970**, *53*, 2823.
- (64) Huzinaga, S. *J. Chem. Phys.* **1965**, *42*, 1293.
- (65) 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.
- (66) Hunstock, E.; Mealli, C.; Calhorda, M. J.; Reinhold, J. *Inorg. Chem.* **1999**, *38*, 5053.
- (67) Aime, S.; Dastrù, W.; Gobetto, R.; Krause, J.; Milone, L. *Organometallics* **1995**, *14*, 4435.
- (68) Farrugia, L. J. *J. Cluster Sci.* **2000**, *11*, 39.
- (69) Wade, K. *Chem. Commun. (Cambridge)* **1971**, 792.
- (70) Wade, K. *Adv. Inorg. Chem. Radiochem.* **1976**, *18*, 1.
- (71) Mingos, D. M. P. *Nat. Phys. Sci.* **1972**, *99*, 236.
- (72) Mingos, D. M. P. *Acc. Chem. Res.* **1984**, *17*, 311.
- (73) Farrar, D. H.; Grachova, E. V.; Lough, A.; Patirana, C.; Poë, A. J.; Tunik, S. P. *J. Chem. Soc., Dalton Trans.* **2001**, 2015.
- (74) Xie, Y.; King, R. B.; Schaefer, H. F., III *Spectrochem. Acta* **2005**, *A61*, 1693.
- (75) Jones, L. H.; McDowell, R. S.; Goldblatt, M. *Inorg. Chem.* **1969**, *8*, 2349.
- (76) Michels, G. D.; Svec, H. J. *Inorg. Chem.* **1981**, *20*, 3445.
- (77) Gregory, M. P.; Poliakov, M.; Turner, J. J. *J. Mol. Struct.* **1985**, *127*, 247.
- (78) Battiston, G. A.; Bor, G.; Dietler, U. K.; Kettle, S. F. A.; Rossetti, R.; Sbrignadello, G.; Stanghellini, P. L. *Inorg. Chem.* **1980**, *19*, 1961.

JCTC Journal of Chemical Theory and Computation

Sparkle/PM3 Parameters for the Modeling of Neodymium(III), Promethium(III), and Samarium(III) Complexes

Ricardo O. Freire,[†] Nivan B. da Costa, Jr.,[‡] Gerd B. Rocha,[§] and Alfredo M. Simas^{*,†}

Departamento de Química Fundamental, CCEN, UFPE, 50590-470 - Recife, PE, Brazil, Departamento de Química, CCET, UFS, 49100-000 - Aracaju, SE, Brazil, and Departamento de Química, CCEN, UFPB, 58.059-970 - João Pessoa, PB, Brazil

Received November 4, 2006

Abstract: The Sparkle/PM3 model is extended to neodymium(III), promethium(III), and samarium(III) complexes. The unsigned mean error, for all Sparkle/PM3 interatomic distances between the trivalent lanthanide ion and the ligand atoms of the first sphere of coordination, is 0.074 Å for Nd(III); 0.057 Å for Pm(III); and 0.075 Å for Sm(III). These figures are similar to the Sparkle/AM1 ones of 0.076 Å, 0.059 Å, and 0.075 Å, respectively, indicating they are all comparable models. Moreover, their accuracy is similar to what can be obtained by present-day ab initio effective potential calculations on such lanthanide complexes. Hence, the choice of which model to utilize will depend on the assessment of the effect of either AM1 or PM3 on the quantum chemical description of the organic ligands. Finally, we present a preliminary attempt to verify the geometry prediction consistency of Sparkle/PM3. Since lanthanide complexes are usually flexible, we randomly generated 200 different input geometries for the samarium complex QIPQOV which were then fully optimized by Sparkle/PM3. A trend appeared in that, on average, the lower the total energy of the local minima found, the lower the unsigned mean errors, and the higher the accuracy of the model. These preliminary results do indicate that attempting to find, with Sparkle/PM3, a global minimum for the geometry of a given complex, with the understanding that it will tend to be closer to the experimental geometry, appears to be warranted. Therefore, the sparkle model is seemingly a trustworthy semiempirical quantum chemical model for the prediction of lanthanide complexes geometries.

Introduction

Recently, we introduced Sparkle/AM1,¹ a new paradigm for lanthanide complexes semiempirical calculations, at a level of accuracy useful for coordination compounds design. And, subsequently, we presented Sparkle/AM1 parameters for neodymium(III),² promethium(III), and samarium(III).³ Recent research on lanthanide complexes has indeed indicated that Sparkle/AM1 coordination polyhedron geometries are comparable to, if not better than, geometries obtained with

the best contemporary ab initio calculations with effective core potentials (ab initio/ECP) on complexes of a size large enough to be of value to practical use.^{1,3,4} Besides, sparkle model calculations are hundreds of times faster.

Sparkle/AM1 lanthanides function as new elements to the semiempirical molecular orbital model AM1.⁵ That is, when a lanthanide complex is calculated, the lanthanide ion is modeled as a sparkle, whereas the ligands are modeled by AM1.

Another very popular semiempirical model is PM3,^{6,7} which mainly gives enthalpies of formation with lower average errors than AM1. PM3 is presently available in a variety of quantum chemical softwares, both commercial and noncommercial.^{8–17} The usefulness of PM3 has been recently

* Corresponding author phone: +55 81 2126-8447; fax: +55 81 2126-8442; e-mail: simas@ufpe.br.

[†] Departamento de Química Fundamental, UFPE.

[‡] Departamento de Química, UFS.

[§] Departamento de Química, UFPB.

expanded due to the availability of parameters for many elements, such as for the transition metals,¹⁸ for sodium,¹⁹ and for all nonradioactive elements of the main group, excluding the noble gases.²⁰ Specific parameters for some types of chemical interactions are also available, such as the parameters for zinc for the calculation of metalloenzyme active sites,¹⁹ or the parameter set to describe iron-sulfur proteins.²¹

Novel lanthanide complexes of neodymium^{22–27} and samarium^{28–32} are always emerging, and new applications are frequently been reported.^{33–37} Promethium, on the other hand, does not have any stable isotopes. However, a few of the unstable isotopes, mainly ¹⁴⁷Pm and ¹⁴⁹Pm, find a variety of applications, mainly in medicine.^{38–40} Recently, PM3 semiempirical calculations were carried out on the ligands of lanthanide(III) double decker complexes, illuminating the role of ligand substituents on the electrochemical properties of such complexes.⁴¹ However, calculations were not performed on the complexes themselves, due to a lack of parameters for the lanthanide ions involved.

Therefore, in order to expand the bounds of applications of our sparkle model, we advance, in this paper, Sparkle/PM3 parameters for Nd(III), Pm(III), and Sm(III) ions. We further present a preliminary attempt to attest the geometry prediction coherence of Sparkle/PM3.

The Sparkle Model

Modeling lanthanide complexes is challenging because the ions lack stereochemical preference, possess a handful of high coordination numbers, and display small energy variations among their various coordination geometries.

The Sparkle model recognizes the contracted nature of the 4-f orbitals of the lanthanide trications, of electronic configuration [Xe]4fⁿ, coexisting with a poor overlap with the orbitals of the ligands, which assigns a predominantly ionic character to organolanthanide complexes.⁴² Accordingly, the angular effects of the f orbitals are shielded from external perturbations by the filled 5s² and 5p⁶ orbitals and are not taken into account. As such, the sparkle model regards the lanthanide trications as triple positively charged closed shell inert gas electron densities without any angular steric properties. The sparkle model thus replaces the trivalent rare earth ion by a Coulombic charge of +3e superimposed to a repulsive exponential potential of the form exp(-αr), which accounts for the size of the ion; provides three electrons to the orbitals of the ligands; adds two Gaussian functions to the core-core repulsion energy term; and includes the lanthanide atomic mass.

Parametrization

The parametrization procedure used for obtaining the Sparkle/PM3 parameters for Nd(III), Pm(III), and Sm(III) was essentially the same as the one of our previous works.^{2,3} Accordingly, for neodymium and samarium, we only used high quality crystallographic structures (R-factor <5%) of complexes taken from the “Cambridge Structural Database 2004” (CSD),^{43–45} having found a total of 57 structures of complexes of Nd(III) and 42 of Sm(III). Thus, as training sets for Nd(III) and Sm(III), we used the same two subsets

of 15 complexes each, already chosen for the Sparkle/AM1 parametrization for the same ions, and presented in Figure 1 of the article on Nd(III)² and in Figure 2 of the Sm(III) article.³

Again, since there are no crystallographic structures of promethium coordination compounds available in CSD, we followed the same procedure as for Sparkle/AM1: (i) we picked for both training and validation the same set of 15 representative samarium complexes previously chosen;³ (ii) we then replaced samarium with promethium; and (iii) fully optimized the geometries with RHF/STO-3G/ECP using the quasirelativistic MWB ECP of ref 46. We defined a special code for the promethium parametrization set which we presented in Figure 9 of the Pm(III) article.³ XILGOO{Pm}, for example, would be the samarium complex of CSD code XILGOO with promethium instead of samarium. And we used RHF/STO-3G/ECP because, apparently, this is the most efficient model chemistry in terms of coordination polyhedron crystallographic geometry predictions from isolated lanthanide complex ion calculations, as repeatedly reported.^{1,3,4}

The Sparkle/PM3 parameters found for the three lanthanide ions are shown in Table 1.

Table 1. Parameters for the Sparkle/PM3 Model for the Nd(III), Pm(III), and Sm(III) Ions

	Sparkle/PM3		
	Nd(III)	Pm(III)	Sm(III)
GSS	57.4944898977	59.2924444913	54.8086404668
ALP	4.7057677595	3.1490918074	3.6813938335
a ₁	1.0715972265	1.6572814674	0.7706615984
b ₁	6.9565346287	9.2529413759	6.6020324700
c ₁	1.7812099249	1.7412637448	1.7636673188
a ₂	0.0886417116	0.1851223683	0.0936188340
b ₂	10.8664473398	7.4186533283	9.3136737687
c ₂	3.0992613820	3.0623727738	2.9879390071
EHEAT ^a (kcal·mol ⁻¹)	962.8	976.9	974.4
AMS (amu)	144.2400	145.0000	150.3600

^a The heat of formation of the Nd(III), Pm(III), and Sm(III) ions in Sparkle/PM3 and Sparkle/AM1 models was obtained by adding to the heat of atomization of each respective lanthanide their first three ionization potentials.

Validation

Unlike ab initio model chemistries, semiempirical ones do not have strong theorems behind them. As such, their validation as useful tools must be established statistically.

Accordingly, as geometry accuracy measures, we used the average unsigned mean error for each complex *i*, UME_{*i*}, defined as

$$\text{UME}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} |R_{ij}^{\text{CSD}} - R_{ij}^{\text{calc}}| \quad (1)$$

where *n_i* is the number of ligand atoms directly coordinating the lanthanide ion. Two cases have been examined: (i) UME_{(Ln-L)_s} involving only the interatomic distances *R_j* between the lanthanide central ion, Ln, and the atoms of the coordination polyhedron, L, important to luminescent complex design and (ii) UMEs of the interatomic distances *R_j*

between the lanthanide central ion and the atoms of the coordination polyhedron as well as all the interatomic distances R_j between all atoms of the coordination polyhedron. Tables 1S–3S of the Supporting Information present the $UME_{(Ln-L)}$ s and UMEs and for both Sparkle/PM3 and Sparkle/AM1 for Nd(III), Pm(III), and Sm(III), respectively.

Assuming that the sparkle model is a well founded representation of the lanthanide ions as well as of their interactions with the ligands the distribution of these UMEs should be random around a mean, whose value can be used as a measure of the accuracy of the model. Since the UMEs are positive, defined in the domain $(0, \infty)$, they should follow the gamma distribution which has the probability density function $f(x; k, \theta)$

$$f(x; k, \theta) = x^{k-1} \frac{e^{-x/\theta}}{\theta^k \Gamma(k)} \quad (2)$$

where $x > 0$ stands for the UMEs, $k > 0$ is the shape parameter, $\theta > 0$ is the scale parameter of the gamma distribution, and $\Gamma(k)$ is the gamma function of k . The expected value of the gamma distribution is simply $k\theta$.

The gamma distribution fits of the UME data were obtained by estimates of the shape and scale parameters by the method of maximum likelihood.

The quality of a gamma distribution fit can be assessed via the one-sample nonparametric Kolmogorov-Smirnov test⁴⁷ in order to verify statistically whether the distribution of the UME values is really a gamma distribution indexed by the estimated parameters. In this case, the null hypothesis is that the UME values do follow that gamma distribution. In order for the null hypothesis not to be rejected at the usual level of 5%, the p -value of the test must thus be larger than 0.05. And the higher the p -value, whose maximum possible value is 1, the higher the probability that the UMEs are random, the more the sparkle model captured the deterministic aspects of the problem, and the more justifiable is the use of the statistical tools employed here.

If the p -value is indeed larger than 0.05, then one can compute, from the gamma distribution fit, the probability of the UME, for an arbitrary lanthanide complex, to belong to an interval.

We now examine results for both the already published Sparkle/AM1 model for Nd(III), Pm(III), and Sm(III) as well as for the Sparkle/PM3 model being presented in this article for the same lanthanide ions.

Figure 1a presents a gamma distribution fit of the $UME_{(Ln-L)}$ data for the Sparkle/PM3 model for Nd(III). As indicated in the figure, the p -value is 0.667, thus indicating that the UMEs are indeed significantly randomly distributed around the mean and correctly follow a gamma distribution. Figure 1b shows the gamma distribution fit for the Nd(III) Sparkle/AM1 model, with a p -value of 0.704. We also superimposed to the gamma distribution fits, histograms of the actual data—the number of bars in each being chosen to best adjust the histogram to the curve obtained from the fit—in order to simply give a pictorial idea of where and how the actual UMEs occurred.

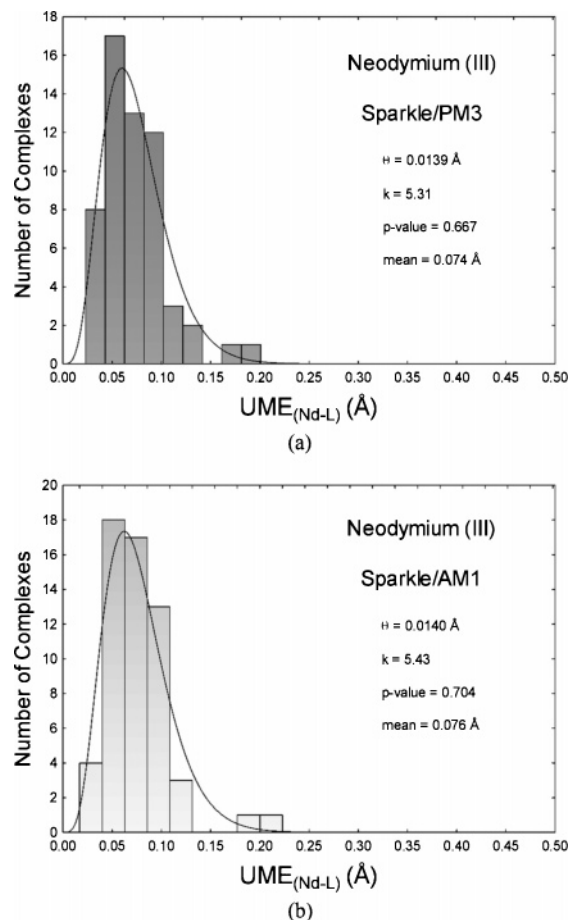


Figure 1. Probability densities of the Gamma distribution fits of the $UME_{(Ln-L)}$ s for the Nd(III) Sparkle/PM3 and Sparkle/AM1 models, superimposed to histograms of the same data for all 57 Nd(III) complexes considered; where k is the shape parameter, and θ is the scale parameter of the gamma distribution; the p -value is a measure of the significance of the gamma distribution fit; and mean is the expected value of the fitted gamma distribution, which is set to be equal to the arithmetic mean value of the 57 $UME_{(Ln-L)}$ s.

Figures 2 and 3 present, each, a gamma distribution fit of the respective $UME_{(Ln-L)}$ for the present Sparkle/PM3 as well as for previously published Sparkle/AM1 models for promethium and samarium. In all cases the respective p -values were well above the critical value of 0.05, ranging from 0.911 to 0.968, thus validating the usage of the sparkle model for both PM3 and AM1 for the prediction of lanthanide complexes geometries.

Equivalent analysis for the whole UMEs, with similar conclusions, can be found in Figures S1–S3 of the Supporting Information.

Comparison with *ab Initio*/ECP Calculations

Repeated studies by our research group^{1,3,4} have confirmed the unanticipated fact that RHF/STO-3G/ECP appears to be the most efficient model chemistry in terms of coordination polyhedron crystallographic geometry predictions from isolated lanthanide complex ion calculations. Contrary to what would normally be expected, either an increase in the basis set or inclusion of electron correlation, or both, consistently

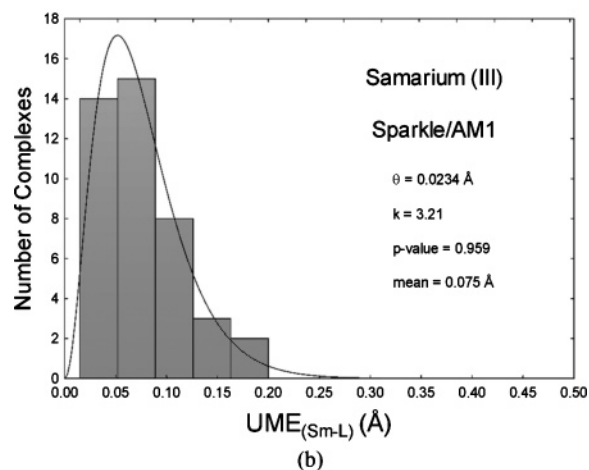
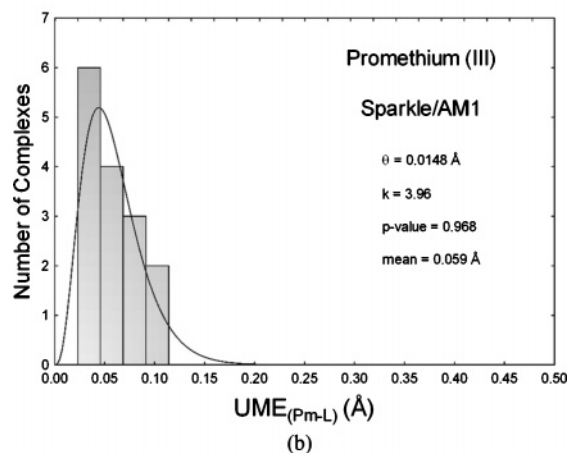
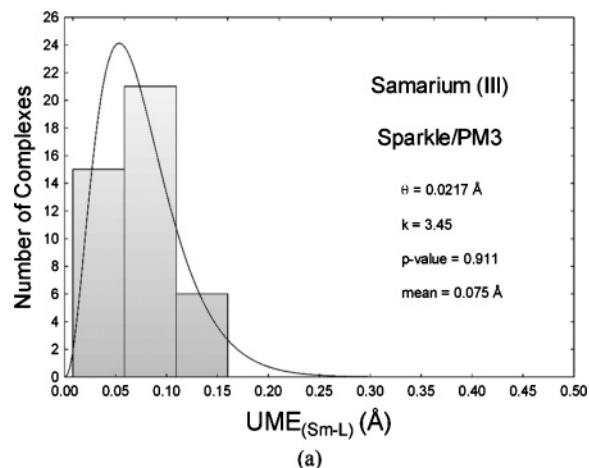
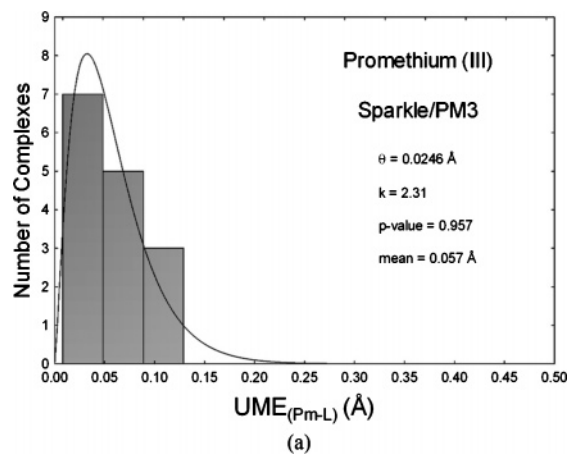


Figure 2. Probability densities of the Gamma distribution fits of the $UME_{(Ln-L)}$ s for the Pm(III) Sparkle/PM3 and Sparkle/AM1 models, superimposed to histograms of the same data for all 15 Pm(III) complexes considered; where k is the shape parameter, and θ is the scale parameter of the gamma distribution; the p -value is a measure of the significance of the gamma distribution fit; and mean is the expected value of the fitted gamma distribution, which is set to be equal to the arithmetic mean value of the 15 $UME_{(Ln-L)}$ s.

augmented the deviations and impaired the quality of the predicted coordination polyhedron geometries.

In the Sm(III) Sparkle/AM1 paper,³ we presented RHF/STO-3G/ECP full geometry optimizations of seven representative samarium complexes of known crystallographic geometries, including a disamarium, of CSD code MEWGOQ. We are therefore in a position to compare the accuracy of Sparkle/PM3 with Sparkle/AM1 and RHF/STO-3G/ECP for the same set.

Figure 4 presents the average $UME_{(Sm-L)}$ and UME values for Sparkle/PM3, Sparkle/AM1, and RHF/STO-3G/ECP full geometry optimizations of the complexes presented in Figure 7 of ref 3. Clearly, all three model chemistries are comparable, with Sparkle/AM1 being on average slightly superior to both Sparkle/PM3 and to RHF/STO-3G/ECP.

Since there are no reports of ab initio full geometry optimization of neodymium(III) complexes, we chose seven of these complexes to have their geometries fully optimized with the model chemistry RHF/STO-3G/ECP. The chosen complexes, shown in Figure 5, were selected to be representative of the various classes of ligands (β -diketones,

Figure 3. Probability densities of the Gamma distribution fits of the $UME_{(Ln-L)}$ s for the Sm(III) Sparkle/PM3 and Sparkle/AM1 models, superimposed to histograms of the same data for all 42 Sm(III) complexes considered; where k is the shape parameter, and θ is the scale parameter of the gamma distribution; the p -value is a measure of the significance of the gamma distribution fit; and mean is the expected value of the fitted gamma distribution, which is set to be equal to the arithmetic mean value of the 42 $UME_{(Ln-L)}$ s.

nitrate, monodentates, bidentates, tridentates, polydentates, and dilanthanides) present in the validation set.

Figure 6 presents the average $UME_{(Nd-L)}$ and UME values for Sparkle/PM3, Sparkle/AM1, and RHF/STO-3G/ECP full geometry optimizations of the complexes presented in Figure 5. Clearly, all three model chemistries are comparable, this time, with Sparkle/PM3 being on average slightly superior to both Sparkle/AM1 and to RHF/STO-3G/ECP. Indeed, $UME_{(Nd-L)}$ s of the order of 0.056 Å are small enough to be useful to the luminescent neodymium complex design.

Geometry Prediction Coherence

The variational theorem applies to Sparkle model calculations simply because the semiempirical models AM1 and PM3 retain the algebraic structure of the Hartree–Fock method. However, since the Sparkle model parameters are the result of a sophisticated fit of experimental values, in principle there are no guarantees that a minimized sparkle model geometry will function as an estimate of the true experimental

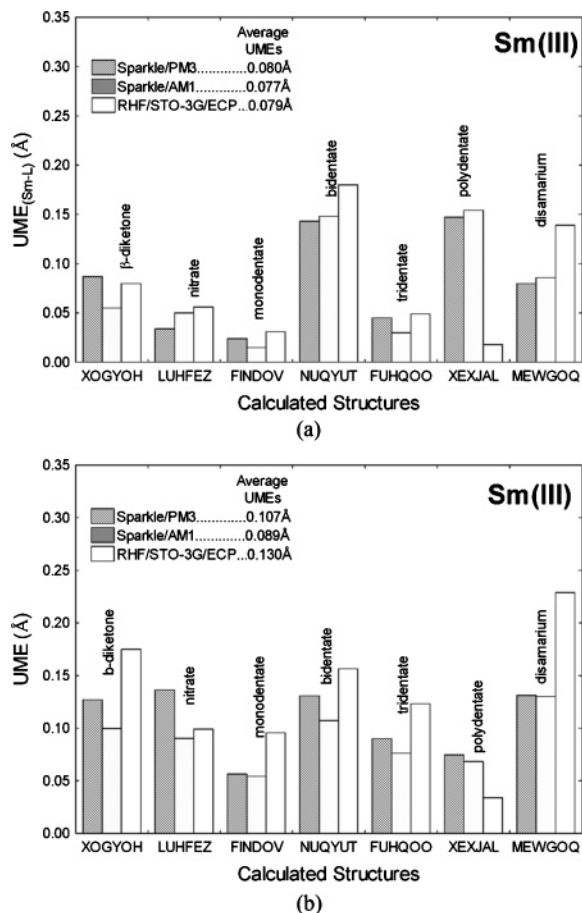


Figure 4. Unsigned mean errors, $UME_{(Ln-L)}$ s (in Å), between the samarium central ion and the atoms of the coordination polyhedron and between the samarium central ion and the atoms of the coordination polyhedron as well as all the interatomic distances R_i between all atoms of the coordination polyhedron, obtained from Sparkle/PM3, Sparkle/AM1, and ab initio RHF/STO-3G/ECP calculations of the ground-state geometries, for each of the representative Sm(III) complexes, identified by their respective Cambridge Structural Database 2004 codes.

geometry. By extension, in principle, there are also no guarantees that a sparkle model global minimum geometry will be closer to the experimental geometry than any of the other sparkle model local minima found.

On the other hand, the sparkle model was parametrized from experimental geometries as input data, and its parameters have been thoroughly minimized to lead to optimized geometries with the lowest deviations possible from the starting geometries. Thus, we can reasonably conjecture that the sparkle model optimized geometry that can be obtained starting from the experimental geometry should be the sparkle model global minimum or at least very close to it. And if this conjecture is true, then trying to find the geometry corresponding to the global minimum in the sparkle model nuclear potential energy hypersurface for an arbitrary complex whose experimental geometry is unknown is a legitimate procedure.

In order to verify in a preliminary manner this conjecture, a samarium(III) complex, of CSD code QIPQOV, was selected as a case study.

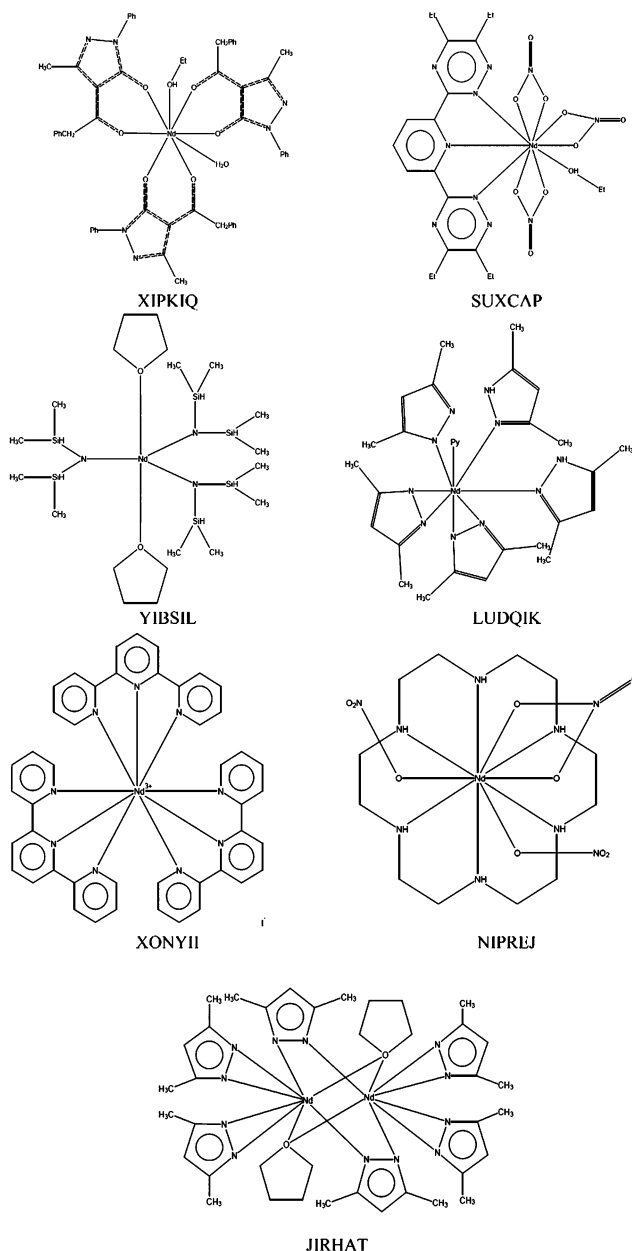


Figure 5. Schematic two-dimensional representation of the structures of neodymium(III) complexes, used for comparison between ab initio model geometries and their crystallographic counterparts, identified by their respective Cambridge Structural Database 2004. The ab initio calculations have been performed using the Hartree–Fock method with the STO-3G basis set for all atoms, except for the neodymium(III) ion, in which case we used the quasirelativistic ECP of ref 46.

We then generated 200 different input geometries for this complex. Each of the geometries resulted from the application of a procedure to each and every one of its ligands in an independent manner. In this procedure, the ligands are considered to be rigid and independent of each other and of the central samarium ion. Starting with the experimental geometry, for each ligand we proceeded as follows: (i) we defined a randomly oriented Cartesian coordinate system whose origin is located at the center of mass of the ligand; (ii) we then randomly chose one of the three axes of this Cartesian coordinate system; (iii) we rotated the ligand

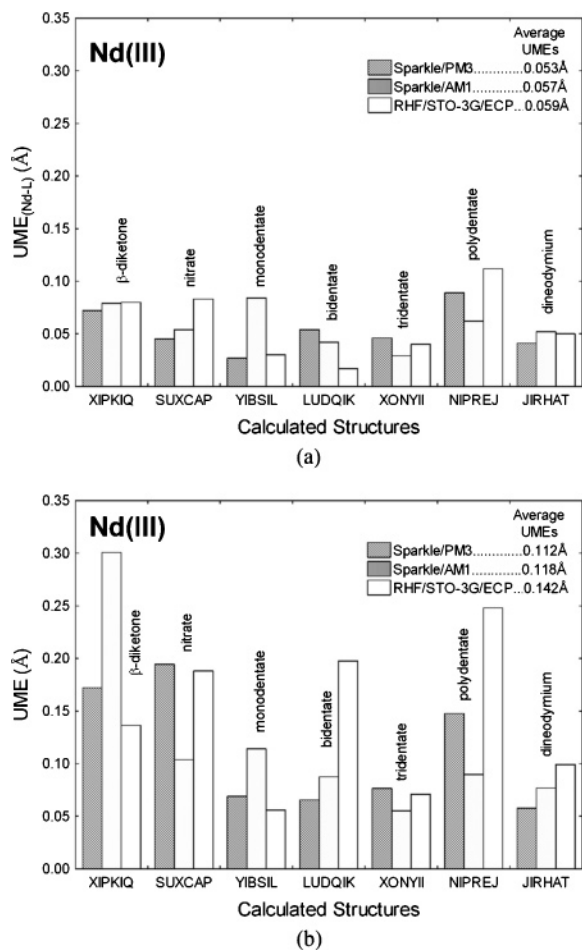


Figure 6. Unsigned mean errors, $UME_{(Ln-L)}$ s (in Å), between the neodymium central ion and the atoms of the coordination polyhedron and between the neodymium central ion and the atoms of the coordination polyhedron as well as all the interatomic distances R_i between all atoms of the coordination polyhedron, obtained from Sparkle/PM3, Sparkle/AM1, and ab initio RHF/STO-3G/ECP calculations of the ground-state geometries, for each of the representative Nd(III) complexes, identified by their respective Cambridge Structural Database 2004 codes.

around this chosen axis by a random angle belonging to the interval $[+30^\circ, -30^\circ]$; (iv) subsequently, one of the atoms of the ligand, we called atom R, was randomly chosen to define the axis connecting it to the samarium ion; and (v) a random translation, in the direction of this axis, was finally applied to the whole ligand—the magnitude of this translation belonging to the interval $[-15\%, +15\%]$ of the interatomic distance between the samarium ion and atom R of the ligand.

For each of the 200 different input geometries, we performed a full Sparkle/PM3 geometry optimization.

For some of the input geometries, the starting distances of the originally coordinating atoms were so far away from the samarium atom that the corresponding Sparkle/PM3 geometry optimizations converged to one or more uncoordinated ligands. A total of 45 of the outputs were then discarded for this reason.

First, consider $UME_{(Ln-L)}$ s as accuracy measure. As can be clearly seen in Figure 7a, the 155 remaining output geometries grouped into 5 clusters. The cluster with the

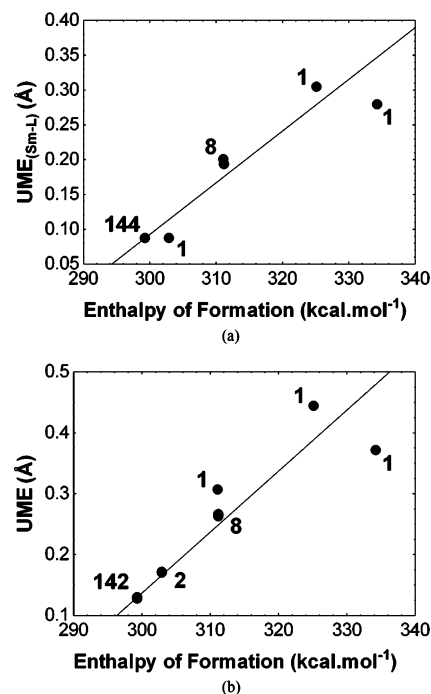


Figure 7. Clusters of output geometries obtained from Sparkle/PM3 full geometry optimizations of random input geometries for the samarium complex of CSD code QIPQOV, showing that the group of clusters with the highest enthalpies of formation was also the group of clusters with the highest $UME_{(Sm-L)}$ s and UME. The number of optimized geometries comprising each group of clusters is also shown. The trendline is present just to guide the eye.

lowest energy also has 144 essentially identical output geometries. And this cluster also has the smallest value of $UME_{(Ln-L)}$. And considering that 200 random initial geometries is a reasonable number, likely the geometry corresponding to the 144 outputs cluster is also the global minimum.

Now consider UMEs as the accuracy measure. Figure 7b shows that in this case the geometries grouped into six clusters, the one with the lowest energy, with 142 essentially identical output geometries, being also the one with the lowest UME.

The trendlines in both parts a and b of Figure 7 are present just to guide the eye and to make it easy to confirm that trying to find the geometry corresponding to the global minimum, in the sparkle model nuclear potential energy hypersurface for an arbitrary complex whose experimental geometry is unknown, is warranted.

Moreover, apparently, it is not hard to find the geometry corresponding to the global minimum of this lanthanide complex. Indeed, out of 155 attempts, 144 arrived at the $UME_{(Sm-L)}$ minimum, a 93% chance; and 142 arrived at the UME minimum, a 92% chance. So, at least in this case, a few different input geometries should be enough to arrive to the global minimum with a high degree of certainty.

Conclusion

Sparkle/PM3 presents a level of accuracy equivalent to Sparkle/AM1 and also to the most accurate ab initio full

geometry optimization calculations that can be nowadays carried out on complexes of a size large enough to be of relevance to complex design.

Besides, both Sparkle/PM3 and Sparkle/AM1 seem to have captured the deterministic aspects involved in the prediction of the geometries of the complexes, as indicated by the statistically significant gamma distribution fits of the unsigned mean errors data.

The preliminary results presented in this article unveiled a significant trend: that more accurate geometry local minima do tend to cluster at lower total energies. And, as the energies of the local minima increase, their $UME_{(Sm-L)S}$ or UMEs also tend to increase.

This trend adds to the validity of Sparkle/PM3 as a trustworthy lanthanide complexes geometry prediction tool.

Finally, the decision of which of the equivalent models to use either Sparkle/PM3 or Sparkle/AM1 rests with the user who must choose based on an appraisal of the influence of either AM1 or PM3 on the quantum chemical description of the specific ligands under investigation and the likely ensuing impact of this choice on the property of interest.

Acknowledgment. We appreciate the financial support from CNPq (Brazilian agency) and also a grant from PADCT and the Instituto do Milênio de Materiais Complexos. We also wish to thank CENAPAD (Centro Nacional de Processamento de Alto Desempenho) at Campinas, Brazil, for having made available to us their computational facilities. We gratefully acknowledge the Cambridge Crystallographic Data Centre for the Cambridge Structural Database.

Supporting Information Available: Instructions and examples on how to implement the Nd(III), Pm(III), and Sm(III) Sparkle/PM3 model in Mopac93r2; parts of the codes of subroutines Block.f, Calpar.f, and Rotate.f that need to be changed as well as their modified versions for Nd(III), Pm(III), and Sm(III); examples of Mopac93r2 crystallographic geometry input (.dat) and optimized geometry summary output (.arc) files from the Sparkle/PM3 calculations for the Nd(III) complex XONYII, for the dineodymium complex JIRHAT, for the Pm(III) complex XEXJAL{Pm}, for the dipromethium complex SOXKAR{Pm}, for the Sm(III) complex XAGVOQ, and for the disamarium complex QQQEMA01; tables of $UME_{(Ln-L)S}$ and UMEs for both Sparkle/PM3 and Sparkle/AM1 for Nd(III), Pm(III), and Sm(III), respectively; figures with gamma distribution fits of the UME data for both Sparkle/PM3 and Sparkle/AM1 models for Nd(III), Pm(III), and Sm(III); and table with UMEs for various types of distances involving the lanthanide ion and the ligand atoms, for both Sparkle/PM3 and Sparkle/AM1, for all complexes, and for each of the ions considered in this article. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Freire, R. O.; Rocha, G. B.; Simas, A. M. Sparkle Model for the Calculation of Lanthanide Complexes: AM1 Parameters for Eu(III), Gd(III) and Tb(III). *Inorg. Chem.* **2005**, *44*, 3299.
- Bastos, C. C.; Freire, R. O.; Rocha, G. B.; Simas, A. M. Sparkle Model for AM1 Calculation of Neodymium (III) Coordination Compounds. *J. Photochem. Photobiol., A* **2006**, *117*, 225.
- Freire, R. O.; Costa, N. B., Jr.; Rocha, G. B.; Simas, A. M. Sparkle/AM1 Parameters for the Modeling of Samarium (III) and Promethium (III) Complexes. *J. Chem. Theory Comput.* **2006**, *2*, 64.
- Freire, R. O.; Rocha, G. B.; Simas, A. M. Lanthanide Complex Coordination Polyhedron Geometry Prediction Accuracies of *ab-initio* Effective Core Potential Calculations. *J. Mol. Model.* **2006**, *12*, 373.
- Dewar, M. J. S.; Zoebish, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods 1. Method. *J. Comput. Chem.* **1989**, *10*, 209–220.
- Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods. 2. Applications. *J. Comput. Chem.* **1989**, *10*, 221–264.
- Stewart, J. J. P. *MOPAC2007, version 7.058*; Stewart Computational Chemistry: Colorado Springs, CO, U.S.A., 2007.
- Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98, revision A.7*; Gaussian, Inc.: Pittsburgh, PA, 1998.
- HyperChem(TM) Professional, version 7.51*; Hypercube Inc.: Gainesville, FL, U.S.A., 2006.
- Spartan, version 04*; Wavefunction Inc.: Irvine, CA, U.S.A., 2004.
- Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER, version 8*; University of California: San Francisco, CA, 2004.
- Field, M. J.; Albe, M.; Bret, C.; Proust-De Martin, F.; Thomas, A. The Dynamo Library for Molecular Simulations Using Hybrid Quantum Mechanical and Molecular Mechanical Potentials. *J. Comput. Chem.* **2000**, *21* (12), 1088.
- VAMP, version 8*; Accelrys Corporate Headquarters: San Diego, CA, U.S.A., 2001.
- Rowley, C.; Hassinen, T. *Ghemical for GNOME, version 1.00*; University of Iowa, Iowa City, IA, U.S.A., 2002.

- (16) Jorgensen, W. L. BOSS - Biochemical and Organic Simulation System. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; John Wiley & Sons Ltd.: Athens, U.S.A., 1998; Vol. 5, pp 3281–3285.
- (17) AMPAC, version 8; Semichem Inc.: Shawnee, U.S.A., 2004.
- (18) Cundari, T. R.; Deng, J.; Fu, W. T. PM3(tm) Parameterization Using Genetic Algorithms. *Int. J. Quantum Chem.* **2000**, *77*, 421.
- (19) Brothers, E. N.; Merz, K. M. Sodium Parameters for AM1 and PM3 Optimized Using a Modified Genetic Algorithm. *J. Phys. Chem. B* **2002**, *106* (10), 2779.
- (20) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods IV: Extension of MNDO, AM1, and PM3 to More Main Group Elements. *J. Mol. Model.* **2004**, *10* (2), 155.
- (21) Sundararajan, M.; McNamara, J. P.; Hillier, I. H.; Wang, H.; Burton, N. A. The Development of a PM3 Parameter set to Describe Iron-Sulfur Proteins. *Chem. Phys. Lett.* **2005**, *404* (1–3), 9.
- (22) Strobel, S.; Hartenbach, I.; Schleid, T. In *Synthesis and Crystal Structure of the Neodymium Sulfate Nitrate Trihydrate Nd[SO₄][NO₃]·3H₂O*; Proceedings of the XVIIIth Tage der Seltenen Erden (Terrae Rarae 2005), Bonn-Röttgen, Germany, 2005.
- (23) Song, Y. M.; Yao, X. Q.; Deng, T.; Wu, J. X.; Wu, Q. Structure of Tetrakis(pyridinioacetate) Neodymium(III) Tetrahydrate Perchlorate. *Chem. Pap.* **2006**, *60*, 302.
- (24) Reddmann, H.; Apostolidis, C.; Walter, O.; Amberger, H. D. Electronic Structures of Highly Symmetrical Compounds of f Elements. 40. Parametric Analysis of the Crystal Field Splitting Pattern of Tris(hydrotris(1-pyrazolyl)borato)neodymium(III). *Z. Anorg. Allg. Chem.* **2006**, *632* (8–9), 1405.
- (25) Rohde, A.; Urland, W. Catena-Poly[[2,2'-bipyridine-kappa N-2,N']neodymium(III)]-mu-dichloroacetato-1 kappa(2) O:O':2 kappa O-di-mu-dichloroacetato-kappa O-4:O']. *Acta Crystallogr., Sect. E: Struct. Rep. Online* **2006**, *62* (7), M1618.
- (26) Bowman, L. J.; Izod, K.; Clegg, W.; Harrington, R. W. Heteroleptic Complexes of Lanthanum(III) and Neodymium(III) with Oxygen- or Nitrogen-functionalized Tris(triorganosilyl)methyl Ligands. *Organometallics* **2006**, *25* (12), 2999.
- (27) Mondry, A.; Starynowicz, P. Ten-coordinate Neodymium(III) Complexes with Triethylenetetraaminehexaacetic Acid. *Eur. J. Inorg. Chem.* **2006**, *9*, 1859.
- (28) Peluffo, F.; Torres, J.; Kremer, C.; Dominguez, S.; Mederos, A.; Kremer, E. Phosphodiesterolytic Activity of Samarium(III) Mixed Ligand Complexes Containing Crown Ethers and Alpha-amino Acids. *Inorg. Chim. Acta* **2006**, *359* (7), 2107.
- (29) Wu, A. Q.; Zheng, F. K.; Liu, X.; Guo, G. C.; Cai, L. Z.; Dong, Z. C.; Takano, Y.; Huang, J. S. A Novel Bi-layered Samarium Complex With an Unprecedented Coordination Mode of Orotic Acid [Sm-2(HL)(2)(ox)(H₂O)(2)](n) center dot 2.5nH(2)O (H₃L = orotic acid, ox(2-) = oxalate(2-)): Synthesis, Crystal Structure and Physical Properties. *Inorg. Chem. Commun.* **2006**, *9* (4), 347.
- (30) Yuan, F. G.; Qian, H.; Min, X. An Ate Samarium Complex Containing Two Different Kinds of Reduced Azobenzene Species. *Inorg. Chem. Commun.* **2006**, *9* (4), 391.
- (31) d'Hardemare, A. D.; Philouze, C.; Jarjays, O. [11,23-Dimethyl-15,19-diaza-3,7-diazonia-tricyclo [19.3.1.1(9,13)]-tetracos-1(25),2,7,9,11,-13(26),14,19,21,23-decaene-25,26-diolato-kappa N-4(15),N-19,O,O']tris(nitrato-kappa O-2,O')-samarium(III). *Acta Crystallogr., Sect. E: Struct. Rep. Online* **2006**, *62* (2), M227.
- (32) Zhu, Y. J.; Chen, J. X.; Zhang, W. H.; Zhang, Y.; Lang, J. P. Catena-Poly[[[diapua[trans-3-(4-pyridyl)acrylato]samarium(III)]-di-mu-trans-3-(4-pyridyl)acrylato] dihydrate]. *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.* **2005**, *61*, M491.
- (33) Kostova, I.; Rastogi, V. K.; Kiefer, W.; Kostovski, A. New Cerium(III) and Neodymium(III) Complexes as (III) Cytotoxic Agents. *Appl. Organomet. Chem.* **2006**, *20* (8), 483.
- (34) Yang, L. F.; Gong, Z. L.; Nie, D. B.; Lou, B.; Bian, Z. Q.; Guan, M.; Huang, C. H.; Lee, H. J.; Baik, W. P. Promoting Near-infrared Emission of Neodymium Complexes by Tuning the Singlet and Triplet Energy Levels of Beta-diketonates. *New J. Chem.* **2006**, *30* (5), 791.
- (35) O'Riordan, A.; O'Connor, E.; Moynihan, S.; Nockemann, P.; Fias, P.; Van Deun, R.; Cupertino, D.; Mackie, P.; Redmond, G. Penetration of Radionuclides Across the Skin: Rat Age Dependent Promethium Permeation Through Skin in Vitro. *Thin Solid Films* **2006**, *497* (1–2), 299.
- (36) Salem, A. A. Fluorimetric Determinations of Nucleic Acids Using Iron, Osmium and Samarium Complexes of 4,7-diphenyl-1,10-phenanthroline. *Spectrochim. Acta, Part A* **2006**, *65* (1), 235.
- (37) Morandea, L.; Remaud-Le Saec, P.; Ouadi, A.; Bultel-Riviere, K.; Mouglin-Degrad, M.; de France-Robert, A.; Faivre-Chauvet, A.; Gestin, J. F. Synthesis and Evaluation of a Novel Samarium-153 Bifunctional Chelating Agent for Radioimmunotargeting Applications. *J. Labelled Compd. Radiopharm.* **2006**, *49* (2), 109.
- (38) Kassai, Z.; Koprda, V.; Bauerová, K.; Harangozó, M.; Bendová, P.; Bujnová, A.; Kassai, A. Penetration of radionuclides across the skin: Rat age dependent promethium permeation through skin in vitro. *J. Radioanal. Nucl. Chem.* **2003**, *258*, 669.
- (39) Li, W. P.; Smith, C. J.; Cutler, C. S.; Ketring, A. R.; Jurisson, S. S. Development of Receptor-Based Radiopharmaceuticals Using Carrier-free Promethium-149: Syntheses, in Vitro Stability Studies, and in Vivo Biodistribution Studies of DTPA, DOTA, and DTPA-Octreotide Complexes. *J. Nucl. Med.* **2000**, *41* (5), 246.
- (40) Lewis, M. R.; Zhang, J. L.; Jia, F.; Owen, N. K.; Cutler, C. S.; Embree, M. F.; Schultz, J.; Theodore, L. J.; Ketring, A. R.; Jurisson, S. S.; Axworthy, D. B. Biological Comparison of Pm-149-, Ho-166-, and Lu-177-DOTA-biotin Pretargeted by CC49 ScFv-streptavidin Fusion Protein in Xenograft-Bearing Nude Mice. *Nucl. Med. Biol.* **2004**, *31* (7), 973.
- (41) Lu, G.; Bai, M.; Li, R.; Zhang, X.; Ma, C.; Lo, P.-C.; Ng, D. K. P.; Jiang, J. Lanthanide(III) Double-decker Complexes with Octaphenoxy- or Octathiophenoxyphthalocyaninato Ligands - Revealing the Electron-Withdrawing Nature of the Phenoxy and Thiophenoxy Groups in the Double-decker Complexes. *Eur. J. Inorg. Chem.* **2006**, *18*, 3703.
- (42) Anwander, R. Lanthanides: Chemistry and Use in Organic Synthesis. In *Topics in Organometallic Chemistry*, 1st ed.; Kobayashi, S., Eds.; Springer: Berlin, Germany, 1999; Vol. 2, pp 1–62.

- (43) Allen, F. H. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, 58, 380.
- (44) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, 58, 389.
- (45) Allen, F. H.; Motherwell, W. D. S. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, 58, 407.
- (46) Dolg, M.; Stoll, H.; Savin, A.; Preuss, H. *Theor. Chim. Acta* **1989**, 75, 173.
- (47) Conover W. J. Statistics of the Kolmogorov-Smirnov type. In *Practical nonparametric statistics*, 3rd ed.; Wiley, B., II, Ed.; John Wiley & Sons: New York, U.S.A., 1999; pp 428–473.

CT600326M

Toward the Prediction of Organic Hydrate Crystal Structures

Ashley T. Hulme[†] and Sarah L. Price*

*Department of Chemistry, University College London, 20 Gordon Street,
London WC1H 0AJ, United Kingdom*

Received February 22, 2007

Abstract: Lattice energy minimization studies on four ordered crystal structures of ice and 22 hydrates of approximately rigid organic molecules (along with 11 corresponding anhydrate structures) were used to establish a model potential scheme, based on the use of a distributed multipole electrostatic model, that can reasonably reproduce the crystal structures. Transferring the empirical repulsion–dispersion potentials for organic oxygen and polar hydrogen atoms to water appears more successful for modeling ice phases than using common water potentials derived from liquid properties. Lattice energy differences are reasonable but quite sensitive to the exact conformation of water and the organic molecule used in the rigid molecule modeling. This potential scheme was used to test a new approach of predicting the crystal structure of 5-azauracil monohydrate (an isolated site hydrate) based on seeking dense crystal packings of 66 5-azauracil···water hydrogen-bonded clusters, derived from an analysis of hydrate hydrogen bond geometries involving the carbonyl- and aza-group acceptors in the Cambridge Structural Database. The known structure was found within 5 kJ mol⁻¹ of the global minimum in static lattice energy and as the third most stable structure, within 1 kJ mol⁻¹, when thermal effects at ambient temperature were considered. Thus, although the computational prediction of whether an organic molecule will crystallize in a hydrated form poses many challenges, the prediction of plausible structures for hydrogen-bonded monohydrates is now possible.

1. Introduction

Hydrate formation is common for organic molecules, with estimates varying from a third of organic molecules^{1,2} to perhaps three-quarters of pharmaceutical compounds forming hydrates.³ Many manufacturing processes provide an opportunity for hydrates to form,² and the state of hydration can be changed with environmental humidity and time.^{2,4} The state of hydration of an active pharmaceutical ingredient can significantly affect the solubility and dissolution rate and therefore its bioavailability.⁴ Thus intimate control of the hydration state of an active pharmaceutical ingredient is required. Such control can only follow detailed investigations into the existence of hydrated states, their formation, and

their dehydration products.⁵ The aim of this study is to take the first steps toward the use of computational crystal structure prediction to aid investigations into hydrate as well as polymorph screening.⁶

The roles of the water molecules in hydrates have been classified into three categories:² those in which each water molecule is isolated in the lattice (isolated site hydrate) and only hydrogen bonded to the organic molecule; those where water occupies channels in the crystal structure in which the water content is invariant (stoichiometric channel hydrate) or can readily vary in stoichiometry with the environmental humidity (nonstoichiometric channel hydrate); and those where the water is associated with ionic species (ion associated hydrates). Attempts have also been made to classify the extended water hydrogen-bonded substructures in hydrates,^{3,7} in a manner analogous to graph set analysis,⁸ defining chains, rings, tapes, and layers of water molecules.

* Corresponding author e-mail: s.l.price@ucl.ac.uk.

[†] Current address: Pharmorphix Ltd., 250 Cambridge Science Park, Milton Rd., Cambridge CB4 0WE, U.K.

Well-defined hydrate crystal structures in the Cambridge Structural Database (CSD)⁹ show a range of possible water hydrogen-bonding geometries¹⁰ from a single hydrogen bond up to the maximum of four. It has been contended that hydrate formation is more prevalent for those organic molecules in which there is a hydrogen bond donor/acceptor imbalance,¹¹ particularly when there are fewer donors than acceptors, with the inclusion of the water addressing this imbalance. This contention has been investigated by a statistical analysis of hydrate structures in the CSD,¹² which concluded that the sum of donors and acceptors of a molecule, rather than their ratio, influences hydrate formation, which is also increased with the polarity of the surface of the organic molecule. Hydrates can also be formed with more extensive, disordered water filling channels and voids in the crystal structure, extending to highly solvated protein structures where only a minority of waters in close proximity to the protein are in fixed positions.¹³ While a computational method of predicting hydrate formation would be a useful complement to the formulation and process design for pharmaceuticals, it is clearly a major challenge which will be very dependent on the molecule involved: a comprehensive method of hydrate prediction would have to search through a range of stoichiometries for possible hydrate crystal structures and consider entropic effects for disordered water as well as incorporating conformational flexibility. More organic hydrate crystal structures are known than all other solvates combined,¹⁴ emphasizing water's unique role in solvation and crystallization and being a solvent of choice in industrial processes.

The most fundamental requirement for hydrate prediction is the accurate modeling of the balance of organic molecule...organic molecule, organic molecule...water, and water...water intermolecular interactions. This study has investigated whether current models for intermolecular forces that are used for the crystal structure prediction of rigid organic molecules¹⁵ are suitable for hydrate prediction. Most simple water potentials have been parametrized against a wide range of liquid properties,¹⁶ with only one explicit attempt to modify a water potential for use with ice,¹⁷ which reparameterized the TIP4P potential to reproduce the density of several forms of ice. A wide range of computationally inexpensive models for water...water interactions were tested for their ability to reproduce four of the ordered crystal structures of ice, to assess their ability to model water...water interactions in the crystalline state. The repulsion–dispersion model potentials tested range from several that have been developed and are widely used for simulating liquid water to those derived from the empirical model potentials commonly used in organic crystal structure prediction (CSP).¹⁸ Both atomic charge and atomic multipole¹⁹ descriptions of the dominant electrostatic interactions were tested in conjunction with each repulsion–dispersion potential. Given the computational expense required for crystal structure prediction, more realistic but more complex model intermolecular pair potentials for water, which include further terms such as flexibility, anisotropic repulsion, and polarization,^{16,20,21} were not considered. The most promising water model was then tested, in conjunction with a commonly used

intermolecular potential for organic molecules, for its ability to reproduce the organic molecule...water interactions in the crystal structures of a range of 22 hydrate structures of rigid organic molecules. Corresponding anhydrate crystal structures were also available for seven of these compounds, allowing the lattice energies of the hydrates to be compared with those of the corresponding anhydrates and ice.

This model potential scheme is then used in a proof-of-concept test to predict the crystal structure of 5-azauracil monohydrate with the foreknowledge that 5-azauracil does indeed form a monohydrate. In order to avoid searching through the whole multidimensional space for crystals with two independent molecules in the unit cell, we developed a specific search strategy, analogous to one recently applied to diastereomeric salts,²² based on the assumption that the water would be hydrogen bonded to the organic molecule, and using an analysis of the CSD to determine the likely water hydrogen-bonding geometries. This approach, assuming the approximate hydrogen-bonding geometry and dense packing, significantly reduces the number of structures that have to be considered compared with more mathematically complete search methods appropriate for two independent molecules in the asymmetric unit cell.^{23,24} The random search method used in the only previously published work on the prediction of hydrate structures²⁵ found the experimental structure for only five of the nine polyalcohol and carbohydrate monohydrates considered, though the difficulty of these searches was considerably increased by the flexibility of the molecules.

The crystal structure prediction of 5-azauracil monohydrate tests the ability of our search strategy to find the known structure and that of the model potential scheme to successfully model the energy of the known structure relative to the alternative structures. Thus, we can assess the current possibilities of hydrate prediction on the basis of static lattice energy minimization.

2. Method

2.1. Testing Intermolecular Potentials for Their Ability To Model Ice. The phase diagram of ice presently contains 14 distinct crystalline phases. Phases I_h (common ice), III, IV, V, VI, VII, and XII are disordered and are unsuitable for testing the suitability of potentials for modeling crystalline phases through static lattice energy minimization. However, ices II^{26,27} and VIII,²⁸ which are high pressure ordered phases with no disordered analogues, ice IX,²⁹ a nearly ordered modification of ice III, and ice XI,^{30,31} a low temperature, ambient pressure proton-ordered modification of ice I_h, are suitable for this task. Recently, after the completion of the current study, an ordered version of ice V and a partially ordered version of ice XII have been reported.³² The structural properties of ices II, VIII, IX, and XI are reported in Table 1 and were used to test a range of water intermolecular potentials. These crystal structures were determined for D₂O using neutron diffraction and the accurate location of the hydrogen atoms allowed the experimental water geometries to be used. However, the variations in these molecular geometries (Table 1) from the isolated molecule³³ bond length (0.9572 Å) and bond angle (104.52°) demon-

Table 1. Summary of the Ice Structures Used To Test the Water Intermolecular

structure	space group (SG for energy minimization)	Z	intramolecular geometry		hydrogen bonds	
			O–H length (Å)	H–O–H angle (°)	O···O length (Å)	O–H···O angle (°)
ice II ²⁷	$R\bar{3}$	2	0.958	103	2.805	166
			0.972	107	2.767	167
			0.942	-	2.779	178
			1.014	-	2.845	168
ice VIII ²⁸	$I4_1/amd$ ($C112_1$)	0.5 (1)	0.968	106	2.879	178
ice IX ²⁹	$P4_12_12$ ($P2_1$)	1.5 (6)	0.977	106	2.75	167
			0.971	105	2.797	175
			0.979	-	2.763	165
ice XI ³¹	$Cmc2_1$ ($P\bar{1}$)	1 ^a (4)	0.976	108	2.74	177
			1.054	114	2.803	178
			0.947	-	2.737	176

^a Asymmetric unit contains two-half molecules – one located on a 2-fold axis and one on a mirror plane.

strate the flexibility of the molecular structure to distort within the tetrahedral hydrogen-bonding coordination upon crystallization and highlights the inherent limitations of our approximation of modeling the water as rigid.

Each ice structure was lattice energy minimized using DMAREL,³⁴ allowing for rigid body rotation, translation, and cell changes within the symmetry constraints of the subgroup of the experimental space group which gave whole molecules in the asymmetric unit (Table 1). After energy minimization, the retention of the higher experimental symmetry by the energy minimized structures was confirmed by the PLATON³⁵ ADDSYM algorithm. Consideration was limited to model potentials where the repulsion–dispersion potential was of the Lennard-Jones 12–6 form, specifically SPC,³⁶ its derivatives SPC/E³⁷ and MSPC/E,³⁸ and the two related potentials TIP3P³⁹ and TIP4P³⁹ or the Buckingham exp-6 form using the NSPC/E⁴⁰ model. Two other Buckingham potentials which had been derived by empirical fitting to organic crystal structures rather than water were also considered: the FIT potential¹⁵ where the oxygen potential was derived from oxohydrocarbons⁴¹ and the polar hydrogen from N–H groups mainly hydrogen bonded to carbonyl groups¹⁵ and a variation of this, FIT(COOH), where the polar hydrogen parameters were optimized to carboxylic acids.⁴² These published water potentials were tested with their own charge model (denoted STD) and in combination with both the CHELPG potential derived charges (ESP)⁴³ and distributed multipoles (DMA)¹⁹ derived from the MP2/6-31G(d,p) charge density calculated for the molecular structure for each ice. Only potential models with interaction sites on the atomic nuclei were considered, thus excluding using the full TIP4P model which has the oxygen partial charge at a non-nuclear position. The electrostatic contribution to the lattice energy was evaluated by Ewald summation for all charge–charge, charge–dipole, and dipole–dipole terms, with the repulsion–dispersion potential summed to a 15 Å atom–atom cutoff, and the higher multipole–multipole interactions in the DMA models up to R⁻⁵ summed to a 15

Å center of mass cutoff. All minima were confirmed to be stable by considering their second derivative rigid body properties. Thus a total of 21 intermolecular potential combinations for water were tested for their ability to reproduce the four ordered ice structures, by static lattice energy minimization. While the reproduction of the lattice parameters and density of ice XI is of most interest, as its stability domain of under 73 K at ambient pressure most closely matches the static lattice energy approximation of zero temperature and pressure, the lattice parameters of the other forms should be only underestimated by a few percent with the structures substantially unchanged. The overall structure reproduction for both ice structures and organic hydrates was quantified by the usual weighted *F*-value ‘figure of shame’⁴⁴

$$F = \left(\frac{100\Delta a}{a}\right)^2 + \left(\frac{100\Delta b}{b}\right)^2 + \left(\frac{100\Delta c}{c}\right)^2 + \Delta\alpha^2 + \Delta\beta^2 + \Delta\gamma^2 + (10\text{rms}\Delta x)^2 + \left(\frac{\text{rms}\Delta\theta}{2}\right)^2$$

where Δa is the error in cell length a (Å), $\Delta\alpha$ is the error in cell angle α (°), and the root-mean-square (rms) values of the space-group symmetry-allowed rigid-body center of mass translations Δx (Å) and rotations $\Delta\theta$ (°) are calculated over all the molecules in the unit cell. Since there are more symmetry-unconstrained molecular translations with two (or more) independent molecules in the asymmetric unit, *F*-values are expected to be larger for hydrates than for $Z' = 1$ anhydrate crystal structures.

2.2. Testing Model Intermolecular Potentials for Organic Hydrates. A test set of hydrate structures was constructed by searching the CSD (May 2004) for suitable hydrate crystal structures containing only the atomic species C, H, N, O, and F and excluding structures containing ions, polymers, disordered structures, structures without three-dimensional coordinates determined, and structures solved from powder X-ray diffraction data. Each structure in the subsequent set was examined manually to eliminate those with undetermined water hydrogen positions and those in which the parent molecule was deemed too flexible, i.e., contained groups with greater conformational flexibility than methyl, nitro, and amino substituents. Of the resulting test set of 22 hydrates, seven were found to have one or more corresponding anhydrous crystal structures which were also modeled for comparison.

The repulsion–dispersion potential tested was the FIT potential with the exp-6 form

$$U = \sum_{i \in M, k \in N} (A_u A_{kk})^{1/2} \exp(-(B_u + B_{kk})R_{ik}/2) - (C_u C_{kk})^{1/2}/R_{ik}^6$$

where atom i of molecule M is of type ι and separated by an intermolecular distance R_{ik} from atom k of molecule N of type κ . Following the success of this model for reproducing the ice structures (section 3.1), the parameters used for C, N, and H_{np} (bonded to C) were those that had originally been fitted to azahydrocarbons,⁴⁵ for F to perfluorocarbons,⁴⁶ and for O (in the organic molecule and water) to oxohydrocar-

bons.⁴¹ The parameters used for H_p, bonded to either N, O or in water, had been fitted to a range of polar and hydrogen bonded organic crystal structures¹⁵ in combination with the same C, H_{np}, N, and O parameters and a DMA electrostatic model. This model potential combined with a DMA model has been widely used in organic crystal structure prediction⁴⁷ including fluorinated compounds.^{48,49} This validation against hydrate crystal structures was to assess the ability of the geometric combining rules to correctly model organic molecule...water interactions by extrapolating from models for water...water and organic molecule...organic molecule crystal structures. This empirical model effectively represents all contributions to the intermolecular potential except the electrostatic contribution, which was represented by the distributed multipole representation of the MP2/6-31G(d,p) wavefunctions of the isolated molecules.

In order to test sensitivity to the molecular structure, for each hydrate crystal structure the lattice energy minimum using both the experimental molecular conformation (denoted ExpMinExp) and the ab initio optimized molecular conformation (denoted ExpMinOpt) were determined using the same DMAREL methodology described in 2.1. To allow for the systematic error in X-ray location of hydrogens, the ExpMinExp molecular structures had all bond lengths to hydrogen extended to neutron values,⁵⁰ with the standard SPC length for water (1.0 Å).³⁶ For the ExpMinOpt structures, the organic molecules were optimized at the MP2/6-31G(d,p) level, and a standard water geometry was used with the MP2/6-31G(d,p) optimized water molecule (angle 103.8° and bond length of 0.961 Å) (cf. Table 1). For three molecules, the crystal structure reproduction was sufficiently sensitive to certain flexible torsion angles that additional studies were performed with molecular structures in which just these torsion angles were constrained to experimental values and all other molecular parameters optimized (ExpMinConOpt). The ExpMinConOpt set of calculations constrained the NO₂ torsion for anhydrous 5-nitouracil and its monohydrate, one hydroxyl group proton torsion in dialuric acid monohydrate and the ring atom positions in anhydrous 5-fluorocytosine form 1. All ab initio calculations were performed using GAUSSIAN03,⁵¹ and the distributed multipoles for the specific molecular charge density were obtained using GDMA1.0.⁵²

2.3. Crystal Structure Prediction for 5-Azuracil Monohydrate. The model potential, as validated in the previous sections, was then tested for its ability to predict a monohydrate in which the water is hydrogen bonded to the organic molecule, by generating densely packed crystals from rigid molecule...water bimolecular clusters and then minimizing the lattice energy, allowing the water and molecule to independently adjust their relative orientation and position within the crystal lattice. The test system chosen was 5-azauracil monohydrate as both the anhydrate and monohydrate crystal structures are both reproduced with typical accuracy by the model intermolecular potential scheme (sections 3.1 and 3.2). The anhydrate structure has already been predicted as the global minimum in its lattice energy⁵³ and is well reproduced by Molecular Dynamics simulations at 310 K⁵⁴ with the same model potential. However, the main

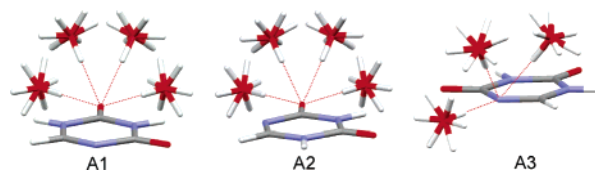


Figure 1. The 66 5-azauracil-water clusters used to generate initial monohydrate crystal structure, with notation for acceptors A1 and A2 and A3.

reason for choosing 5-azauracil is that this rigid molecule has three hydrogen bond donors and two acceptors, providing considerable variety of possible 5-azauracil...water, 5-azauracil...5-azauracil, and water...water hydrogen bonds in the crystals.

The search strategy employed required sufficient initial crude monohydrate crystal structures to be generated to ensure that all plausible hydrogen-bonding geometries could be sampled during the lattice energy optimization. Since the lattice energy minimization procedure was unlikely to significantly move the water molecules when optimizing from a densely packed crystal structure, it was necessary to establish a range for common hydrogen-bonded geometries for water (H_w–O_w–H_w) around the carbonyl and nitrogen acceptor groups by analysis of hydrate structures present in the CSD as described in the Supporting Information. The distributions for the H_w...O and H_w...N bond lengths and O_w–H_w...O and O_w–H_w...N angles were sufficiently sharp (Supporting Information, Figures S2 and S3) that clusters were defined with these parameters set to 1.9 Å, 1.9 Å, 135°, and 120°, respectively, and the hydrogen bonds were fixed to be linear. However, the torsion angles required to define the positions of the water molecule relative to the 5-azauracil molecular plane and the orientation of the non-hydrogen-bonded H_w atom were found to be fairly evenly distributed in both cases (Supporting Information, Figures S2 and S3). Hence clusters were defined at each acceptor (shown in Figure 1) so that the O_w was positioned by a torsion angle with values of 0°, 60°, 120°, or 180° (excluding one geometry where the water molecule physically overlapped the 5-azauracil molecule), and then the non-hydrogen-bonded H_w was defined by its torsion angle being either 0°, 60°, 120°, 180°, 240°, or 300°. Figure 1 shows all the resulting 66 starting point clusters which were built from the optimized molecular conformations. The implicit assumption that these clusters would automatically generate hydrogen bonds to the two N–H acceptors in the densely packed crystal structures was carefully monitored during the results analysis and appeared to be appropriate in this specific case because of the proximity of the donor and acceptor sites.

Each rigid 5-azauracil...water cluster was used in MOLPAK⁵⁵ to generate monohydrate crystal structures with one 5-azauracil...water cluster in the asymmetric unit cell in 37 MOLPAK packing types covering 18 space groups. The MOLPAK packing types were defined to represent the most common modes of packing of organic molecules, and this remains appropriate since the space group distribution of hydrate crystal structures is similar to that for organic structures generally.² The densest 125 of the approximately 5000 structures generated in each packing type were passed

Table 2. Summary of the Average F -Value for Each Repulsion–Dispersion Plus Electrostatic Model Combination Averaged over the Four Ordered Ice Structures

repulsion–dispersion potential	electrostatic model	average F -value
FIT	DMA	21
SPC/E	DMA	56
SPC	DMA	57
TIP4P	DMA	68
TIP3P	DMA	75
TIP3P	STD	94
SPC/E	STD	95
MSPC/E	ESP	96
SPC	STD	99
MSPC/E	STD	101
TIP3P	ESP	105
TIP4P	ESP	109
SPC	ESP	117
SPC/E	ESP	118
MSPC/E	DMA	130
NSPC/E	DMA	173
FIT(COOH)	ESP	175
FIT(COOH)	DMA	192
NSPC/E	ESP	204
NSPC/E	STD	240
FIT	ESP	285

to DMAREL for lattice energy minimization, using the same model potential and methodology as described in 2.1 and 2.2. Thus approximately 300 000 lattice energy minimizations were carried out. All structures that minimized to saddle points were discarded. The optimized structures were compared using COMPACK⁵⁶ to overlay the 15 molecule coordination spheres and powder patterns^{57,58} to determine the unique low-energy structures in each search and then from the combination of all 66 searches. The elastic constants and $k = 0$ phonons for the unique crystal structures were calculated within the rigid-body harmonic approximation^{59,60} and used to estimate the zero-point energy, entropy, and the Helmholtz free energy⁶¹ at 298 K.

3. Results

3.1. The Testing of Water Potentials To Reproduce the Ordered Structures of Ice. The minima in the lattice energy for each of the eight different repulsion–dispersion models in combination with up to three different electrostatic models are compared with the experimentally known ice structures used as the starting point for the minimization in the Supporting Information (Tables S1–S4 for ice II, VIII, IX, and XI, respectively). Although many potentials reproduce one or more structures satisfactorily within the limits of static lattice energy minimization (errors in cell dimensions $< 5\%$, $F < \sim 50$), many minimizations result in grossly distorted structures, for example, with one cell parameter changing by over 10%. The average of the F -values for all four structures for each intermolecular potential model are summarized in Table 2, in order of increasing average F -value.

From this ranked list, it is clear that for the majority of potentials, the electrostatic model is of principal importance in determining the ability to reproduce the ice structures, with the theoretically more accurate distributed multipole model giving superior results compared to atomic charge models. This is in accord with the ability of the distributed multipole model to model the orientation dependence of the dominant term for hydrogen-bonding directionality.^{18,62,63} The NSPC/E and FIT(COOH) dispersion–repulsion potentials performed poorly, irrespective of the electrostatic model. It can be concluded from the poor performance of the FIT(COOH) that the hydrogen repulsion in water is closer to that for an N–H hydrogen than a carboxylic acid O–H hydrogen. The FIT dispersion–repulsion potential, combined with the distributed multipole electrostatic model, gives the overall best reproduction of the four test ice structures (Table 2). For this potential, all of the lattice energy minima are up to 5% denser than the corresponding experimental structure, consistent with the neglect of thermal effects, and with all of the structural variables quantitatively reproduced to reasonably good agreement (Table 3). The sensitivity to the intramolecular geometry was also investigated by contrasting (Table 3) the lattice energy minima for all four ice structures found with their specific molecular structure (ExpMinExp) with those (ExpMinOpt) minima found using the MP2/6-31G(d,p) optimized rigid water geometry, as used in the hydrate modeling. This made little difference to the structural reproductions, though it did destabilize the lattice energy in all cases by up to 2.5 kJ mol⁻¹. The lattice energies for the four polymorphs fell into a small range, -55.53 to -53.27 kJ mol⁻¹ (ExpMinExp), which is typical of the relative energy differences between polymorphs, and compares reasonably well⁶⁴ with the sublimation enthalpy of ice I_h at 0 K, calculated to be 47.34(2) kJ mol⁻¹.⁶⁵

It is perhaps surprising that transferring the repulsion–dispersion model from organic functional groups rather than a specific water model appears more successful; perhaps this reflects that the errors in transferring a simple model from the liquid to the idealized crystalline state are larger. A significant component of the errors in the reproductions of the ice structures are likely to be due to the approximations of lattice energy minimization to model these ice phases, most of which are only stable at high pressures. The success of this relatively simple model for ice structures arises from the realistic modeling of the dominant electrostatic contribution. For comparison we note that an eight-site intermolecular potential for water with 77 fitted-parameters reproduces²¹ 1077 calculated points with negative (stable) energies for the water dimer with an rms error of 0.4 kJ mol⁻¹. The results in Table 3 lead to the conclusion that further empirical refinement of specific water parameters for lattice energy minimization studies by fitting to the ordered ice structures was not warranted.

3.2. Results for the Reproduction of Hydrate Crystal Structures. The reproductions of the hydrate and anhydrate crystal structures on lattice energy minimization vary considerably. The F -values are summarized in Tables 4 and 5, and the detailed results, including an analysis of the most poorly reproduced hydrogen bonds, are included in the

Table 3. Reproduction of the Crystal Structures of the Ordered Polymorphs of Ice by the FIT+DMA Model Potential Used in the Hydrate Modeling

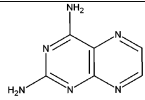
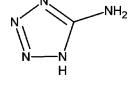
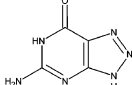
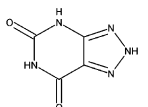
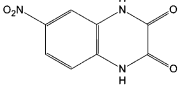
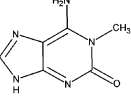
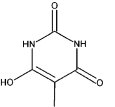
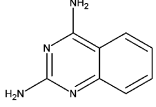
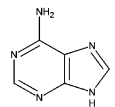
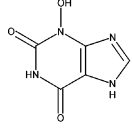
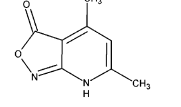
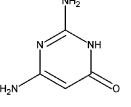
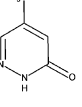
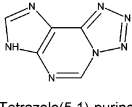
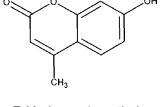
	a (Å)	% error	b (Å)	% error	c (Å)	% error	density (g cm ⁻³)	% error	energy (kJ mol ⁻¹)	F
Ice II										
experimental ²⁷	12.983		12.983		6.254		1.18			
ExpMinExp	12.773	-1.6	12.773	-1.6	6.190	-1.0	1.23	4.4	-53.61	14
ExpMinOpt	12.709	-2.1	12.709	-2.1	6.226	-0.5	1.24	4.8	-53.13	18
Ice VIII										
experimental ²⁸	4.656		4.656		6.775		1.63			
ExpMinExp	4.583	-1.6	4.583	-1.6	6.635	-2.1	1.72	5.4	-55.53	10
ExpMinOpt	4.551	-2.3	4.551	-2.3	6.732	-0.6	1.72	5.4	-54.81	12
Ice IX										
experimental ²⁹	6.73		6.83		6.73		1.16			
ExpMinExp	6.639	-1.4	6.701	-1.9	6.639	-1.4	1.22	4.7	-53.27	16
ExpMinOpt	6.64	-1.3	6.696	-2.0	6.64	-1.3	1.22	4.8	-52.59	17
Ice XI										
experimental ³¹	4.5019		7.7979		7.328		0.93			
ExpMinExp	4.696	4.3	7.484	-4.0	7.194	-1.8	0.95	1.7	-55.47	44
ExpMinOpt	4.483	-0.4	7.760	-0.5	7.320	-0.1	0.94	1.0	-53.07	6

Supporting Information, Tables S5 and S6. Examination of the lattice energy minimized structures using the standard hydrogen bond criteria within PLUTO⁶⁶ led to the conclusion that in only 5 of the 22 hydrate structures had the hydrogen bonding been altered upon energy minimization. The worst result of these five (THYMMH, CYTOSM, AZGUAN, CIMMEQ, DIALAC02) structures is shown in Figure 2, from which it is clear that the huge F_{opt} value of 1633 arises from large changes in the relative water positions, with corresponding large cell deviations, but the thymine hydrogen-bonded ribbon remains intact. This contrasts with the majority of the monohydrate minimizations, which resulted in $F_{\text{opt}} \leq 120$ and the hydrogen-bonding motif being well reproduced, as exemplified by the overlay in Figure 3 for 5-azauracil monohydrate.

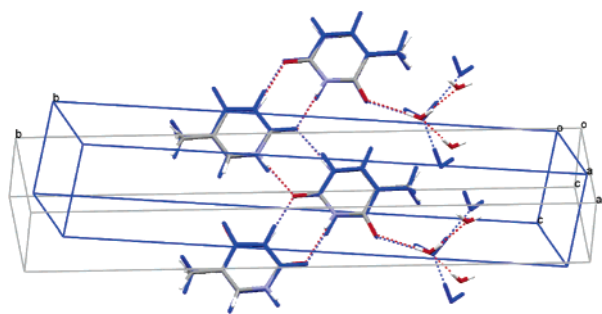
From the successful reproduction of the majority of hydrates structures it was concluded that the FIT potential would be adequate for use in a crystal structure prediction search for monohydrates of a small organic molecule, but, as generally found in organic crystal structure modeling, there are some pathological cases. The seven cases where both the hydrate and the anhydrate are modeled allowed comparison of the hydrate lattice energies with the weighted mean of the anhydrate and ice energy calculated with the same computational model (Table 5). Crudely, at 0 K there would be no thermodynamic reason for the monohydrate to form if its lattice energy was less stable than the sum of the lattice energies of the anhydrate and ice. However, the results show that the calculated hydrate lattice energies are commonly close to this value: i.e. the calculations do not show any great thermodynamic driving force for hydrate formation, with the differences generally comparable with the differences in lattice energy between ExpMinExp and ExpMin(Con)Opt energy minimizations. The former are generally more stable, implying that the very small conformational distortions probably arise from the crystal packing, but the energies are very sensitive to the exact positions of the protons in this rigid body modeling.⁶⁴

3.3. Crystal Structure Prediction Applied to 5-Azauracil Monohydrate. Comparison of the low-energy monohydrate structures found in the search with the ExpMinOpt minimized experimental 5-azauracil monohydrate structure showed that it was found as the 23rd ranked structure, 4.3 kJ mol⁻¹ above the global minimum in the lattice energy. All the more stable predicted structures are shown in Table 6. Inspection of these structures showed that the majority corresponded to sheets with hydrogen bonds solely between the 5-azauracil and water molecules, with none between pairs of 5-azauracil molecules. The experimental structure has the water hydrogen bonded only within the sheet (Figure 3) and only weak van der Waals interactions between sheets, whereas the majority of lower energy structures contained one of two sheet motifs (Figure 4) with the out-of-plane water proton hydrogen bonding to an adjacent sheet. The hypothetical sheet 1 can be derived from the experimental sheet structure (Figure 5), by the rotation of the water molecule out of the plane and a compensating adjustment of the 5-azauracil molecules to give a closer contact between the carbonyl oxygens. The more stable sheet 2 differs significantly from sheet 1 (and the experimental sheet) in that alternating 5-azauracil molecules have to be rotated by about 120° in the plane of the molecule (Figure 4), which is unlikely to occur as a solid-state transformation. The lowest energy nonsheet structure (A3_1_c ad/9) has a water molecule doubly hydrogen bonded to N1-H3 and O2 in a ring motif, and A3_1_f ad/31 has the water hydrogen bonded to the acceptor atoms N3 and O3, both being considerable distortions from the initial bimolecular cluster hydrogen-bonding motifs. However, the two other possible doubly hydrogen-bonded water...5-azauracil motifs do not appear to pack sufficiently well to appear in this low-energy region. The other nonsheet structures only have single hydrogen bonds to water. Thus only the experimental sheet structure shows the most common hydrogen-bonding motif for water in hydrates¹² (donor-donor-acceptor with the water ac-

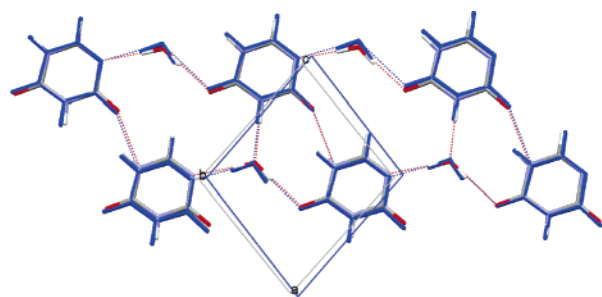
Table 4. Some Hydrate Structures Used To Test the Intermolecular Potential^a

 2,4-Diaminopteridine monohydrate AMPTRA10 $F_{\text{exp}} = 96$; $F_{\text{opt}} = 56$	 5-Aminotetrazole monohydrate AMTETZ01 $F_{\text{exp}} = 32$; $F_{\text{opt}} = 115$	 8-Azaguanine monohydrate AZGUAN $F_{\text{exp}} = 458$; $F_{\text{opt}} = 295$
 Xanthazole monohydrate XANAZH01 $F_{\text{exp}} = 114$; $F_{\text{opt}} = 95$	 6-Nitro-2,3-dihydroxy-quinoxaline monohydrate BAKGOJ01 $F_{\text{exp}} = 27$; $F_{\text{opt}} = 57$	 1-Methyl-isoguanine dihydrate CIMMEQ $F_{\text{exp}} = 504$; $F_{\text{opt}} = 1366$
 Dialuric acid monohydrate DIALAC02 $F_{\text{exp}} = 198$; $F_{\text{opt}} = 237$	 2,4-Diaminoquinazoline monohydrate DUPYIW $F_{\text{exp}} = 16$; $F_{\text{opt}} = 43$	 Adenine trihydrate FUSVAQ01 $F_{\text{exp}} = 111$; $F_{\text{opt}} = 106$
 3-Hydroxyxanthine dihydrate HXANTH10 $F_{\text{exp}} = 37$; $F_{\text{opt}} = 84$	 4,6-Dimethyl-isoxazolo-(3,4-b)pyridin-3-one monohydrate MIOZPO $F_{\text{exp}} = 106$; $F_{\text{opt}} = 208$	 2,6-Dimano-4-pyrimidinone monohydrate SEYDIJ $F_{\text{exp}} = 134$; $F_{\text{opt}} = 184$
 5-Methylpyridazine-3-one monohydrate TEKVIO $F_{\text{exp}} = 58$; $F_{\text{opt}} = 46$	 Tetrazolo(5,1)-purine monohydrate TRZPUR $F_{\text{exp}} = 136$; $F_{\text{opt}} = 118$	 7-Hydroxy-4-methylchromen-2-one monohydrate WIKDAV $F_{\text{exp}} = 25$; $F_{\text{opt}} = 36$

^a The CSD refcode for the specific crystal structure used is given, with the F -values for the ExpMinExp (F_{exp}) and ExpMinOpt (F_{opt}) or ExpMinConOpt (F_{copt}) lattice energy minimizations with the different rigid molecule structures.

**Figure 2.** Overlay of the experimental crystal structure (colored by element) with that of the ExpMinOpt energy minimized structure (colored blue) for thymine monohydrate, the worst reproduction.

cepting one hydrogen bond and acting as a hydrogen bond donor to two independent acceptors).

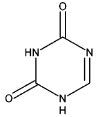
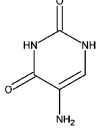
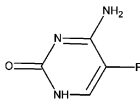
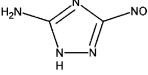
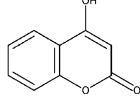

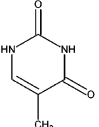
**Figure 3.** Overlay of the experimental crystal structure (colored by element) with that of the ExpMinOpt energy minimized structure (colored blue) for 5-azauracil monohydrate.

Thus, the search strategy was certainly successful in locating the known monohydrate structure and a range of plausible alternatives. Indeed the frequency with which all the sheet structures were found suggests a high level of redundancy in generation of these structures, though this has to be balanced by two of the nonsheet structures being found only once. Detailed analysis (Supporting Information, Table S7) shows that one of the sheet 1 structures could be found starting from all 66 clusters and that the experimental structure was found starting from 50 cluster geometries. Thus the exploration of possible hydrogen-bonding geometries appears reasonably complete for this system.

The known structure is not the lowest in static lattice energy, though the energy gap to the global minimum structure of less than 5 kJ mol^{-1} is small. Its relative stability is improved by considering entropic effects, as the experimental structure is the third most stable according to the estimated Helmholtz free energy at 298 K (Table 6). While a sheet 2 structure remains the most stable, the difference is reduced to only 0.8 kJ mol^{-1} , though certain sheet 1 structures and nonsheet structures remain thermodynamically competitive. It is notable that there are many different ways of stacking both hypothetical sheet structures with very similar lattice energies but different susceptibility to shearing forces. The known structure is particularly susceptible to shear, as might be expected from the lack of hydrogen bonding between the sheets.

The energy differences between the known and the other low-energy structures in Table 6 are small compared with the many approximations in the computational model. In this case, the most important approximation is probably that the conformation of water and 5-azauracil has not been optimized within each crystal structure, particularly since there is an 18 kJ mol^{-1} difference in lattice energy between the ExpMinExp and ExpMinOpt minimizations of the experimental 5-azauracil monohydrate crystal structure (Table 5). Figure 3 shows how these small molecular structure differences lead to minor changes in the hydrogen-bonding geometry and yet significant energy differences. Also, it is notable that all hypothetical crystal structures based on sheets (Figure 4), other than the experimental structure, have head-to-head carbonyl $\text{O} \cdots \text{O}$ distances less than 3 \AA , which are rather short according to a recent estimate of the oxygen van der Waals radius⁶⁸ of 1.58 \AA , though within previous estimates.⁶⁴ Since such head-to-head interactions differ from

Table 5. Some Hydrates with Corresponding Anhydrate Structures and Results of the Lattice Energy Minimization Calculations of the Intermolecular Potential^a

Compound	Experimental information		ExpMinExp Minimizations			ExpMin(Con)Opt Minimizations		
	Anhydrate	Monohydrate	Anhydrate	Monohydrate	Stab. E	Anhydrate	Monohydrate	Stab. E
5-azauracil 	XERBEB $F_{\text{exp}} = 38$; $F_{\text{opt}} = 72$	HOQHAW $F_{\text{exp}} = 86$; $F_{\text{opt}} = 99$	-117.41	-178.26	-5.38	-109.3	-160.7	1.68
5-nitouracil 	NIMFOE $F_{\text{exp}} = 45$; $F_{\text{opt}} = 51$	NURMAH $F_{\text{exp}} = 29$; $F_{\text{opt}} = 72$	-124.92	-183.7	-3.30	<i>-116.6</i>	<i>-163.8</i>	5.88
	NIMFOE01 $F_{\text{exp}} = 75$; $F_{\text{opt}} = 98$		-124.26	-183.7	-3.96	<i>-115.54</i>	<i>-163.8</i>	4.78
	NIMFOE01 $F_{\text{exp}} = 12$; $F_{\text{opt}} = 20$		-129.94	-183.7	1.72	<i>-119.34</i>	<i>-163.8</i>	8.62
5-fluorocytosine 	MEBQE01 $F_{\text{exp}} = 4$; $F_{\text{opt}} = 7$	BIRMEU01 $F_{\text{exp}} = 41$; $F_{\text{opt}} = 60$	-127.59	-181.2	1.88	<i>-123.68</i>	<i>-163.9</i>	12.86
	MEBQE01 $F_{\text{exp}} = 4$; $F_{\text{opt}} = 11$		-132.84	-181.2	7.12	-117.12	-163.9	6.29
3-amino-5-nitro-1,2,4-triazole 	JOWWIB $F_{\text{exp}} = 58$; $F_{\text{opt}} = 187$	JYWET $F_{\text{exp}} = 187$; $F_{\text{opt}} = 109$	-137.06	-193.16	-0.62	-117.82	-169.84	1.06
4-hydroxycoumarin 	Anhydrate 2 ⁶⁷ $F_{\text{exp}} = 131$; $F_{\text{opt}} = 178$	HOXCUM01 $F_{\text{exp}} = 73$; $F_{\text{opt}} = 89$	-112.5	-156.54	11.43	-106.85	-146.44	13.48
	Anhydrate 3 ⁶⁷ $F_{\text{exp}} = 38$; $F_{\text{opt}} = 55$		-110.24	-156.54	9.17	-105.98	-146.44	12.61
Cytosine 	CYTSIN $F_{\text{exp}} = 67$; $F_{\text{opt}} = 10$	CYTOSM $F_{\text{exp}} = 183$; $F_{\text{opt}} = 812$	-135.75	-182.46	8.76	-121.54	-161.3	13.32
Thymine 	THYMIN01 $F_{\text{exp}} = 28$; $F_{\text{opt}} = 16$	THYMMH $F_{\text{exp}} = 1308$; $F_{\text{opt}} = 1633$	-110.15	-162.32	3.30	-104.1	-155.22	1.95

^a The CSD refcode for the specific crystal structure used is given, with the F -values for the ExpMinExp (F_{exp}) and ExpMinOpt (F_{opt}) or ExpMinConOpt (F_{opt}) lattice energy minimization with the different rigid molecule structures. Stab. E is the difference between the calculated hydrate lattice energy and the sum of the anhydrate and ice XI lattice energies, with a negative sign implying that the monohydrate is more stable than this extrapolated value. Note that the lattice energies of ice XI from Table 3 are used, i.e., the water geometry differs between the monohydrates and ice XI for the ExpMinExp values. Values in italics are derived using ExpMinConOpt lattice energies, with the torsions specified in section 2.2 fixed.

commonly found attractive carbonyl...carbonyl geometries⁶⁹ (which are well modeled by this potential in cases such as alloxan),⁷⁰ there could well be an underestimate of the repulsion in this contact in the two unobserved sheet motifs.

4. Discussion

The computational search strategy and lattice energy modeling developed here are certainly successful at generating a range of plausible crystal structures for 5-azauracil monohydrate, including the known one, within a small energy range. It is tempting to assume that crystallographic experience might lead to the discounting of the alternative structures based on either the observed preferences for

hydrate hydrogen-bonding geometries¹² or the unusual carbonyl interactions in the alternative sheets and so favor the observed sheet structure. However, crystallographic intuition in selecting plausible structures is far from reliable.⁷¹ While this study did not predict the known monohydrate structure as the most thermodynamically stable in free energy, it may well have been selected as one of the three most plausible structures on this basis and hence would have been a successful prediction by the rules of the international blind tests.⁷²

It is highly likely that the observed structure is the thermodynamically most stable at ambient conditions, though it is possible that the observed structure is kinetically favored,

Table 6. Summary of the Low-Energy 5-Azauracil Monohydrate Predicted Structures, in Order of Lattice Energy Stability

structure ^a	N_{find}	lattice energy (kJ mol ⁻¹)	free energy ^b	free energy rank	space group	a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)	shear ^c (GPa)	hydrogen bond motif
A1_3_c ad/32	38	-165.0	-166.1	1	<i>Pc</i>	5.064	7.084	7.275	90	90.61	90	0.4	sheet 2
A2_3_b ad/1	31	-164.2	-165.7	2	<i>Pc</i>	5.499	6.972	6.844	90	93.71	90	0.3	sheet 2
A1_2_d da/92	29	-164.1	-164.2	10	<i>Cc</i>	7.337	7.039	10.04	90	90.31	90	1.5	sheet 2
A3_1_c ad/9	1	-162.5	-165.3	4	<i>Pc</i>	3.683	5.706	12.646	90	100.22	90	1.1	not sheet
A2_1_d aa/82	65	-162.4	-165.1	6	<i>P1</i>	3.701	5.762	6.172	102.43	94.7	91.35	3.8	sheet 1
A2_2_e da/110	46	-162.3	-164.7	8	<i>Cc</i>	11.934	7.403	10.183	90	144.497	90	1.5	sheet 1
A3_1_f ad/17	25	-162.3	-164.7	7	<i>Pc</i>	5.04	7.413	7.007	90	93.38	90	1.1	sheet 1
A3_1_d aa/93	66	-162.2	-163.8	13	<i>P1</i>	4.895	4.995	5.361	92.44	92.01	90.78	5.5	sheet 1
A1_2_b ah/65	64	-162.1	-165.2	5	<i>P2₁</i>	3.625	12.149	5.802	90	91.11	90	1.7	not sheet
A3_1_f ad/31	1	-162.0	-163.3	16	<i>P2₁</i>	4.716	5.261	10.623	90	96.65	90	7.5	not sheet
A3_1_c da/66	17	-161.9	-163.2	17	<i>Cc</i>	6.998	7.406	10.113	90	93.09	90	1.2	sheet 1
A2_4_b da/75	51	-161.9	-163.0	18	<i>Cc</i>	12.858	6.906	6.789	90	119.13	90	1.7	sheet 1
A2_2_d da/97	29	-161.4	-162.7	22	<i>Cc</i>	7.186	7.012	10.349	90	90.88	90	1.5	sheet 2
A2_2_b da/84	41	-161.3	-162.4	23	<i>Cc</i>	12.66	6.796	7.194	90	123.67	90	1.1	sheet 1
A2_2_b ab/123	63	-161.1	-163.9	11	<i>P-1</i>	6.92	6.871	6.911	104.06	109.32	113.49	3.0	sheet 1
A3_1_e ad/104	11	-161.1	-163.4	14	<i>Pc</i>	3.811	7.475	9.066	90	91.11	90	2.8	sheet 1
A2_3_e de/8	21	-160.9	-164.3	9	<i>C2/c</i>	11.535	7.57	15.747	90	131.81	90	0.6	sheet 1
A2_4_b ai/7	15	-160.8	-163.9	12	<i>P2₁/c</i>	7.227	6.946	10.125	90	92.25	90	1.8	not sheet
A1_3_d dc/13	40	-160.8	-163.3	15	<i>C2/c</i>	11.598	7.55	11.786	90	93.94	90	2.8	sheet 1
A2_4_a fa/78	34	-160.8	-163.0	19	<i>P2₁/c</i>	6.705	11.706	9.281	90	130.7	90	2.9	not sheet
A3_1_e da/57	12	-160.8	-162.8	21	<i>Cc</i>	11.676	7.482	17.437	90	160.208	90	2.4	sheet 1
A2_2_c am/86	5	-160.7	-163.0	20	<i>P2₁/c</i>	6.935	12.468	6.966	90	120.13	90	1.4	sheet 2
A1_1_a ab/19	50	-160.7	-165.3	3	<i>P2₁/m</i>	6.584	5.775	7.065	90	101.8	90	0.4	sheet expt

^a Structures are denoted by the acceptor, the placement of the water, and the orientation of the water, MOLPAK packing type and number of one example of the cluster, with N_{find} being the number of searches (max 66) in which this crystal structure was found at least once. ^b An estimate of the Helmholtz free energy at 298 K derived from the second derivative properties at the lattice energy minimum, with free energy rank being the order of this stability. ^c The lowest eigenvalue of the shear submatrix of the elastic tensor. Bold type corresponds to the experimental structure (ExpMinOpt).

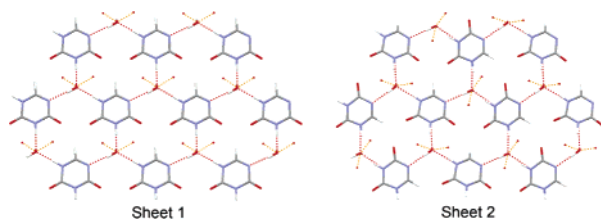


Figure 4. Hydrogen bonding in hypothetical sheets 1 and 2. Orange dashed lines show hydrogen bonds out of the plane of the sheet to molecules in the sheet below.

if fragments of the observed sheet formed during the nucleation process. However, the idea that water of hydration gets trapped into the hydrate structure is less compelling for this sheet than for a structure in which the water had multipoint hydrogen bond contacts to a single solute molecule.⁷³ Thus, we should assess the uncertainties in calculating the relative thermodynamic stabilities of the structures in Table 6. The application of the rigid body approximation is a significant limitation even for the hydrate molecules considered here, as demonstrated by the ExpMinOpt vs ExpMinExp energies in Tables 3 and 5. While methods of *ab initio* optimization of the molecular structure under the influence of the packing forces have been successfully applied for variations in single bond torsion angles to optimize similar strength intermolecular hydrogen bonds,⁷⁴ getting the inter- and intramolecular energy balance correct for these more rigid molecules would be challenging, particularly because of the limited confidence that can be

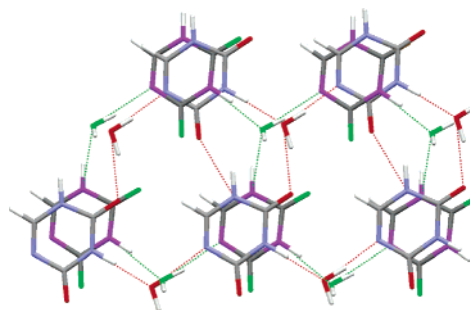


Figure 5. Overlay of the experimental ExpMinOpt sheet with the predicted sheet 1 motif. Carbon and hydrogen atoms are colored gray and white in both structures, with oxygen red and nitrogen blue in the experimental ExpMinOpt and green and purple in the lowest energy sheet 1 structure, A2_1_d aa/82.

placed in the intermolecular potential used. While the superiority of the water...water model to simple models developed for liquid simulation has been clearly demonstrated for lattice energy minimization modeling, the repulsion—dispersion potential has a poor basis compared with more modern water potentials with more complex potential forms. The close carbonyl...carbonyl head on contact found in the most stable predicted structures is unlikely to have been sampled during the initial empirical parameter fitting to oxohydrocarbons or during subsequent validation in predominantly heteroatomic environments within organic crystal structures. Since this oxygen repulsion potential

and all the water...organic molecule interactions are derived from the geometric combining rules, with their poor theoretical basis,⁶³ it is encouraging that this starting point for development of the model potential has worked so well. The predominant reason for this success is surely that the distributed multipoles electrostatic model reflects the charge distribution of the molecules realistically. Developing more accurate intermolecular potentials including the effect of polarization,⁷⁵ and appropriately modeling molecular flexibility will require considerably more research as well as computational resources for their implementation. However, it is plausible that improvements in the model potential and allowing the water and 5-azauracil geometries to relax under the crystal packing forces may well stabilize the observed structure over the alternatives.

There is no reason to expect that the intermolecular potentials developed in this study will not be suitable for similar studies to suggest possible hydrate structures for other small organic molecules, since so many experimental hydrate structures are adequately reproduced. The success of the search strategy employed for 5-azauracil suggests that it could be thoughtfully adapted to a range of other molecules with competing hydrogen bond donors and acceptors to investigate ordered hydrogen-bonded hydrate structures. Indeed, far fewer input clusters would be required to contrast different possible hydrate hydrogen-bonding motifs as a complement to an experimental solid form screen.⁷⁶ More mathematically complete search algorithms for two (or more) independent molecules in the asymmetric unit cell^{23,24} would avoid the need to carefully consider the hydrogen-bonding possibilities for the molecules' functional groups and be capable of finding a wider range of structures. Indeed the crystal structure of 5-azauracil monohydrate can be readily found with the search methods of van Eijck²⁵ and Karamertzanis,²³ although the relative lattice energies of the structures found with point charge electrostatic models are significantly reordered when reminimized with the distributed multipole model developed in this study.

Although this work represents progress in the prediction of the crystal structures of a specific type of hydrate, it clearly demonstrates that the accuracy of the lattice energy modeling is not capable of predicting whether a hydrate will form on energetic grounds. The comparison of the lattice energies of the hydrates compared with the extrapolation from the corresponding lattice energies for the anhydrate and ice (Table 5) does not suggest such a marked stabilization compared with the uncertainties in the thermodynamic modeling that the formation of the observed hydrate is clearly predicted. However, this analysis has necessarily been limited to molecules where both the hydrate and anhydrate crystallize sufficiently readily to produce single crystals suitable for X-ray determination.

5. Conclusion

A methodology currently used for organic crystal structure prediction has been adapted to monohydrates and shown to be successful in generating plausible 5-azauracil monohydrate structures with the known crystal structure found

within 5 kJ mol⁻¹ of the global minimum in lattice energy, 1 kJ mol⁻¹ in free energy. The approach is certainly suitable for determining the range of plausible hydrogen-bonding motifs for a hydrate of a rigid polar molecule. However, further developments in the thermodynamic modeling are required to predict the formation of a hydrate in preference to an anhydrate. Nevertheless, computational studies clearly have potential in understanding hydrate formation.

The computed crystal structures for 5-azauracil monohydrate are stored on CCLRC e-Science Centre data portal and are available from the authors.

Acknowledgment. ATH was funded by the Basic Technology program of the Research Councils UK through the project 'Control and Prediction of the Organic Solid State'. Dr. Panagiotis Karamertzanis is thanked for useful discussions and computational assistance, Dr. Bouke van Eijck for contrasting search methods, and Dr. Louise Price for assistance in preparing the manuscript.

Supporting Information Available: CSD analysis of hydrogen bonding of water to specific acceptors, detailed lattice energy minimizations for the ice and hydrate structures, an analysis of the occurrence of the low-energy structures from the different bimolecular cluster searches, and a list of the final potential parameters. This material 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*, 2435–2460.
- (2) Morris, K. R. Structural aspects of hydrates and solvates. In *Polymorphism in Pharmaceutical Solids*; Brittain, H. G., Ed.; Marcel Dekker, Inc.: New York, 1999; Vol. 95, pp 125–181.
- (3) Infantes, L.; Chisholm, J.; Motherwell, S. Extended motifs from water and chemical functional groups in organic molecular crystals. *CrystEngComm* **2003**, *5*, 480–486.
- (4) Khankari, R. K.; Grant, D. J. W. Pharmaceutical Hydrates. *Thermochim. Acta* **1995**, *248*, 61–79.
- (5) Byrn, S. R.; Pfeiffer, R. R.; Stowell, J. G. *Solid-state Chemistry of Drugs*; 2nd ed.; SSCI Inc.: West Lafayette, IN, 1999.
- (6) Price, S. L. The computational prediction of pharmaceutical crystal structures and polymorphism. *Adv. Drug Delivery Rev.* **2004**, *56*, 301–319.
- (7) Infantes, L. Water Clusters in Organic Molecular Crystals. *CrystEngComm* **2002**, *4*, 454–461.
- (8) Etter, M. C.; MacDonald, J. C.; Bernstein, J. Graph-Set Analysis of Hydrogen-Bond Patterns in Organic- Crystals. *Acta Crystallogr., Sect. B: Struct. Sci.* **1990**, *46*, 256–262.
- (9) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380–388.
- (10) Gillon, A. L.; Feeder, N.; Davey, R. J.; Storey, R. Hydration in molecular crystals - A Cambridge Structural Database analysis. *Cryst. Growth Des.* **2003**, *3*, 663–673.
- (11) Desiraju, G. R. Hydration in Organic Crystals: Prediction from Molecular Structure. *Chem. Commun.* **1991**, 426–428.

- (12) Infantes, L.; Fabian, L.; Motherwell, W. D. S. Organic crystal hydrates: what are the important factors for formation. *CrystEngComm* **2007**, *9*, 65–71.
- (13) Thanki, N.; Thornton, J. M.; Goodfellow, J. M. Distributions of Water Around Amino-Acid Residues in Proteins. *J. Mol. Biol.* **1988**, *202*, 637–657.
- (14) Gorbitz, C. H.; Hersleth, H. P. On the inclusion of solvent molecules in the crystal structures of organic compounds. *Acta Crystallogr., Sect. B: Struct. Sci.* **2000**, *56*, 526–534.
- (15) Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M. Role of Electrostatic Interactions in Determining the Crystal Structures of Polar Organic Molecules. A Distributed Multipole Study. *J. Phys. Chem.* **1996**, *100*, 7352–7360.
- (16) Guillot, B. A reappraisal of what we have learnt during three decades of computer simulations on water. *J. Mol. Liq.* **2002**, *101*, 219–260.
- (17) Abascal, J. L. F.; Sanz, E.; Fernandez, R. G.; Vega, C. A potential for the study of ices and amorphous water: TIP4P/Ice. *J. Chem. Phys.* **2005**, *122*, 234511.
- (18) Price, S. L. Quantifying intermolecular interactions and their use in computational crystal structure prediction. *CrystEngComm* **2004**, *6*, 344–353.
- (19) Stone, A. J.; Alderton, M. Distributed Multipole Analysis - Methods and Applications. *Mol. Phys.* **1985**, *56*, 1047–1064.
- (20) Millot, C.; Soetens, J. C.; Costa, M. T. C. M.; Hodges, M. P.; Stone, A. J. Revised anisotropic site potentials for the water dimer and calculated properties. *J. Phys. Chem. A* **1998**, *102*, 754–770.
- (21) Bukowski, R.; Szalewicz, K.; Groenenboom, G.; van der Avoird, A. Interaction potential for water dimer from symmetry-adapted perturbation theory based on density functional description of monomers. *J. Chem. Phys.* **2006**, *125*, 044301.
- (22) Karamertzanis, P. G.; Price, S. L. Challenges of crystal structure prediction of diastereomeric salt pairs. *J. Phys. Chem. B* **2005**, *109*, 17134–17150.
- (23) Karamertzanis, P. G.; Pantelides, C. C. Ab initio crystal structure prediction - I. Rigid molecules. *J. Comput. Chem.* **2005**, *26*, 304–324.
- (24) Bazterra, V. E.; Thorley, M.; Ferraro, M. B.; Facelli, J. C. A Distributed Computing Method for Crystal Structure Prediction of Flexible Molecules: An Application to N-(2-Dimethyl-4,5-dinitrophenyl) Acetamide. *J. Chem. Theory Comput.* **2007**, *3*, 201–209.
- (25) van Eijck, B. P.; Kroon, J. Structure predictions allowing more than one molecule in the asymmetric unit. *Acta Crystallogr., Sect. B: Struct. Sci.* **2000**, *56*, 535–542.
- (26) Kamb, B. Ice II: A Proton-Ordered Form of Ice. *Acta Crystallogr.* **1964**, *17*, 1437–1449.
- (27) Kamb, B.; Hamilton, W. C.; LaPlaca, S. J.; Prakash, A. Ordered Proton Configuration in ice II, from Single-Crystal Neutron Diffraction. *J. Chem. Phys.* **1971**, *55*, 1934–1945.
- (28) Kuhs, W. F.; Finney, J. L.; Vettier, C.; Bliss, D. V. Structure and hydrogen ordering in Ices VI, VII, VIII by neutron powder diffraction. *J. Chem. Phys.* **1984**, *81*, 3612–3623.
- (29) La Placa, S. J.; Hamilton, W. C.; Kamb, B.; Prakash, A. On a nearly proton-ordered structure for ice IX. *J. Chem. Phys.* **1972**, *58*, 567–580.
- (30) Jackson, S. M.; Nield, V. M.; Whitworth, R. W.; Oguro, M.; Wilson, C. C. Single-Crystal Neutron Diffraction Studies of the Structure of Ice XI. *J. Phys. Chem. B* **1997**, *101*, 6142–6145.
- (31) Leadbetter, A. J.; Ward, R. C.; Clark, J. W.; Tucker, P. A.; Matsuo, T.; Suga, H. The equilibrium low-temperature structure of ice. *J. Chem. Phys.* **1985**, *82*, 424–428.
- (32) Salzmann, C. G.; Radaelli, P. G.; Hallbrucker, A.; Mayer, E.; Finney, J. L. The Preparation and Structures of Hydrogen Ordered Phases of Ice. *Science* **2006**, *311*, 1758–1761.
- (33) Finney, J. L. What's so special about water? *Philos. Trans. R. Soc. London, Ser. B* **2004**, *359*, 1145–1165.
- (34) Willock, D. J.; Price, S. L.; Leslie, M.; Catlow, C. R. A. The Relaxation of Molecular Crystal Structures Using a Distributed Multipole Electrostatic Model. *J. Comput. Chem.* **1995**, *16*, 628–647.
- (35) Spek, A. L. *PLATON - a multipurpose crystallographic tool*; Utrecht University: Utrecht, 2001.
- (36) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Dordrecht, 1981; p 331.
- (37) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
- (38) Boulougouris, G. C.; Economou, I. G.; Theodorou, D. N. Engineering a Molecular Potential for Water Phase Equilibrium over a Wide Temperature Range. *J. Phys. Chem. B* **1998**, *102*, 1029–1035.
- (39) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (40) Errington, J. R.; Panagiotopoulos, A. Z. A Fixed Point Charge Model for Water Optimised to the Vapour-Liquid Coexistence Properties. *J. Phys. Chem. B* **1998**, *102*, 7470–7475.
- (41) Cox, S. R.; Hsu, L. Y.; Williams, D. E. Nonbonded Potential Function Models for Crystalline Oxohydrocarbons. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1981**, *37*, 293–301.
- (42) Beyer, T.; Price, S. L. Dimer or catemer? Low-energy crystal packings for small carboxylic acids. *J. Phys. Chem. B* **2000**, *104*, 2647–2655.
- (43) Breneman, C. M.; Wiberg, K. B. Determining Atom-Centered Monopoles From Molecular Electrostatic Potentials - the Need For High Sampling Density in Formamide Conformational-Analysis. *J. Comput. Chem.* **1990**, *11*, 361–373.
- (44) Filippini, G.; Gavezzotti, A. Empirical Intermolecular Potentials For Organic-Crystals: the '6-Exp' Approximation Revisited. *Acta Crystallogr., Sect. B: Struct. Sci.* **1993**, *49*, 868–880.
- (45) Williams, D. E.; Cox, S. R. Nonbonded Potentials For Azahydrocarbons: the Importance of the Coulombic Interaction. *Acta Crystallogr., Sect. B: Struct. Sci.* **1984**, *40*, 404–417.
- (46) Williams, D. E.; Houpt, D. J. Fluorine Nonbonded Potential Parameters Derived From Crystalline Perfluorocarbons. *Acta Crystallogr., Sect. B: Struct. Sci.* **1986**, *42*, 286–295.
- (47) Price, S. L.; Price, L. S. Modelling Intermolecular Forces for Organic Crystal Structure Prediction. In *Intermolecular Forces and Clusters I*; Wales, D. J., Ed.; Springer-Verlag: Berlin, Heidelberg, Germany, 2005; pp 81–123.

- (48) Hulme, A. T.; Price, S. L.; Tocher, D. A. A New Polymorph of 5-Fluorouracil Found Following Computational Crystal Structure Predictions. *J. Am. Chem. Soc.* **2005**, *127*, 1116–1117.
- (49) Hulme, A. T.; Tocher, D. A. The Discovery of New Crystal Forms of 5-Fluorocytosine Consistent with the Results of Computational Crystal Structure Prediction. *Cryst. Growth Des.* **2006**, *6*, 481–487.
- (50) Allen, F. H.; Kennard, O.; Watson, D. G.; Brammer, L.; Orpen, A. G.; Taylor, R. Tables of Bond Lengths Determined By X-Ray and Neutron-Diffraction .1. Bond Lengths in Organic-Compounds. *J. Chem. Soc., Perkin Trans. 2* **1987**, S1–S19.
- (51) Frisch, M. J. et al. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2003.
- (52) Stone, A. J. *GDMA: A Program for Performing Distributed Multipole Analysis of Wave Functions Calculated Using the Gaussian Program System, version 1.0*; University of Cambridge: Cambridge, United Kingdom, 1999.
- (53) Potter, B. S.; Palmer, R. A.; Withnall, R.; Chowdhry, B. Z.; Price, S. L. Aza analogues of nucleic acid bases: experimental determination and computational prediction of the crystal structure of anhydrous 5-azauracil. *J. Mol. Struct.* **1999**, *486*, 349–361.
- (54) Gray, A. E.; Day, G. M.; Leslie, M.; Price, S. L. Dynamics in crystals of rigid organic molecules: contrasting the phonon frequencies calculated by molecular dynamics with harmonic lattice dynamics for finidazole and 5-azauracil. *Mol. Phys.* **2004**, *102*, 1067–1083.
- (55) Holden, J. R.; Du, Z. Y.; Ammon, H. L. Prediction of Possible Crystal-Structures For C-, H-, N-, O- and F-Containing Organic Compounds. *J. Comput. Chem.* **1993**, *14*, 422–437.
- (56) Chisholm, J. A.; Motherwell, S. COMPACK: a program for identifying crystal structure similarity using distances. *J. Appl. Crystallogr.* **2005**, *38*, 228–231.
- (57) de Gelder, R.; Wehrens, R.; Hageman, J. A. A generalized expression for the similarity of spectra: Application to powder diffraction pattern classification. *J. Comput. Chem.* **2001**, *22*, 273–289.
- (58) van de Streek, J.; Motherwell, S. Searching the Cambridge Structural Database for polymorphs. *Acta Crystallogr., Sect. B: Struct. Sci.* **2005**, *61*, 504–510.
- (59) Day, G. M.; Price, S. L.; Leslie, M. Elastic constant calculations for molecular organic crystals. *Cryst. Growth Des.* **2001**, *1*, 13–26.
- (60) Day, G. M.; Price, S. L.; Leslie, M. Atomistic calculations of phonon frequencies and thermodynamic quantities for crystals of rigid organic molecules. *J. Phys. Chem. B* **2003**, *107*, 10919–10933.
- (61) Anghel, A. T.; Day, G. M.; Price, S. L. A study of the known and hypothetical crystal structures of pyridine: why are there four molecules in the asymmetric unit cell? *CrystEngComm* **2002**, *4*, 348–355.
- (62) Buckingham, A. D.; Fowler, P. W.; Stone, A. J. Electrostatic Predictions of Shapes and Properties of Vanderwaals Molecules. *Int. Rev. Phys. Chem.* **1986**, *5*, 107–114.
- (63) Stone, A. J. *The Theory of Intermolecular Forces*; Oxford University Press: Oxford, U.K., 1996.
- (64) Gavezzotti, A. *Molecular Aggregation: Structure Analysis and Molecular Simulation of Crystals and Liquids*; Oxford University Press: Oxford, U.K., 2007; Vol. 19.
- (65) Whalley, E. The difference in the intermolecular forces on H₂O and D₂O. *Trans. Faraday Soc.* **1957**, *53*, 1578–1585.
- (66) Motherwell, W. D. S.; Shields, G. P.; Allen, F. H. Visualization and characterization of non-covalent networks in molecular crystals: automated assignment of graph-set descriptors for asymmetric molecules. *Acta Crystallogr., Sect. B: Struct. Sci.* **1999**, *55*, 1044–1056.
- (67) Hulme, A. T. Experimental and Computational Studies of Polymorphism of Small Organic Molecules, Ph.D. Thesis, University College London, 2007.
- (68) Rowland, R. S.; Taylor, R. Intermolecular Nonbonded Contact Distances in Organic-Crystal Structures - Comparison With Distances Expected From Van-Der-Waals Radii. *J. Phys. Chem.* **1996**, *100*, 7384–7391.
- (69) Allen, F. H.; Baalham, C. A.; Lommerse, J. P. M.; Raithby, P. R. Carbonyl-carbonyl interactions can be competitive with hydrogen bonds. *Acta Crystallogr., Sect. B: Struct. Sci.* **1998**, *54*, 320–329.
- (70) Coombes, D. S.; Nagi, G. K.; Price, S. L. On the lack of hydrogen bonds in the crystal structure of alloxan. *Chem. Phys. Lett.* **1997**, *265*, 532–537.
- (71) Day, G. M.; Motherwell, W. D. S. An experiment in crystal structure prediction by popular vote. *Cryst. Growth Des.* **2006**, *6*, 1985–1990.
- (72) Day, G. M. et al. A third blind test of crystal structure prediction. *Acta Crystallogr., Sect. B: Struct. Sci.* **2005**, *61*, 511–527.
- (73) Nangia, A.; Desiraju, G. R. Pseudopolymorphism: occurrences of hydrogen bonding organic solvents in molecular crystals. *Chem. Commun.* **1999**, 605–606.
- (74) Karamertzanis, P. G.; Price, S. L. Energy Minimization of Crystal Structures Containing Flexible Molecules. *J. Chem. Theory Comput.* **2006**, *2*, 1184–1199.
- (75) Stone, A. J.; Misquitta, A. J. Atom-atom potentials from ab initio calculations. *Int. Rev. Phys. Chem.* **2007**, *26*, 193–222.
- (76) Johnston, A.; Florence, A. J.; Shankland, N.; Kennedy, A. R.; Shankland, K.; Price, S. L. Crystallization and crystal energy landscape of hydrochlorothiazide. *Cryst. Growth Des.* **2007**, *7*, 705–712.

CT700045R

JCTC

Journal of Chemical Theory and Computation

Assessment of Semiempirical Quantum Mechanical Methods for the Evaluation of Protein Structures

Andrew M. Wollacott[†] and Kenneth M. Merz, Jr.*

University of Florida, Gainesville, Florida 32611-8435

Received November 4, 2006

Abstract: The ability to discriminate native structures from computer-generated misfolded ones is key to predicting the three-dimensional structure of a protein from its amino acid sequence. Here we describe an assessment of semiempirical methods for discriminating native protein structures from decoy models. The discrimination of decoys entails an analysis of a large number of protein structures and provides a large-scale validation of quantum mechanical methods and their ability to accurately model proteins. We combine our analysis of semiempirical methods with a comparison of an AMBER force field to discriminate decoys in conjunction with a continuum solvent model. Protein decoys provide a rigorous and reliable benchmark for the evaluation of scoring functions, not only in their ability to accurately identify native structures but also to be computationally tractable to sample a large set of non-native models.

Introduction

The three-dimensional structure of a protein is determined primarily by its amino acid sequence,¹ yet, while this principle is well established, reliable methods for the prediction of protein tertiary structure from primary structure have not yet been developed.² Current efforts for protein structure prediction have focused on homology modeling, threading/fold recognition, and ab initio folding, all of which share the thermodynamic hypothesis that the native three-dimensional conformation has the lowest free energy in comparison to non-native or misfolded structures.³ While current progress has proved extremely promising,⁴ ab initio folding methods cannot be consistently applied to successfully predict the fold of any given sequence.^{5,6} In ab initio folding methods, not only must a very large conformational space be sampled but also it is particularly important to be able to identify native folds from those non-native structures that are generated.⁷

The development of scoring functions to be used in the studies of biological macromolecules is still ongoing, with

the focus being on either fast, less accurate methods or on high accuracy methods. The choice of a potential energy function in protein modeling depends on the type of simulation performed and the size of the system being modeled. Developing reliable tests for a scoring function, therefore, remains an important aspect of computational modeling.

In order to provide an objective manner with which to evaluate scoring potentials, sets of computationally misfolded models of proteins are typically used. The ability to compare native structures to available decoy sets allows for a common and relatively unbiased benchmark for evaluating scoring functions.⁸ Analyzing decoys allows a potential energy function to be tested for its ability to discriminate native protein structures from a large ensemble of non-native models,⁹ which is important for structure prediction methods such as homology modeling, threading/folding recognition, and ab initio folding. In these cases, the native conformation is expected to have the lowest free energy.

Atomic-based protein scoring potentials are employed for modeling structures at higher resolutions.¹⁰ These potentials are physics-based and fall under either a molecular mechanics (MM) type scoring function or a quantum mechanical (QM) based function. MM based functions for studying proteins include AMBER,¹¹ CHARMM,¹² and OPLS,¹³ among others. These potentials can be used for performing either molecular

* Corresponding author e-mail: merz@qtp.ufl.edu. Corresponding author address: Department of Chemistry, Quantum Theory Project, 2328 New Physics Building, P.O. Box 118435, University of Florida, Gainesville, FL 32611-8435.

[†] Current address: University of Washington, Seattle, WA.

Table 1. Select Decoy Sets Used

decoy set	PDB	description	type	N_{decoys}	N_{res}	N_{atoms}	rmsd range (Å)	%H/%E ^b
4-state-reduced	1ctf	C-terminal domain of ribosomal protein L7/L12	X-ray	630	68	1005	2.1–9.8	53/26
	1r69	N-terminal domain of phage 434 repressor	X-ray	675	63	997	2.2–9.4	70/0
	1sn3	scorpion toxin variant 3	X-ray	660	65	948	2.5–10.3	12/22
	2cro	phage 434 Cro protein	X-ray	674	65	1081	2.0–9.5	66/0
	4pti	trypsin inhibitor	X-ray	687	58	892	2.8–10.7	16/24
	4rxn	rubredoxin	X-ray	677	54	794	2.5–9.2	0/20
Rosetta	1gb1	immunoglobulin binding domain of streptococcal protein G	NMR	999	54	823	3.1–18.0	28/33
	1hsn	high mobility group protein I box	NMR	999	62	1014	4.1–17.6	79/0
	1orc	Cro repressor (mutant)	X-ray	999	56	877	4.0–14.1	46/27
	1pgx	protein G (B2 domain)	X-ray	999	57	873	3.4–20.4	26/46
	1uxd	fructose repressor DNA-binding domain	NMR	999	43	690	2.2–12.5	81/0
	2fow	RNA binding domain of ribosomal protein LII	NMR	999	66	1009	4.0–21.6	52/9
	1hc8 ^a	RNA binding domain of ribosomal protein LII	X-ray	999	66	1009	4.0–21.6	53/9
	1r69	N-terminal domain of phage 434-repressor	X-ray	999	61	976	3.1–15.5	70/0

^a 1hc8 is the X-ray structure for the 2fow decoy set. ^b Percentage of helices and sheets in the native.

dynamics simulations or energy minimizations of structures. MM potentials have been parametrized on small molecules modeled at the quantum mechanical level or from liquid simulations, and these parameters can subsequently be extended to larger biological systems, with generally satisfactory results.

Until recently, a full QM treatment of protein structures has not been feasible. The linear-scaling semiempirical package, DivCon,¹⁴ has been developed in our laboratory and has allowed for large biological systems to be studied at the QM level. The major problem with applying higher levels of theory to biologically relevant systems is their poor scaling. Semiempirical methods scale as N^3 , where N is the number of basis functions used to represent electrons in the system.¹⁵ With judicious use of cutoffs, and by applying a divide-and-conquer¹⁶ approach to large systems, calculations performed with DivCon are able to scale linearly with system size.¹⁷

The QM treatment of proteins has several advantages over MM based methods. The point charge model used in MM packages ignores higher order effects such as charge transfer and polarization.¹⁸ These effects have not been widely incorporated into MM functions so charge interactions can be more accurately modeled with QM methods. The utility of semiempirical methods, such as those used in DivCon, has already been shown for studying the electrostatic interactions in protein-folding and protein–ligand interactions.^{19,20} However, this study marks the first large-scale investigation into the utility of semiempirical quantum mechanical methods for studying protein structures.

Approach

Decoy Sets. There are various protein decoy sets readily available for public use, and these vary in quality depending upon the method of generation. Two of the more popular decoy sets include the 4-state-reduced set models created by Park and Levitt⁷ and the Rosetta decoy set, produced by Simmons and Baker.²¹

The 4-state-reduced decoy set from Levitt has been widely used as a means of evaluating scoring potentials and is considered to be a high-quality decoy set.²² The set consists

of six small proteins whose structures have been solved by X-ray crystallography and are considered well refined. During the generation of decoys, the secondary structure of the native structure was held constant by altering only hinge regions. A set of loop residues between segments of defined secondary structure was chosen and their conformation was enumerated by varying the ϕ, ψ torsion angles. Only a subset of 10 residues was modified, with only four predefined states of the backbone torsional angles sampled. The large number of conformations generated was reduced by applying a radius of gyration cutoff as well as removing structures that contained bad clashes in the reduced representation.

The proteins in the 4-state-reduced set, listed in Table 1, possessed small compact native folds with between 54 and 68 residues and represent a diverse set of small proteins. The original set also included a calcium-binding protein (3icb) which chelates two calcium ions with carbonyl and carboxyl groups from side chains and the protein backbone. Since decoy structures were generated without calcium ions bound, this set was removed as the unbound native structure would be unfairly penalized by having a large concentration of negative charge near the binding sites.

Another high quality collection of decoys is the Rosetta decoy set, published by Baker and co-workers.²³ In the Rosetta set, nativelike structures were created by assembling fragments of nonhomologous protein structures containing similar local sequences to that of the native structure. Relative to the 4-state-reduced set, this allows for a broader range of conformational space to be explored. During decoy generation for both data sets, the structures were created in a reduced representation, with all heavy atoms for the protein backbone, and a C^β atom for the side chains. A subset of these structures, possessing more than 40 residues having structures with an rmsd less than 5.0 Å to the native, was chosen for this study (Table 1).

Scoring Functions. For this study, we are interested in the energetic differences between individual protein states, more specifically the energy gap between a protein decoy and its native structure. Since protein decoys have the same sequence as their native structure the two will have the same unfolded state, so the calculated energy of each structure

can be represented as an effective free energy as described in previous decoy studies.¹⁰ Thus, the energy gap between a native structure and its decoy can be represented as

$$\Delta\Delta G_{\text{eff}} = \Delta G_{\text{eff}}^{\text{native}} - \Delta G_{\text{eff}}^{\text{decoy}} \quad (1)$$

Using an MM potential, in this case AMBER,¹¹ with the PARM94 parameter set, the energy of each structure was given as the sum of its geometric (bond lengths, bond angles, and torsions), van der Waals, electrostatic, and solvation energies (using a generalized Born approximation).

The heat of formation of a model protein as calculated using semiempirical methods was used as part of the energy term in the DivScore potential. The DivScore potential is a linear combination of the heat of formation, the solvation energy, and the attractive term of the Lennard-Jones energy function (eq 2).

$$\Delta G_{\text{tot}} = \Delta H_{\text{f}} + \Delta G_{\text{solv}} + \sum \text{LJ}_6 \quad (2)$$

The solvation energy was determined by solving the finite difference Poisson Boltzmann equation,²⁴ and the attractive portion of the Lennard-Jones term was taken from AMBER. This attractive potential was used to compensate for the poor treatment of dispersive effects by semiempirical methods. The heat of formation was determined at the AM1²⁵ and PM3²⁶ levels.

Here we compare the ability of AMBER and DivScore to identify native structures from non-native models. AMBER was chosen, not because it is the best MM-based method for decoy discrimination, but because it is a commonly used force field for studying proteins. While there may be other MM force fields or statistical-based potentials that outperform AMBER, this comparison is only used to validate the use of semiempirical methods for evaluating protein structures.

Results

Scoring Structures “As-Is”. We have scored the available protein decoy sets using decoy heavy atom coordinates as-is (referred to as hydrogen-minimized since only hydrogen atoms were optimized). The results of scoring hydrogen-minimized decoys with both the AMBER and DivScore potentials are listed in the Supporting Information, Tables S1 and S2. Overall, both AMBER and DivScore demonstrate poor discrimination of decoys by score as a function of rmsd. Since the protein decoys have been created computationally using various force fields, scoring the decoy structures relative to the unmodified crystal structure introduces bias. The decoys may have geometries that are better suited to the force field, improving their score relative to the native structure.

In these studies, we aim to obtain a ranking of 1 for the native structure, which signifies that the native structure can be reliably identified from its decoys. The Z-score (eq 3), which provides a measure of the native structure’s energy compared to all decoys, is also tabulated. It is desirable that the energy gap between the best scoring decoy and the native structure be sufficiently large to clearly identify the native structure as the best model. Ideally, a good scoring function

Table 2. Decoy Ranking for All-Atom Minimized Structures Using AMBER

decoy set	system	rank _{nat}	Z-score _{nat}	$\Delta\Delta G_{\text{eff}}$ (kcal/mol)	rmsd _{best} ^a (Å)
4-state-reduced	1ctf	1	-3.85	-71.59	2.62
	1r69	1	-2.04	-77.79	3.38
	1sn3	1	-5.97	-102.70	3.29
	2cro	1	-3.90	-64.01	2.07
	4pti	1	-4.93	-51.62	2.78
	4rxn	1	-4.66	-61.61	2.61
Rosetta	1gb1	1	-4.53	-29.15	8.31
	1hsn	36	-1.66	38.20	10.20
	1orc	11	-2.45	14.56	8.65
	1pgx	1	-5.00	-32.45	4.55
	1uxd	4	-2.46	4.76	2.81
	2fow	80	-1.39	34.76	7.73
	1hc8	7	-2.34	13.06	7.73
1r69	1	-6.81	-63.71	7.80	

^a rmsd of the best scoring decoy.

should score native-like decoys more favorably than structurally dissimilar models.

$$Z_{\text{score}} = \frac{\Delta G_{\text{nat}} - \overline{\Delta G}}{\sigma} \quad (3)$$

Scoring Minimized Structures with AMBER. Because of the potential for bias when comparing structures generated through different means, an all-atom gradient-based minimization was performed on all decoys and native structures. Minimizing structures serves to clean up the models from any structural anomalies in a consistent fashion. van der Waals clashes are removed during optimization due to the large forces resulting from the steepness of the repulsive terms of the potential. In addition, bond lengths and angles are minimized with a consistent parametrized potential, removing any bias in the force field toward the native structure or its decoys.

The results of scoring with AMBER for the all-atom minimized structures are shown in Table 2 and illustrated in Figure 1. As indicated, the native structure was identified in all cases for the 4-state-reduced decoy sets. The native structure had a very favorable Z-score compared to the ensemble of decoys, in addition to a large energy gap separating the native structure from the highest scoring decoy. The rmsd of the best scoring decoy compared to the native state was low, being only 2.79 Å on average, while the lowest rmsd possible was, on average, 2.35 Å. Overall AMBER performs very well for identifying the native structure for the 4-state-reduced set, in agreement with previous studies of MM-based potentials.^{10,27,28} This suggests that MM potentials with implicit solvent perform well for decoy discrimination, regardless of the parameter set used.

While AMBER performed well for Levitt’s 4-state-reduced set, it was unable to identify the native structure in four cases for the Rosetta set: 1hsn, 1orc, 1uxd, and 2fow. Previous studies have shown that Rosetta decoys are a more challenging set for MM-based potentials.²⁹ Additionally, for the cases where AMBER was able to identify the native

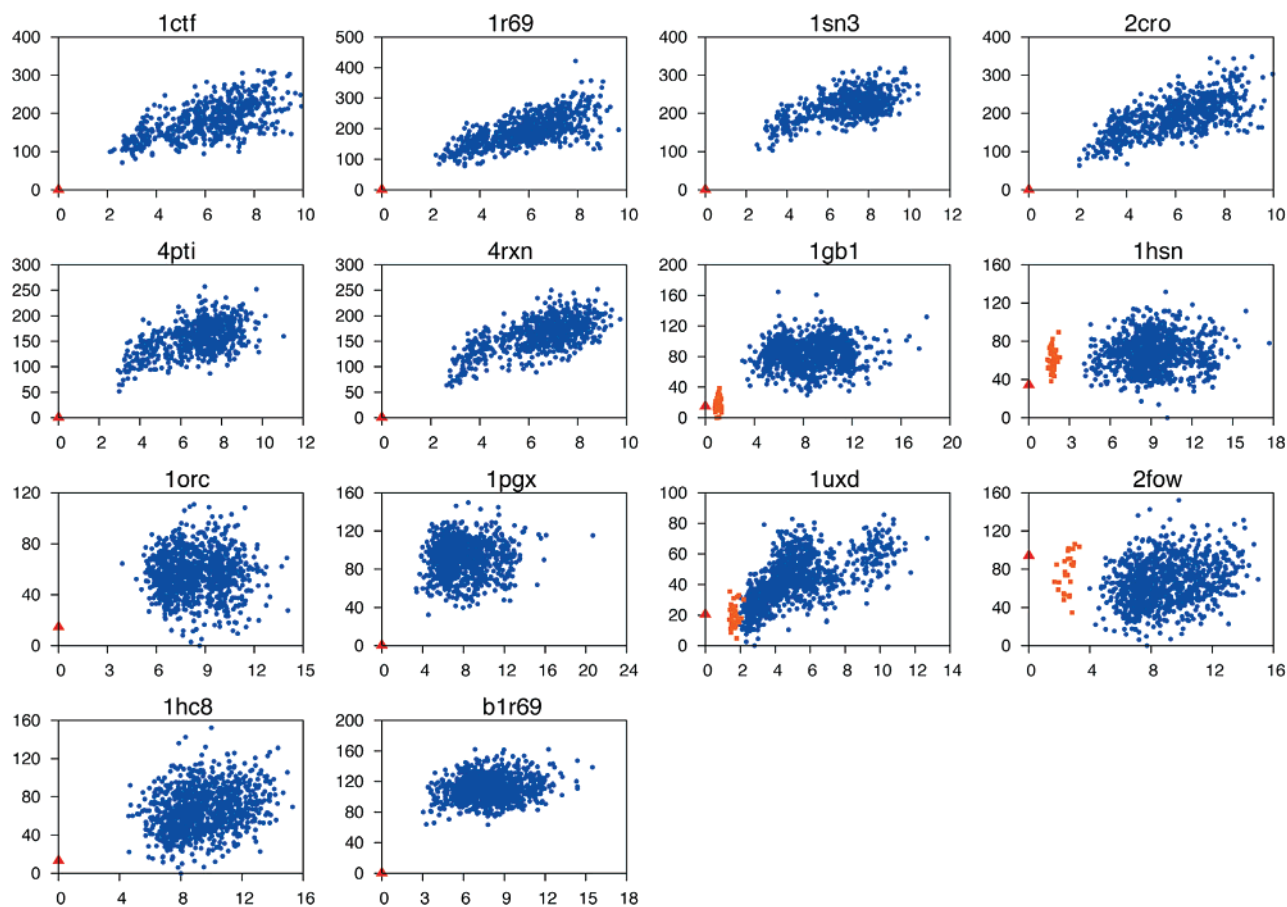


Figure 1. Energy vs rmsd plots for systems scored using the AMBER potential: (red triangle) native structure or NMR minimized mean where applicable, (orange square) individual NMR model, and (blue circle) decoy structure (Y-axis $\Delta\Delta G$ [kcal/mol] and X-axis rmsd [Å]). Energies are reported as the difference in energy for a state compared to the lowest energy structure in the decoy set.

structure, the energy gap was not as favorable as it was for the 4-state-reduced decoys. The rmsd values of the lowest scoring decoy are also generally far from the native structure.

In comparing the 4-state-reduced set to the Rosetta set, there are several possible explanations for AMBER's inability to perform well on the Rosetta decoys. All of the native structures in the 4-state-reduced set were solved using X-ray crystallography, while those in the Rosetta set are a combination of X-ray and NMR structures. Indeed, of the four cases where AMBER is unable to identify the native structure, three were NMR structures. This may suggest that AMBER structures generally score better for X-ray structures. Indeed the X-ray structure of the RNA binding domain of ribosomal protein LII (1hc8) scores significantly better than any NMR model for this system (2fow), as shown in Figure 1. It has been observed that structures are generally more stable in molecular dynamics simulations when started from an X-ray structure in comparison to an NMR structure.³⁰ It has also been demonstrated that NMR structures are more difficult to identify from among a set of decoys, compared to X-ray structures.²⁹ Rosetta decoys may also possess more nativelike structural characteristics since they were generated using fragments from the PDB.

An energy decomposition has been performed for decoys scored using the AMBER potential. As shown in Table 3, the internal geometries of the 4-state-reduced set generally

favor the native state, as opposed to the Rosetta set. The energy gap for the internal geometric energy is fairly large for the Rosetta set, although most of the difference arises from the torsional energy term, with only a small fraction resulting from the bond and angle terms. Since decoys are generally less well packed than native structures, their side chains may be able to adopt more favorable torsional orientations, whereas side chains in native structures may remain strained to improve overall packing. The Lennard-Jones energy provides a reliable indicator of the native structure for the 4-state-reduced set, though less so for the Rosetta set. For the cases where the van der Waals energy was unable to identify the native structure, the energy gap to the overall-best scoring decoys was generally small. The electrostatic energy reported is a combination of the Coulombic interaction and the solvation energy. The two are generally antagonistic²⁸ and so have not been separated. AMBER scores well for the native structure with respect to the electrostatic energy for the 4-state-reduced set, while again performing poorly for the Rosetta set.

Scoring Minimized Structures with DivScore. As with AMBER, performing an all-atom minimization prior to scoring structures accentuates DivScore's ability to discriminate native structures from decoys. For this study, the AMBER optimized geometries are used, as it is too computationally intensive to minimize all structures at a quantum

Table 3. Energy Decomposition for the AMBER Force Field Applied to All-Atom Minimized Structures

decoy set	system	rank _{tot}	rank _{int} ^a	$\Delta\Delta G_{\text{int}}$	rank _{vdw} ^b	$\Delta\Delta G_{\text{vdw}}$	rank _{ele} ^c	$\Delta\Delta G_{\text{ele}}$
4-state-reduced	1ctf	1	4	2.47	1	-6.07	1	-24.29
	1r69	1	15	7.10	1	-35.23	1	-4.00
	1sn3	1	1	-0.53	1	-8.58	1	-26.31
	2cro	1	15	14.47	1	-24.43	2	2.79
	4pti	1	1	-7.71	2	0.29	1	-0.45
	4rxn	1	1	-11.11	4	19.21	1	-13.49
Rosetta	1gb1	1	559	23.13	1	-1.46	14	8.24
	1hsn	36	963	21.05	137	17.52	12	0.37
	1orc	11	808	27.69	162	25.75	11	14.17
	1pgx	1	865	33.74	1	-21.27	3	7.66
	1uxd	4	959	15.48	1	-19.28	37	8.56
	2fow	80	967	34.72	180	35.72	8	10.39
	1hc8	7	891	30.31	2	0.27	199	40.21
	1r69	1	977	40.30	1	-41.97	2	12.93

^a Internal geometric energy (sum of bond, angle, and torsional energy terms). ^b van der Waals energy. ^c Electrostatic energy.

Table 4. Decoy Rankings for All-Atom Minimized Structures, Scored with DivScore Using the PM3 Hamiltonian

decoy set	system	rank _{nat}	Z-score _{nat}	$\Delta\Delta G_{\text{eff}}$ (kcal/mol)	rmsd _{best} ^a (Å)
4-state-reduced	1ctf	1	-4.56	-119.08	3.15
	1r69	1	-5.53	-153.98	3.37
	1sn3	1	-6.62	-145.77	4.11
	2cro	1	-5.49	-132.71	2.64
	4pti	1	-5.39	-80.18	3.02
	4rxn	1	-4.03	-55.05	3.21
Rosetta	1gb1	1	-3.03	-4.22	5.71
	1hsn	15	-2.16	22.95	7.74
	1orc	1	-4.16	-40.80	6.67
	1pgx	1	-4.64	-25.61	6.94
	1uxd	45	-1.71	24.58	2.41
	2fow	5	-2.26	26.71	6.42
	1hc8	1	-3.76	-25.77	6.37
	1r69	1	-7.04	-112.26	5.94

^a rmsd of the best scoring decoy.

mechanical level. Previous studies have shown that AMBER-minimized protein structures should be sufficient for scoring with SE-QM methods.³¹

These results for scoring with DivScore are summarized in Table 4 for the PM3 Hamiltonian. DivScore is also able to correctly identify the native structure for all of the 4-state-reduced decoys using both PM3 (Table 4) and AM1 (data not shown). Overall, PM3 shows a slight improvement over AM1 in its ability to discriminate decoys, not only in ranking but also in Z-score and energy gap. The rmsd of the best scoring decoys as calculated with PM3 are closer to the native in general than those scored with AM1. Overall, the results indicate that PM3 performs slightly better than AMBER in regards to scoring protein structures.

Figure 2 illustrates the correlation of heats of formation as calculated by PM3 and AM1 for the 1uxd decoy set. As expected, there is a tight correlation between the two and this trend is seen for all decoy sets. Since the two Hamiltonians behave so similarly, only PM3 will be discussed below. Interestingly, there is also a very good correlation between the PM3 heats of formation and the AMBER energy,

indicating that semiempirical methods may be reliably used in many of the situations where classical potentials have been successfully applied.

An energy decomposition has also been performed for the DivScore energies using the PM3 Hamiltonian (Table 5). The heat of formation of the native structure usually scores very well compared to the collection of misfolded structures. All native models, however, score very poorly with respect to the solvation energy in comparison to decoy models. While this energy gap may seem very large, the gap is reported for the native structure in relation to the decoy with the lowest free energy of solvation (not the overall best scoring decoy). Indeed, the best scoring decoys also have very unfavorable free energies of solvation, like the native state, which is usually balanced by the favorable heats of formation and dispersive interactions. All native structures possess very favorable LJ6 terms suggesting tighter packing in the native structures compared to the decoys.

In the previous discussion, the DivScore was calculated by a simple addition of the heat of formation, the solvation energy, and an attractive term to account for dispersive effects. The attractive term is taken from the AMBER force field, while the heat of formation was calculated with the PM3 Hamiltonian. Since these terms were taken from different theoretical treatments, weighted coefficients for each term in the DivScore equation have been assigned.

The coefficients were parametrized to maximize the Z-score of the native structure in relation to all decoys (or in the case of NMR structures, the lowest scoring model). The Z-score was chosen because it best represents the improvement of the native structure over the entire ensemble of decoys as opposed to $\Delta\Delta G_{\text{eff}}$, which only measures the energy gap between the native structure and the best scoring decoy. From eq 3, it can be seen that because the Z-score involves the standard deviation of the data set, it is not a linear function of the individual energy terms thus preventing a linear fitting of coefficients. A Monte Carlo method was therefore used to search the parameter space.

Of the 13 decoy sets in our study, six were chosen at random for the training set of the parametrization. A Monte Carlo method with a metropolis accepting scheme and simulated annealing was applied to optimize the parameters

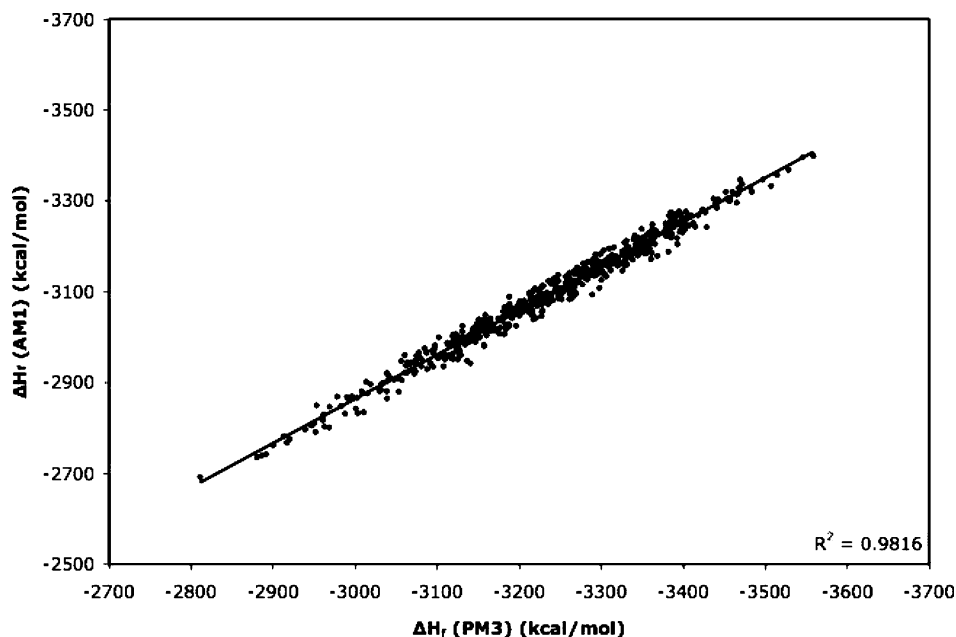


Figure 2. Correlation of heats of formation as calculated for the 1uxd decoy set using the PM3 and AM1 Hamiltonians with DivCon.

Table 5. Energy Decomposition for All-Atom Minimized Decoy Sets as Calculated with DivScore Using the PM3 Hamiltonian

decoy type	system	rank _{tot}	rank _{Hf} ^a	gap _{Hf} ^d	rank _{solv} ^b	gap _{solv} ^d	rank _{LJ6} ^c	gap _{LJ6} ^d
4-state-reduced	1ctf	1	4	44.62	471	570.51	1	-22.04
	1r69	1	2	19.53	594	559.40	1	-58.94
	1sn3	1	1	-63.47	569	613.26	5	22.06
	2cro	1	16	72.01	524	515.20	1	-59.28
	4pti	1	16	95.99	440	455.73	25	58.75
	4rxn	1	1	-32.07	646	925.86	10	34.17
Rosetta	1gb1	1	3	4.07	976	677.13	1	-13.66
	1hsn	15	154	152.73	857	447.77	1	-1.42
	1orc	1	64	183.35	812	570.97	7	6.07
	1pgx	1	1	-10.12	978	627.83	1	-60.63
	1uxd	45	21	118.11	972	343.64	1	-23.04
	2fow	5	56	173.82	894	618.72	4	16.15
	1hc8	1	24	124.08	920	644.35	1	-12.19
	1r69	1	1	-10.07	977	828.78	1	-141.21

^a Heat of formation. ^b Solvation energy. ^c Dispersive term of the classic LJ6-12 potential. ^d Energy gaps are reported as the energetic difference between the native structure and the decoy with the lowest value for the given energetic term.

that would improve the Z-score of the native structure in the training sets. The training set was chosen at random, and the procedure was repeated several times to ensure that the optimal parameter set was reproduced. Equation 4 shows the weighted DivScore with appropriate coefficients for the PM3 calculations. The decoy sets that were used in the training set were 1ctf, 1gb1, 1hsn, 1pgx, and both 1r69 sets.

$$E_{\text{tot}} = 0.250 * \Delta H_f + 0.225 * \Delta G_{\text{solv}} + 0.525 * \text{LJ}_6 \quad (4)$$

Using the weighted individual energy terms, the decoy sets were rescored and reranked. Table 6 shows the results of using weighted energy terms with the PM3 Hamiltonian. As illustrated, there is a marked improvement of the rankings for 1hsn, 1uxd, and 2fow. As expected, there is a general improvement in Z-scores for native structures, consistent with the weighting coefficients designed to improve the discrimi-

natory ability of the function. The energy distributions for each decoy set scored with the weighted DivScore are shown in Figure 3.

For comparison purposes, the energetic terms in the AMBER potential were also reweighted using the same procedure. The reweighted terms are shown in eq 5. The weights for the geometric energy terms (bond length, angle, and torsion) were zero and so are neglected in the equation.

$$E_{\text{tot}} = 0.378 * E_{\text{vdw}} + 0.310 * E_{\text{ele}} + 0.312 * E_{\text{solv}} \quad (5)$$

The results of scoring with the reweighted AMBER potential are shown in Table 7, and it performs well at identifying native structures from decoy models. In general, however, the Z-scores and free energy gaps are not as favorable as those for the weighted DivScore.

Table 6. Decoy Rankings for All-Atom Minimized Structures as Calculated with Weighted DivScore Using the PM3 Hamiltonian

decoy set	system	rank _{nat}	Z-score _{nat}	$\Delta\Delta G_{\text{eff}}$ (kcal/mol)	rmsd _{best} (Å)
4-state-reduced	1ctf	1	-5.43	-55.41	3.15
	1r69	1	-6.36	-71.66	4.02
	1sn3	1	-5.70	-47.15	7.81
	2cro	1	-5.73	-58.67	2.07
	4pti	1	-4.35	-4.35	7.46
	4rxn	1	-4.39	-22.55	2.86
Rosetta	1gb1	1	-4.37	-16.77	3.66
	1hsn	1	-3.05	-2.23	9.57
	1orc	1	-3.48	-13.22	10.21
	1pgx	1	-5.26	-31.65	7.22
	1uxd	1	-3.19	-5.15	2.73
	2fow	2	-2.96	3.01	6.37
	1hc8	1	-4.30	-18.54	6.37
1r69	1	-8.71	-76.19	6.69	

The scoring results for the ribosomal protein LII decoy set are plotted in Figure 3 using the weighted DivScore with the PM3 Hamiltonian. In addition to NMR models (2fow), there is also an X-ray structure available, 1hc8. The X-ray structure obtains a ranking of 1, correctly identifying it as the native structure. None of the NMR models obtain a native ranking, although some score more favorably than others. It should be noted that, overall, there is only a small difference in energy between NMR structures. Structurally, there does not appear to be any clear defining feature distinguishing the X-ray structure from the best and worst NMR models indicating that the differences between the two may be subtle. Structural verification checks show that the X-ray structure has much better packing than the NMR models, and this trend is seen by the improved Lennard-Jones interactions of the X-ray model.

The results of scoring with DivScore show that this scoring function is particularly well suited for identifying the native structure from among all decoys. All native structures can be correctly identified, although in the case of ribosomal protein LII it is the X-ray structure (1hc8) that is identified rather than the NMR models (2fow). The Z-scores for all native structures are large, indicating that the potential function scores the native structure much better than the set of decoys. The energy gaps between the native structure and the best-scoring decoy are large for the 4-state-reduced set, although noticeably smaller for the Rosetta decoys.

Scoring Near-Native Decoys. The 4-state-reduced and Rosetta decoy sets studied here lack near-native decoys with an rmsd less than 2.0 Å. As the field of structure prediction advances, it is important to clearly identify near-native decoys from those with higher RMSDs. A set of near-native, low rmsd decoys was obtained (Bradley, P., personal communication) that was used in ab initio folding studies⁵ and is listed in Table 8. These decoys sets provide a more stringent test of our scoring methods, as some models are as low as 0.55 Å from the native. The results of scoring this set with DivScore are shown in Figure 4. Encouragingly, DivScore was able to discriminate native from near-native

folds for all four decoy sets. In the case of 1r69 and 1di2, favorable energy funnels are observed, whereas 1af7 and 2reb demonstrate limited ability to discriminate near-native decoys from those that are more structurally divergent.

The results of these studies demonstrate the utility of semiempirical methods for studying protein structures, matching, or exceeding the AMBER force field at discriminating native structures. Given the fact that classical potentials have been parametrized for biological molecules, it is surprising that semiempirical methods performed so well—even given the fact that they are known to give poor conformational profiles. Moreover, semiempirical methods were parametrized at the element level, and some of the functional groups found in proteins were not included in the parametrization set.²⁶

This study hints at the possibility of using quantum mechanical methods in large-scale folding studies, although to reduce the time taken they should be coupled with lower-resolution scoring models to remove obviously poor structures. As expected, correctly scoring protein structures is dependent upon first minimizing the structure, preferably with the same potential being used to score the model. Performing an all-atom minimization with the AMBER force field cleaned up all structures and ensured that structures could be compared without bias to their starting structures. While it would have been an interesting study to minimize all decoys at the semiempirical level before scoring with DivCon, such calculations are too costly at present.

It is worth considering why semiempirical models score protein decoys as well as we have found in the present study. Semiempirical methods are known not to give phi-psi plots that agree with high quality ab initio results,³² while force fields are generally parametrized to reproduce these plots at some level of accuracy. This suggests that other factors play a role like long-range electrostatics or cooperativity effects observed in the folding of secondary structural elements.³³ Possibly these effects are overwhelming the conformational effects when using semiempirical methods in scoring native and decoy protein structures.

Conclusion

Here we have validated the capability of using semiempirical quantum mechanics in detecting misfolded proteins relative to the natively folded target protein. The ability of semiempirical methods to detect the native structure from a collection of decoys is quite remarkable and hints that the use of ab initio or density functional methods would also have significant potential in this regard. While the present test was for decoy detection one can envision using semiempirical approaches to facilitate the refinement of homology models as well as having an impact on the protein folding problem. Further studies along these lines are underway.

Evaluating native structures from a collection of misfolded states provides a challenging test for scoring functions and provides insights into the manner in which they fail. Energy decompositions are particularly useful because they highlight the terms in a scoring function that may require reparameterization or further study. AMBER energy decompositions indicate that for most minimized structures, there will be

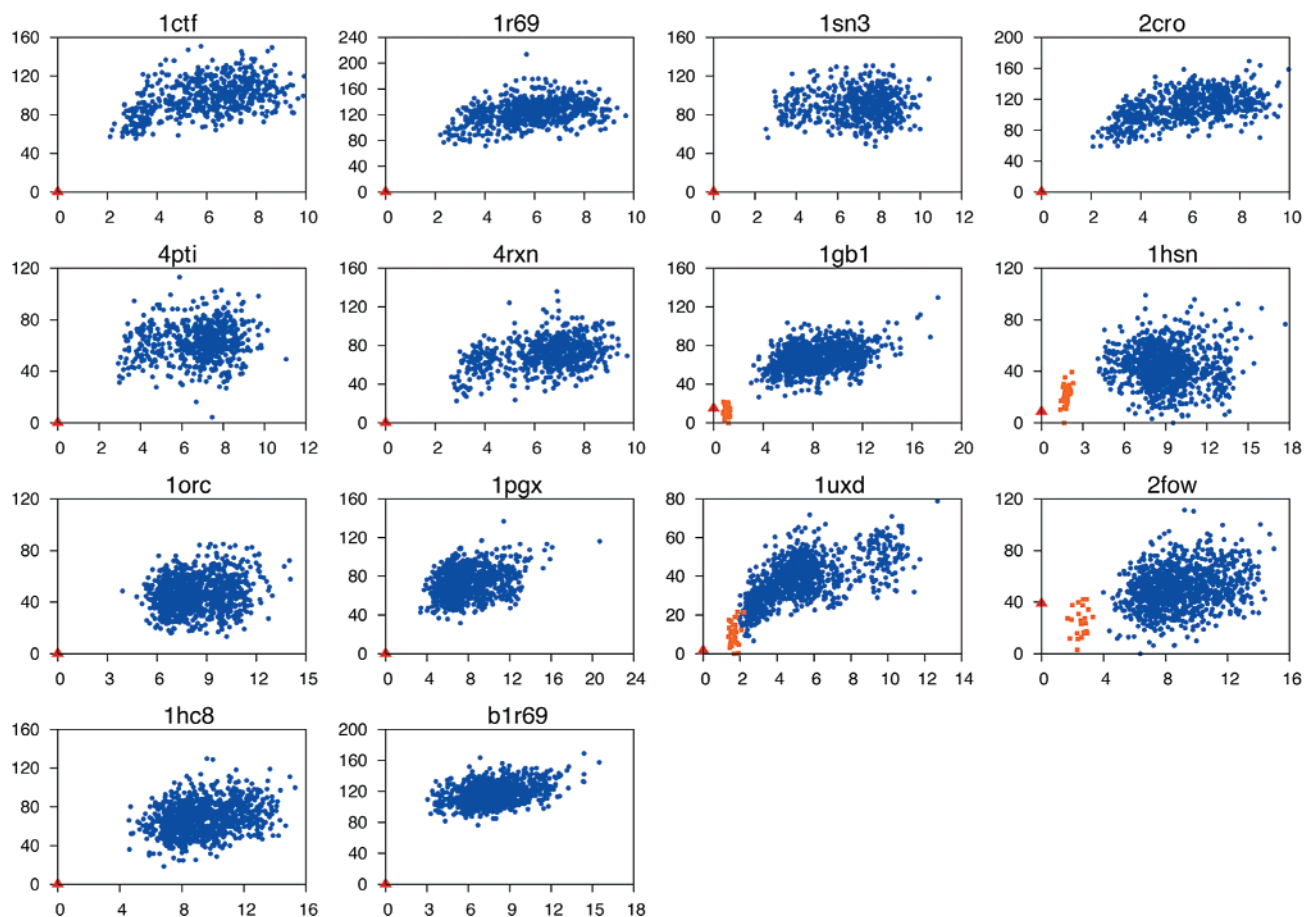


Figure 3. Energy vs rmsd plots for systems scored using the DivScore potential: (red triangle) native structure or NMR minimized mean where applicable, (orange square) individual NMR model, and (blue circle) decoy structure (Y-axis $\Delta\Delta G$ [kcal/mol] and X-axis rmsd [Å]). Energies are reported as the difference in energy for a state compared to the lowest energy structure in the decoy set.

Table 7. Decoy Rankings for All-Atom Minimized Structures as Calculated with the Reweighted AMBER Potential

decoy set	system	rank _{nat}	Z-score _{nat}	$\Delta\Delta G_{\text{eff}}$ (kcal/mol)	rmsd _{best} (Å)
4-state-reduced	1ctf	1	-4.85	-23.94	2.93
	1r69	1	-3.91	-26.81	2.65
	1sn3	1	-6.41	-26.72	2.87
	2cro	1	-4.97	-23.94	2.46
	4pti	1	-4.45	-9.04	3.67
Rosetta	4rxn	1	-4.52	-9.46	2.86
	1gb1	1	-4.06	-5.42	4.15
	1hsn	3	-3.05	3.48	9.52
	1orc	1	-2.71	-0.14	10.90
	1pgx	1	-5.37	-15.74	4.99
	1uxd	1	-3.03	-1.74	3.09
	2fow	1	-3.01	-0.10	8.04
	1r69	1	-7.95	-30.49	4.20

little energetic difference resulting from internal geometric differences. Rather, van der Waals and electrostatic contributions to the energetic state appear to be most important in specifying native folds. However, the solvation energy of the native structures usually scores poorly with respect to the decoy ensemble. While this might initially signify that the solvation energetic terms may need to be revisited, it

actually highlights an important point in scoring proteins; native structures are marked by a competition of various energetic terms. No single term dominates in the energetic treatment of a protein³ as the native state rarely has the tightest packing, the most favorable charge interactions, or the most favorable solvation energy. Instead, the native structure seems tuned to be the best combination of various energetic terms so that it can perform its function.

Methods

In this study, we investigated the ability of semiempirical methods to identify native structures from decoys in subsets of both the 4-state-reduced and Rosetta decoy sets, listed in Table 1. The Rosetta decoy set is composed of 92 different decoy sets, with systems ranging from 17 to 146 residues in length. The decoy sets also differed in their distribution of rmsd values from the native structure with some sets having all decoys over 8 Å away from the native fold. For this study, only small proteins (>45 residues long) were chosen whose decoys were distributed over a large range of conformational space and had structures with an rmsd < 5.0 Å. For sets whose native structure was solved by NMR, the minimized mean structure was taken to be the “native” structure for calculating the rmsd of a given decoy to the native conformation. However, when scoring NMR structures, each structure in the reported ensemble was treated, and the best

Table 8. Near-Native Rosetta Decoy Sets

PDB	description	type	N_{decoys}	N_{res}	N_{atoms}	rmsd range (Å)	%H/%E ^a
1af7	N-terminal domain of Chemotaxis receptor methyltransferase CheR	X-ray	999	72	1217	1.6–12.1	68/0
1di2	RNA binding protein A	X-ray	999	69	1096	0.8–9.4	42/33
1r69	N-terminal domain of phage 434 repressor	X-ray	999	61	976	1.0–7.5	70/0
2reb	RecA	X-ray	999	60	884	0.5–3.9	20/57

^a Percentage of helices and sheets in the native.

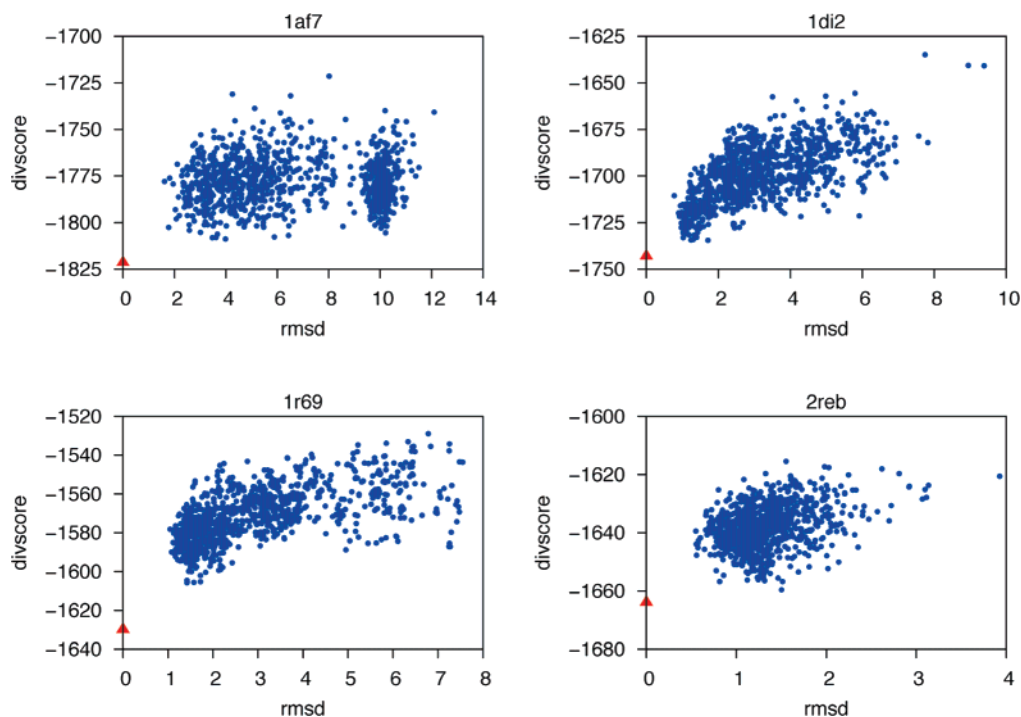


Figure 4. DivScore vs rmsd plots for systems with near-native decoys: (red triangle) native X-ray structure and (blue circle) decoy structure.

scoring conformation was reported. The RNA binding domain of ribosomal protein LII (2fow) has also been solved by X-ray crystallography, and so this structure (1hc8) has been treated as another native reference for this system.

Using the available decoy sets, several analyses have been performed: (1) using protein decoy heavy atom coordinates as-is (referred to as hydrogen-minimized since only hydrogen atoms were optimized) and (2) minimizing all structures with the AMBER force field prior to analysis (referred to as all-atom minimized structures). In all cases, hydrogen atoms were removed from all structures and then added with the LEaP module of AMBER ensuring that all hydrogen atoms were modeled in a consistent manner. All histidine residues were treated as a singly protonated species with the N^δ atom being protonated. Several decoy systems contained multiple cysteine residues. These were treated as disulfide bonds when the geometry was perceived to be favorable to disulfide bond formation. Upon hydrogen addition, hydrogen atoms were minimized using the Sander package in AMBER using the PARM94 force field for 300 steepest descent steps followed by 700 conjugate gradient steps. The resulting structures were then scored using the AMBER MM potential as well as with semiempirical QM methods found in DivCon, at both the AM1²⁵ and PM3³⁴ levels. On current hardware (2.4 GHz

AMD opteron) a DivCon single point energy evaluation takes on average 15 min.

Beyond scoring decoys “as-is”, an all-atom minimization was performed on all decoys. First, a restrained all-atom minimization was performed for 1500 steps with a force constant of 2.0 kcal/mol Å² applied, followed by an unrestrained minimization for 5000 steps with implicit solvation. The initial restrained minimization was performed to limit instabilities during the minimization caused by nonoptimal starting structures. Minimization with the AMBER/GBSA module resulted in structures that did not deviate significantly from the starting structure, with an average heavy-atom rmsd of 1.1 Å.

Abbreviations: molecular mechanics (MM); quantum mechanics (QM); root mean squared deviation (rmsd).

Acknowledgment. We thank Kenneth Ayers for his assistance with managing the computational resources and for valuable discussions. We would also like to thank Philip Bradley for supplying us with near-native decoy sets. We thank the NSF (MCB-0211639) and the NIH (GM44974 and GM066859) for financial support of this research and the National Center for Supercomputer Applications (NCSA) and the Pittsburgh Supercomputer Center for generous allocations of supercomputer time.

Supporting Information Available: Rankings of preminimized decoy sets scored with AMBER and DivScore are available. Scoring decoys as-is accentuated differences between the scoring force field and the method used to generate the structures. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **1973**, *181* (96), 223–30.
- (2) Schonbrun, J.; Wedemeyer, W. J.; Baker, D. Protein structure prediction in 2002. *Curr. Opin. Struct. Biol.* **2002**, *12* (3), 348–54.
- (3) Lazaridis, T.; Karplus, M. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **2000**, *10* (2), 139–45.
- (4) Moult, J. A. decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* **2005**, *15* (3), 285–9.
- (5) Bradley, P.; Misura, K. M.; Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science* **2005**, *309* (5742), 1868–71.
- (6) Zhang, Y.; Arakaki, A. K.; Skolnick, J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins* **2005**, *61 Suppl 7*, 91–8.
- (7) Park, B.; Levitt, M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* **1996**, *258* (2), 367–92.
- (8) Park, B. H.; Huang, E. S.; Levitt, M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* **1997**, *266* (4), 831–46.
- (9) Hendlich, M.; Lackner, P.; Weitckus, S.; Floeckner, H.; Froschauer, R.; Gottsbacher, K.; Casari, G.; Sippl, M. J.; Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **1990**, *216* (1), 167–80.
- (10) Felts, A. K.; Gallicchio, E.; Wallqvist, A.; Levy, R. M. Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the Surface Generalized Born solvent model. *Proteins* **2002**, *48* (2), 404–22.
- (11) Cornell, W. D.; Cieplak, P.; Baylay, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field For the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (12) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy Minimization and Dynamical Calculations. *J. Comput. Chem.* **1983**, *4*, 182–217.
- (13) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (14) Wang, B.; Liao, K. R. N.; Peters, M. B.; Kim, H.; Westerhoff, L. M.; Wollacott, A. M.; van der Vaart, A.; Gongonea, V.; Suarez, D.; Dixon, S. L.; Vincent, J. J.; Brothers, E. N.; Merz, K. M., Jr. *DivCon* 2005.
- (15) Dixon, S. L.; Merz, K. M. Semiempirical molecular orbital calculations with linear system size scaling. *J. Chem. Phys.* **1996**, *104* (17), 6643–6649.
- (16) Yang, W. T.; Lee, T. S. A Density-Matrix Divide-and-Conquer Approach for Electronic-Structure Calculations of Large Molecules. *J. Chem. Phys.* **1995**, *103* (13), 5674–5678.
- (17) Dixon, S. L.; Merz, K. M. Fast, accurate semiempirical molecular orbital calculations for macromolecules. *J. Chem. Phys.* **1997**, *107* (3), 879–893.
- (18) van der Vaart, A.; Merz, K. M., Jr. The Role of Polarization and Charge Transfer in the Solvation of Biomolecules. *J. Am. Chem. Soc.* **1999**, *121*, 9182–9190.
- (19) van der Vaart, A.; Suarez, D.; Merz, K. M. Critical assessment of the performance of the semiempirical divide and conquer method for single point calculations and geometry optimizations of large chemical systems. *J. Chem. Phys.* **2000**, *113* (23), 10512–10523.
- (20) Raha, K.; Merz, K. M., Jr. A Quantum Mechanics Based Scoring Function: Study of Zinc-ion Mediated Ligand Binding. *J. Am. Chem. Soc.* **2004**, *126*, 1020–1021.
- (21) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **1997**, *268* (1), 209–25.
- (22) McConkey, B. J.; Sobolev, V.; Edelman, M. Discrimination of native protein structures using atom-atom contact scoring. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (6), 3215–20.
- (23) Tsai, J.; Bonneau, R.; Morozov, A. V.; Kuhlman, B.; Rohl, C. A.; Baker, D. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins: Struct., Funct., Genet.* **2003**, *53* (1), 76–87.
- (24) Gogonea, V.; Merz, K. M., Jr. Fully Quantum Mechanical Description of Proteins in Solution. Combining Linear Scaling Quantum Mechanical Methodologies with the Poisson-Boltzmann Equation. *J. Phys. Chem. A* **1999**, *103*, 5171–5188.
- (25) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (26) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods. 2. Applications. *J. Comput. Chem.* **1989**, *10* (2), 221–264.
- (27) Lazaridis, T.; Karplus, M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* **1999**, *288* (3), 477–87.
- (28) Dominy, B. N.; Brooks, C. L. Identifying native-like protein structures using physics-based potentials. *J. Comput. Chem.* **2002**, *23* (1), 147–60.
- (29) Lee, M. R.; Tsai, J.; Baker, D.; Kollman, P. A. Molecular dynamics in the endgame of protein structure prediction. *J. Mol. Biol.* **2001**, *313* (2), 417–30.
- (30) Lee, M. R.; Kollman, P. A. Free-energy calculations highlight differences in accuracy between X-ray and NMR structures and add value to protein structure prediction. *Structure (London)* **2001**, *9* (10), 905–16.

- (31) Wollacott, A. M.; Merz, K. M. Development of a Parametrized Force Field To Reproduce Semiempirical Geometries. *J. Chem. Theory Comput.* **2006**, 2 (4), 1070–1077.
- (32) Mohle, K.; Hofmann, H. J.; Thiel, W. Description of peptide and protein secondary structures employing semiempirical methods. *J. Comput. Chem.* **2001**, 22 (5), 509–520.
- (33) Morozov, a. V.; Tsemekhman, K.; Baker, D. Electron density redistribution accounts for half the cooperativity of alpha helix formation. *J. Phys. Chem. B* **2006**, 110 (10), 4503–4505.
- (34) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods. 1. Method. *J. Comput. Chem.* **1989**, 10 (2), 209–220.

CT600325Q

Sugar Folding: A Novel Structural Prediction Tool for Oligosaccharides and Polysaccharides 1

Junchao Xia,^{†,§} Ryan P. Daly,^{†,§} Feng-Chuan Chuang,[†] Laura Parker,[†]
Jan H. Jensen,[‡] and Claudio J. Margulis^{*,†}

*Department of Chemistry, University of Iowa, Iowa City, Iowa 52242, and
Department of Chemistry, University of Copenhagen, Universitetsparken 5,
2100 Copenhagen, Denmark*

Received February 8, 2007

Abstract: This paper is the first in a series of two articles where we report the development of fast sugar structure prediction software (FSPS). To the best of our knowledge, this is the first automated tool for the systematic study of conformations of complex oligosaccharides in solution. In contrast to previously developed molecular builders such as POLYS (Engelsen, S. B.; Cros, S.; Mackie, W.; Perez, S. *Biopolymers* **1996**, 39, 417–433) where only information about the minimum energy conformation of disaccharide pairs is considered in order to build larger oligosaccharides, this tool is based on a systematic search of dihedral conformational space, optimization of structures using implicit solvation models, explicit molecular dynamics simulations, NOE calculations, and a very powerful substructure recognition algorithm and database. Our FSPS can rapidly find minimum-energy conformers and rank them according to different criteria. Two such criteria are the energy of the conformers in implicit solvent and the root-mean-square deviation (RMSD) of computed NOEs with respect to experimental data. Even though experimental NOEs may result from an average over conformers instead of a single structure, we find that sorting according to NOE RMSD constitutes a better estimator for the global free-energy minimum structure in explicit solvent (i.e., the most likely structure in solution). In contrast, the lowest-energy structure in implicit solvent does not usually correspond to the free-energy minimum. A harmonic approximation to compute free energies of each conformer does not appear to reverse this conclusion, indicating that either explicit hydrogen bonding to the solvent or anharmonic effects in the free energy or both are fundamentally important. In the first article, we discuss our methodology and study, as a proof of concept, a simple substituted disaccharide. In the second article, we focus on two complex human milk oligosaccharides.

1. Introduction

Carbohydrates are powerful biological markers because they contain multiple asymmetric carbon centers and possess unique structures and chemical properties. Complex carbohydrates are involved in numerous molecular recognition phenomena because of their exquisite specificity in interact-

ing with proteins and other recognition agents. Glycoconjugates (glycoproteins and glycolipids) are actively involved in biological functions like tumor immunology,² cell growth and differentiation,³ signal transduction,⁴ apoptosis,⁵ spermatogenesis,⁶ and T-cell activation.⁷ Oligosaccharides are recognized by different enzymes and by a family of proteins called lectins. Usually, both enzymes and lectins only recognize a particular fold of the sugar. This is why being able to predict the conformation of oligosaccharides in solution is of utmost importance. Unfortunately, only a small set of sugar-binding proteins have been cocrystallized

* Corresponding author e-mail: claudio-margulis@uiowa.edu.

[†] University of Iowa.

[‡] University of Copenhagen.

[§] Equal contributions.

with their corresponding oligosaccharides. It is therefore desirable to have a computational tool in place that will predict the conformational structures that sugars can adopt in solution. This software also constitutes a powerful aid in the interpretation of nuclear Overhauser effect (NOE) spectra.

Characterization and a priori prediction of conformations of biologically relevant carbohydrates in solution is difficult. Typical tools such as UV and IR spectroscopy are not suitable to study these molecules, and NMR spectra only give information about sets of statistically averaged conformations on a millisecond time scale. Depending on the free-energy difference between conformers in solution, the NMR will be compatible with either a single structure or an ensemble of flexible structures. Therefore, a theoretical prediction of the oligosaccharide conformation is usually necessary to understand the NMR data. In the past, we^{8,9} and several other groups (for example, see refs 10 and 11) have performed molecular dynamics (MD) simulations in order to predict and understand the conformation of carbohydrates in solution. However, this approach has significant drawbacks. The main problem with using molecular dynamics to study the configuration of complex carbohydrates is that only certain conformations of these molecules are visited during the duration of a typical MD run. The challenge is in some ways similar to that of protein folding. One does not expect to see a protein fold on a time scale accessible by computer simulations. This problem is in fact much more pronounced in the case of sugars because proteins are linear polymers while oligosaccharides are often branched and motion of the different branches is often strongly coupled. It is important to recognize that the problem is not simply related to the potential energy barrier between different sugar conformers. This energy is usually low and compatible with thermal fluctuations at room temperature; the problem for molecular dynamics simulations is related to entropy.^{12,13} This is particularly evident in the case of branched saccharides or saccharides with linkage points that are adjacent.

Although much has been learned from research in the protein field where extensive libraries^{14–28} of peptide rotamers are available, no such tools currently exist for oligosaccharide systems. In fact, the problem of designing a rotameric library is topologically much more complex in the case of sugars than in the case of peptides. Sugars have many different linkage points, and their allowed dihedral space not only depends on the linkages and identities of the two monosaccharide units but also on the possibility of branching. Furthermore, recognizing subtrees of connected rings within a larger tree is in itself a highly complex problem in graph theory. In this article, we describe how our newly developed tool overcomes several of these difficulties and delivers results that are very hard to obtain otherwise with current computational tools. This will become more evident in the second paper where we present our results for a pair of complex human milk sugar oligosaccharides. Several methods have been used for the structural predictions of oligosaccharides. Most commonly, molecular dynamics in explicit solvent is used in order to predict NMR or NOE data in solution. In our experience,^{8,9} only disaccharides readily visit all possible free-energy minima during typical molecular

dynamics runs at room temperature. Larger oligosaccharides are generally trapped in local basins for longer than tens of nanoseconds, the length of a typical MD run. This is particularly true in the case of branched sugars or sugars with adjacent linkage points.

In the past, our group^{8,9} has used the following reasonable scheme to study conformations of complex oligosaccharides in solution: First, a search through dihedral space of each independent isolated disaccharide pair is constructed. Long molecular dynamics runs for each component disaccharide are performed to obtain free-energy minima. Second, in order to build all possible oligosaccharides, a combinatorial approach is used in which all possible free-energy minima of the component disaccharides are combined to form all possible oligosaccharide structures. In most cases, many of the combinations are disallowed because of steric clashes or bad hydrogen-bond energetics, and only several combinations are obtained. In principle, this seems like a very large combinatorial problem, but in fact, for biologically relevant oligosaccharides, only tens or hundreds of structures need to be scrutinized. Unlike the case of proteins or polypeptides, sugar monomers are much bulkier, and therefore many conformations are disallowed, particularly when they are branched. The third and final step is to study the dynamics of the different oligosaccharide conformers in order to find which of these are stable in solution.

Even though the approach described above is reasonable, there are two main problems with it. First, it is very time-consuming. It requires long molecular dynamics for all component disaccharide pairs and further molecular dynamics of the resulting oligosaccharides. Second, this sampling method assumes that no other structure except those that correspond to free-energy minima of each disaccharide pair will be minima in the case of the oligosaccharide. This, although reasonable, may preclude the existence of other free-energy minima that appear due to stabilization through interaction between monomer units that are nonadjacent (i.e., stabilization due to secondary structure). This type of interaction could potentially be very common in the case of branched oligosaccharides particularly near crowded linkage points. In section 2, we describe a fast alternative method that overcomes these difficulties.

2. Simulation Methods

We have developed a completely automatic tool to study sugar molecules. A considerable part of a systematic search program involves the elucidation of the topology of polysaccharides by using ring perception techniques. Much work has gone into the development of algorithms for the determination of the smallest set of smallest rings and other ensembles of rings representative of a chemical structure.²⁹ However, the problem of ring recognition in carbohydrates is simpler in that compound rings are an exception to the norm in carbohydrate chemical structure. As a result, a more efficient ring perception algorithm can be used. The primary motivation in ring perception is to enumerate the dihedral degrees of freedom from glycosidic linkages in the molecule to allow a search of the conformational space. [Our ring perception algorithm is implemented through the use of graph

theoretical methods, and we have developed a C++ graph class to deal specifically with all aspects of saccharide ring topology recognition. Atomic coordinates are initially loaded from a generic XYZ file which contains no residue information. A graph object is initialized with vertices and edges which correspond to the atoms and bonds in an oligosaccharide. Subsequently, we derive the ring topology of a complex oligosaccharide by performing a series of depth-first and breadth-first searches of the graph structure. The linkages between different rings (monosaccharides) and the connectivity of side chains are derived using similar methods. Within the graph object, atoms (vertices) are stored with specific information which helps to expedite these processes.] The proverbial systematic search algorithm attempts to visit every possible conformation. Such an approach quickly becomes unfeasible even for the most efficient algorithms on the fastest machines. One way to dampen the effect of combinatorial explosion is to minimize the search space for each dihedral degree of freedom during iteration. Sugar residues, in particular those oligosaccharides relevant to biology, lend themselves very well to such a procedure because the allowed conformations of the ϕ and ψ dihedral angles of a glycosidic linkage are generally constrained to within approximately 30% or less of the total space. In a pentasaccharide with four glycosidic linkages, for example, the overall required search space is reduced to $0.30^4 = 0.0081 = 0.81\%$ of its unfiltered size. The reduction is substantially more important in the case of complex branched oligosaccharides for which our methodology is intended. In fact, the allowed number of conformers could be smaller for larger sugars than for smaller ones. This is the case for the oligosaccharides discussed in the second paper.

Our systematic approach can be described as a set of sequential steps:

1. The first step involves complex ring perception.²⁹ The input is an arbitrary “xyz” file. No atom typing or residue database is required.

2. A molecule is decomposed into its component oligosaccharides fragments. These fragments are checked against a database (which is currently being populated) using a sophisticated subtree matching algorithm as described in section 2.1. If a fragment of the molecule has already been studied, then no systematic search is carried out on it. This will save significant amounts of computational time in the future when the database has many entries. This obviously amounts to a sophisticated version of a rotameric library in which monomers are not simply linked sequentially, but the effect of branching and adjacency is considered through bonds as well as through space. Rotameric libraries for proteins usually only have information about pairs of residues; our approach will store information about larger sugar subfragments. This is feasible since the number of sugar monomers in a typical biologically relevant oligomer is much smaller than the number of amino acids in a protein and since monosaccharides are in general considerably bulkier than amino acids. This procedure is most useful in the case when branching is present since sterics will significantly restrict conformation space and, consequently, the number of configurations in space that we need to store.

The case of linear sugars with nonadjacent linkages is the least interesting to us since the number of configurations to store becomes exponentially large as the size of the oligosaccharide grows.

3. Items not previously studied or stored are separated into monosaccharide residues and side chains. We then perform a systematic grid search for the allowed ϕ - ψ pairs for each residue linkage and side-chain linkages. The angular increment can be arbitrarily chosen; we have used 10–20° for residue linkages and 60–120° for side-chain linkages. After a clash check using a hard sphere criterion, we obtain corresponding steric Ramachandran maps for all residue and side-chain linkages. Clash checks are only performed between atoms in different residues, not within the same ring.

4. The oligosaccharide is reconstructed by reassembling the linkages one by one at corresponding allowed conformations. At this point, depending on the size of the structure pool, coarse graining can be applied to constrain the number of candidate structures. For example, four neighboring points in Ramachandran space will become a new point which is calculated as the geometric center of the allowed points. As opposed to other clustering schemes, this coarse-graining method is unlikely to miss small isolated regions in configuration space.

5. After obtaining the sterically allowed conformations, we perform energy minimizations using an implicit solvent model. In this paper and in the second paper, we have used different software programs^{30–33} and force fields^{34,35} to achieve this.

6. We pool the minimized structures into unique conformational families. We consider that two conformations have a unique structure if the energy difference $\Delta E < 5.0$ kcal/mol and the difference in each of the dihedral angles is $< 10^\circ$. We keep the structure with lowest energy in each family, and we define this as a “unique” conformer.

7. Unique conformers can be sorted on the basis of different criteria. We have used an energy rank as well as a rank based on the root-mean-square deviation (RMSD) between experimental NOE values for proton pairs on different monosaccharide units and our computed values for the unique structures. Our approach for computing NOEs is the same as that used by Cumming and Carver,^{36,37} which is based on the model-free approach.^{38,39}

8. Finally, we run short 5 ns molecular dynamics simulations in explicit solvent in order to gauge the stability of each of the different unique conformers and in order to compute time-averaged NOEs. [A structure is deemed stable if after 5 ns of simulation glycosidic angles have not changed to a different local minimum. Clearly, these short simulations only indicate whether a structure is in a deep local minimum as compared to KT and not whether the structure is at a global free-energy minimum. Much more expensive procedures such as parallel tempering can be applied if accurate relative free energies between different conformers of complex oligosaccharides generated by our fast sugar structure prediction software (FSPS) are sought.]

2.1. Sub-Tree Recognition and Database. Because of the possibility of structural branching in sugars, the act of querying a database in search of a set of structures most

similar to the molecule of interest is highly complex. A practical carbohydrate database query protocol was presented by Aoki et al.^{40–43} While their method makes use of a scoring algorithm to sort matches, the method we use accomplishes the same task through the use of a slightly more generalized algorithm which effectively solves the maximum common subgraph problem for trees with labeled nodes and edges.

The first methods to solve the graph isomorphism problem were mostly set theoretic in nature.^{44,45} More recently, some researchers have focused on using the eigenvalues and eigenvectors associated with the adjacency matrix of a molecule as a means to distinguish it from others. However, this method has two limitations. First of all, it does not explicitly take vertex labels (i.e., atom types) into account. Second of all, it is only capable of verifying exact matches.

The methodology in our FSPS makes use of association graphs to solve this problem⁴⁶ by using an approach developed by Hopfield which is described in ref 47. In order to make a molecular-structure-based database query practical, in this method, a routine which allows the comparison of two structures to find any substructure in common is used. The structural information stored in our database is a residue graph, which is simply the graph generated by viewing each residue in an oligosaccharide as a vertex and each glycosidic linkage as an edge between vertices. Residue graphs are the objects which are compared when a query is made to the database. This type of comparison is made by solving the graph isomorphism problem, which seeks to find the maximum subgraph (i.e., the subgraph containing the largest number of vertices) present in both graphs. Because the vast majority of biological oligosaccharides contain no compound ring structure, the resulting residue graphs are tree graphs since they are presumed to contain no rings (or cycles). The algorithm used in this work finds the maximum subtree common to both oligosaccharides and is derived from a method by Jain and Wysotski.⁴⁷ This method is dependent upon the generation of an association graph, which is basically a map from one residue graph to another (see Jain and Wysotski⁴⁷). The generation of the association graph is a process in and of itself and can be optimized independently of the actual search. The main criterion in its optimization is to make it as small as possible and with as few edges as possible. Once the association graph is generated, a neural network algorithm is used to find the maximum clique⁴⁷ in the association graph. Maximum cliques correspond to subsets of vertices in the association graph which map residues from a new oligosaccharide to those of the structures already stored in our database. The resulting map points out common substructures between the molecule from the database and the molecule of interest. If a match is returned which is 100% of the size of a molecule in the database, then the sterically allowed conformational space stored with this entry in the database is used as an admissible search space for the mapped portion of the molecule of interest.

The association graph method has the advantage of being highly customizable. In addition to the ability to take into account atom types, other information such as atom chirality and bond type can be used to further eliminate possible matches. This procedure greatly shrinks the size of the

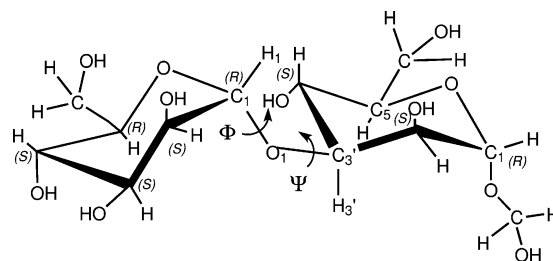


Figure 1. Schematic representation of the α -D-Man-(1 \rightarrow 3)- α -D-Man-O-Me disaccharide molecule. The two dihedral angles are defined as $\phi = \text{H}_1\text{-C}_1\text{-O}_1\text{-C}'_3$ and $\psi = \text{C}_1\text{-O}_1\text{-C}'_3\text{-H}_3'$.

association graph. New conformations are constructed on the basis of stored vectors containing coupled ϕ - ψ information for each linkage of the oligosaccharide in the database. The rest of the oligosaccharide is assembled by avoiding clashes with the database fragment. In the second paper,⁴⁸ we describe the use of this method for the analysis of conformations of complex milk sugars.

3. Results and Discussion: A Simple but very Important Test Problem

The simplest possible example that can be used in order to exemplify our procedure and to test the accuracy and validity of the different approximations involved is a substituted disaccharide. We have chosen α -D-Man-(1 \rightarrow 3)- α -D-Man-O-Me, with the schematic representation shown in Figure 1, because molecular dynamics simulations in explicit solvent can be converged to probe its full free-energy landscape on a several nanosecond time scale. Furthermore, this sugar has been well-studied by means of the nuclear Overhauser effect,⁴⁹ relaxed potential energy surfaces through an extensive molecular mechanics (MM) scheme,^{50,51} and also as a fragment of an oligosaccharide via molecular dynamics.⁵² The full free-energy landscape is not easily accessible for complex oligosaccharides like those studied in the second paper. In the case of the current paper, by having the full free-energy landscape of the molecule as a function of glycosidic dihedral angles, we are able to probe which sorting criteria is best (sequential step 7 in section 2) for our FSPS. Furthermore, because of the small system size, high-level ab initio calculations using an implicit solvent model can be carried out to thoroughly test the accuracy of molecular mechanics energetic predictions.

3.1. Implicit and Explicit Solvent, Force Fields, and ab Initio Calculations. What Matters and What Does Not for the Correct Prediction of Sugar Structures in Solution. *3.1.1. Using MM3 with TINKER.* Figure 2 shows the distribution of unique structures from our systematic search in ϕ - ψ glycosidic space using MM3³⁴ with TINKER.^{30,31} The two dihedral angles are defined as $\phi = \text{H}_1\text{-C}_1\text{-O}_1\text{-C}'_3$ and $\psi = \text{C}_1\text{-O}_1\text{-C}'_3\text{-H}_3'$ as shown in Figure 1. ϕ - ψ torsion angles have been adjusted in steps of 10° over the whole angular space. At each sterically allowed point, an energy minimization was performed using the generalized Born surface area (GBSA) implicit solvent model.^{53,54} Rotations were also performed for the hydroxymethyl group. Figure 2a displays the distribution of unique conformations

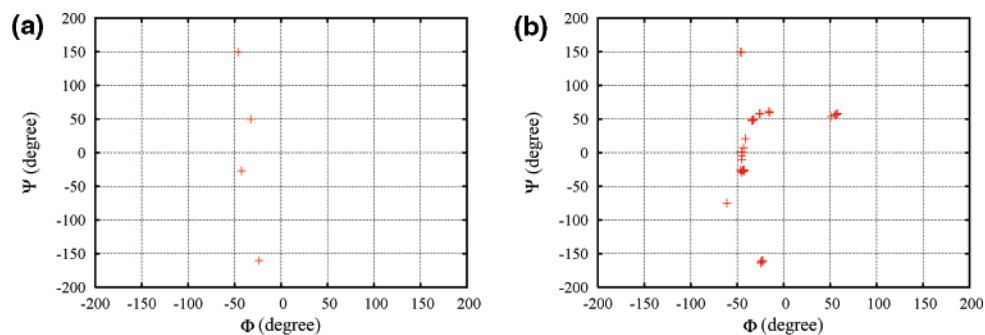


Figure 2. Distribution of unique conformations in ϕ - ψ glycosidic space in the case of (a) no side-chain rotations and (b) with 120° rotations of the first dihedral angle on any side chain with at least two rotatable dihedral angles. Side-chain rotation reveals more local minima. Energy minimizations were performed using TINKER^{30,31} with the MM3 force field³⁴ and GBSA implicit solvent model.^{53,54}

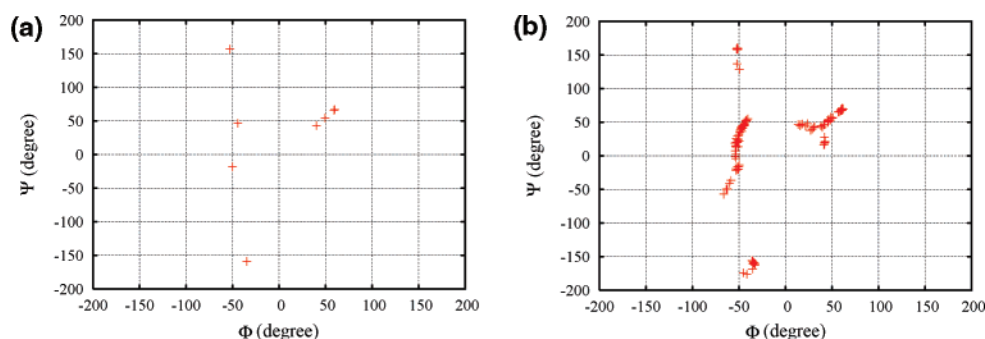


Figure 3. Same as Figure 2 except that the energy minimization is done using GROMACS with the OPLS-AA force field in the gas phase.

Table 1. The Potential Energy Difference (ΔE) and Free Energy Difference (ΔG) (kcal/mol) of Each Unique Conformations in Figure 2a Using Different MM and QM Procedures and Basis Sets^a

		conf. 1	conf. 2	conf. 3	conf. 4
TINKER(MM3) with GBSA	(ϕ, ψ)	(-32.5, 49.6)	(-42.7, -26.8)	(-23.6, -160.4)	(-46.4, 149.4)
	ΔE	0.0	1.44	1.85	4.79
GROMACS(OPLS-AA) gas phase	(ϕ, ψ)	(-44.3, 46.4)	(-50.1, -18.3)	(-34.6, -159.1)	(-52.8, 157.0)
	ΔE	0.0	2.81	3.47	4.51
GAMESS gas phase <i>B3LYP/6 - 31G(d,p)</i>	(ϕ, ψ)	(-36.5, 48.8)	(-50.8, -27.9)	(-28.9, -150.1)	(-42.7, 158.0)
	ΔE	0.0	1.99	0.82	3.62
(nvib = 2)	ΔG	0.0	2.71	0.85	4.63
	(ϕ, ψ)	(-36.1, 45.9)	(-50.7, -29.9)	(-28.2, -150.7)	(-43.5, 159.2)
GAMESS implicit solvent <i>B3LYP/6 - 31G(d,p)PCM</i>	ΔE	0.0	0.63	0.96	4.04
	(ϕ, ψ)	(-38.7, 49.2)	(-52.0, -14.0)		
MD GROMACS(OPLS-AA) with explicit solvent	ΔG	0	-3		

^a MM calculations with MM3 and OPLS-AA show that conformation 1 is the global energy minimum. QM calculations also find that conformation 1 is the global energy minimum. Note that QM calculations in the gas phase change the energy ordering of conformations 2 and 3. On the contrary, MD simulations in explicit solvent and experiments reveal that conformation 2 is in fact the global free-energy minimum.

obtained without any side-chain rotation. Only four minimized conformations are found. However, several other minima are shown in Figure 2b as we rotate the first dihedral angle of the hydroxymethyl group. The latter distribution is consistent with the adiabatically relaxed potential energy surface of Imberty et al. using the MM2 force field.⁵⁰ Including a full dihedral search for all hydroxyl groups instead of only the hydroxymethyl group produces more minima (results not shown); however, this is expensive and does not appear, at least in this particular case, to significantly modify the energy ordering of the conformers.⁵⁰

3.1.2. Using GROMACS with the OPLS-AA Force Field and *ab Initio* Calculations with GAMESS. At the time of

our simulations, GROMACS^{32,33} did not offer an implicit solvent option. The OPLS-AA potential³⁵ appears to show more local minima than MM3 as can be appreciated in Figure 3; however, these extra minima are at much higher energies. The four main unique conformations found from our MM3 TINKER calculations are similar to those predicted by OPLS-AA and appear to keep the same relative energy ordering as shown in Table 1.

We have also analyzed the relative potential energies of these four unique conformations by quantum mechanical (QM) calculations using GAMESS^{55,56} (Table 1). All QM calculations in the gas phase and in implicit solvent^{57,58} appear to indicate that conformation 1 is the lowest-energy

Table 2. Comparison Between Observed and Calculated NOE Values from the α -D-Man-(1 \rightarrow 3)- α -D-Man-O-Me Disaccharide^a

proton 1	proton 2	NOE observed		NOE calculated									
		absolute	relative	conf. 1		conf. 2		conf. 3		conf. 4		MD	
				abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
H ₁	H ₂	0.11	1.0	0.12	1.0	0.12	1.0	0.12	1.0	0.13	1.0	0.09	1.0
	H ₃	0.11	1.0	0.18	1.5	0.13	1.08	0.0	0.0	0.0	0.0	0.13	1.4
	H ₂	0.0	0.0	0.0	0.0	-0.01	-0.08	0.01	0.08	0.14	1.08	0.0	0.0
	H ₄	0.01	0.1	0.04	0.33	0.0	0.0	0.17	1.42	0.0	0.0	0.0	0.0
H ₂	H ₁	0.065	1.0	0.11	1.0	0.11	1.0	0.10	1.0	0.11	1.0	0.10	1.0
	H ₅	0.04	0.60	0.02	0.18	0.08	0.73	0.0	0.0	0.0	0.0	0.08	0.8
RMSD				1.3		0.65		6.56		1.78		0.63	

^a The four conformations are those in Figure 2a, and the experimental data are from Reference 49. Clearly, of all unique structures, conformation 2 has the closest NOE values to experimental data as demonstrated by its RMSD. The time-averaged NOE values from all MD trajectories in Figure 4 are close to that of conformation 2.

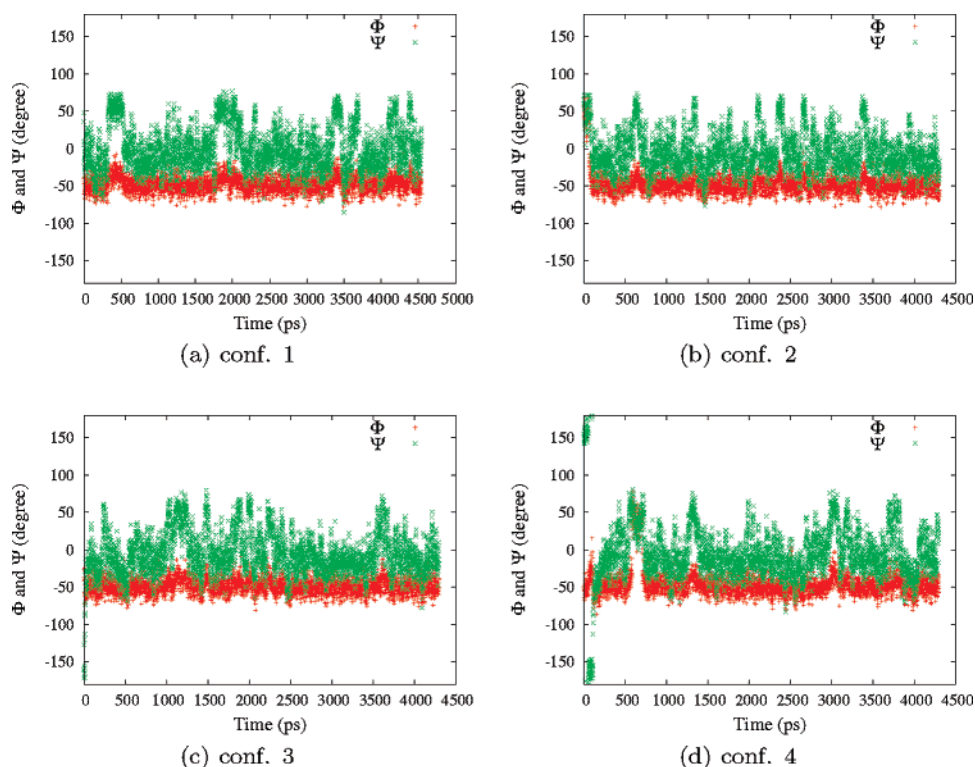


Figure 4. Time evolutions of ϕ and ψ for different unique conformations as shown in Figure 2a using the OPLS-AA force field and explicit SPC water. Conformations 3 and 4 are not preferred, and conformation 2 is visited more frequently than conformation 1.

minimum, similarly to what we found in the case of the MM calculations. We also note that QM calculations in the gas phase change the energy ordering of conformations 2 and 3. This does not occur when we use an implicit solvent model. We have also computed ab initio free energies (Table 1) using a harmonic approximation. These calculations also predict conformation 1 to be the one with the lowest free energy.

3.1.3. The NOE Sorting Criteria. The goal of the FSPS is to predict the most likely structure or structures of oligosaccharides in solution. In order to evaluate and compare the four unique conformations (Figure 2) found by our algorithm, we computed their corresponding NOEs using the procedure described by Cumming and Carver^{36,37} from the model-free approach.^{38,39} We find that, even though, under the approximations used here, conformation 1 is the global-energy

and free-energy minimum in implicit solvent, conformation 2 has in fact a NOE closest to the experimental data as shown in Table 2!

The fact that conformation 2 is indeed the most likely structure in solution is confirmed by our molecular dynamics simulations in explicit solvent. In fact, we predict a free-energy difference of about 3 kcal/mol between conformation 2 and conformation 1 (see subsection 3.1.4). This indicates that gas-phase energies, energies in implicit solvent, or free energies computed using a harmonic approximation may not be an adequate estimator for the most likely structure in solution. This may be due to anharmonic effects or to hydrogen bonding with the solvent. It is well-known that sugars easily form structures stabilized by water-mediated hydrogen bonds.⁸

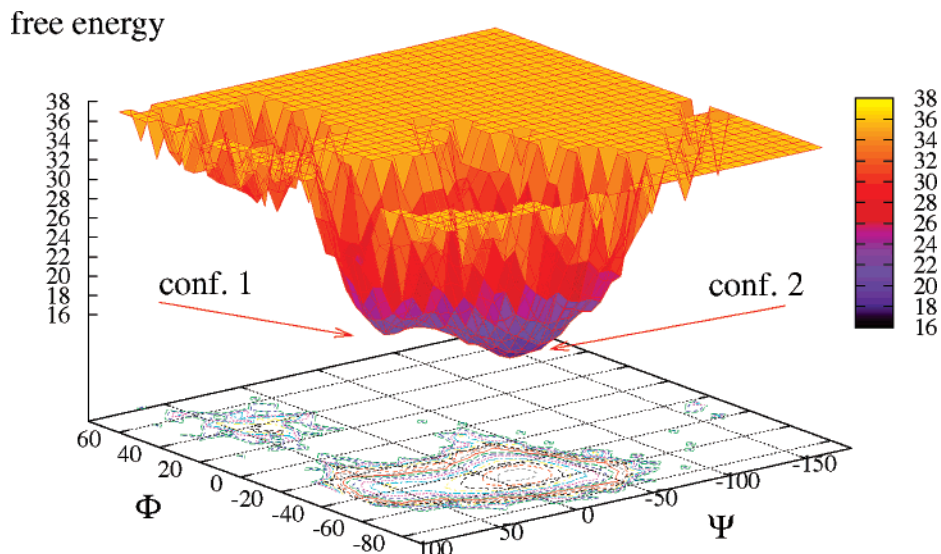


Figure 5. Free energy calculated from the probability distribution of ϕ - ψ obtained from the time evolutions shown in Figure 4. Conformation 2 is the global free-energy minimum in solution, while conformation 1 is the global energy minimum in implicit solvent and the gas phase. The free-energy difference between these two minima is about 3 kcal/mol at 298 K.

Since NOE values correspond to an average over an ensemble of structures, in general, there is not a one-to-one correspondence between a set of NOE values and a particular sugar conformation. However, it appears that, at least in the case of this sugar, free-energy differences in solution between conformers are large enough that the NOE values are dominated by only one conformational structure. Our FSPS can easily enumerate all reasonable minimum-energy structures. It takes a matter of seconds to sort these structures according to their root-mean-square deviation with respect to experimental NOEs. On the basis of our results in this article and those in the second paper, it appears that this sorting criterion is a reliable estimator for the most likely structures in solution.

3.1.4. The Free-Energy Landscape in Explicit Solvent. We used GROMACS^{32,33} with the OPLS-AA force field³⁵ in explicit simple point charge (SPC) water⁵⁹ to model the dynamics of our system. We started runs from each of the four different unique conformations in Table 1. Figure 4 shows the time evolution of ϕ and ψ for each run. Regardless of the initial conformation, it is obvious that the molecule readily transfers between conformations 1 and 2 when the system reaches equilibrium. Conformations 3 and 4 are not preferred in water. It is also clear from the plot that molecules spend more time in conformation 2 than in conformation 1. The time-averaged NOE values from these four MD trajectories shown in Figure 4 are close to that of unique conformation 2 as shown in Table 2 and closely coincide with experiments.

From the time evolution of the dihedral angles, we computed the free energy $f = -KT \ln P(\phi, \psi)$, where $P(\phi, \psi)$ is the probability distribution of $\phi - \psi$. In Figure 5, we see that unique conformation 2 is indeed the global free-energy minimum with a free-energy difference at 298 K of about 3 kcal/mol with respect to unique conformation 1, which is a metastable state. Hence, free-energy calculations from explicit MD coincide with the very inexpensive a priori

prediction of our FSPS on the basis of the deviation of single unique structure NOE values with respect to experiments.

4. Conclusions

Much can be learned from the α -D-Man-(1 \rightarrow 3)- α -D-Man-O-Me system since it has been fully experimentally characterized and since MD time scales are suitable to correctly capture the relative probability of all minima and therefore the corresponding free-energy landscape. It is clear that, in order to predict which conformer is the most likely in solution only on the basis of energetics, the correct relative probability (i.e., the free-energy landscape) of the conformers must be obtained. This probability landscape was accessible in this case because the molecule in question is relatively small and the dynamics is ergodic on the time scale of our simulations. For larger sugars, particularly branched sugars or sugars with adjacent linkage points, this brute-force approach is simply not viable.

Our automatic structure prediction algorithm was able to capture all corresponding energy minima in a tiny fraction of the time required to carry out molecular dynamics simulations long enough to sample them. A simple sorting criterion based on energies or free energies in implicit solvent was not adequate to establish a ranking for these conformers in solution. On the other hand, given the experimental NOEs, a ranking can be devised on the basis of the RMSD between these and those computed from our unique structures. The systematic search algorithm combined with the RMSD sorting criteria provides an accurate definition for the lowest free-energy structure without the need to run any expensive MD simulations. Identifying structure 2 as the most likely configuration in solution (even though its predicted energy in an implicit solvent was higher than that of structure 1) took a minute fraction of the time required to carry out the MD simulations which later confirmed the result.

Our approach provides a viable way to analyze the structure of oligosaccharides since, in our experience, for

sugars with six or seven arbitrarily connected rings, the most relevant energy minima can be obtained within a time scale of hours. By comparing the NOEs of each of these structures against experiments, it is fairly easy to establish a ranking of structures in solution. In the second paper, we show that our algorithm is able to capture many more stable local minima than those previously found by carrying out explicit solvent MD simulations. We will also show that our sorting criteria indeed capture the most likely structures in solution. These results are very promising, and we hope that the study of complex oligosaccharides will become easier as our database of fragments becomes larger.

Acknowledgment. This research was funded by Grant #05-2182 from the Roy J. Carver Charitable Trust awarded to C.J.M. and by the Skou Fellowship from the Danish Natural Sciences Research Council awarded to J.H.J.

References

- (1) Reference deleted in press.
- (2) Wölfl, M.; Batten, W. Y.; Posovszky, C.; Bernhard, H.; Berthold, F. *Clin. Exp. Immunol.* **2002**, *130*, 441–448.
- (3) Schaade, L.; Thomssen, R.; Ritter, K. *Z. Naturforsch., C: J. Biosci.* **2000**, *55*, 1004–1010.
- (4) Simons, K.; Toomre, D. *Nat. Rev. Mol. Cell Biol.* **2000**, *1*, 31–39.
- (5) Simon, B. M.; Malisan, F.; Testi, R.; Nicotera, P.; Leist, M. *Cell Death Differ.* **2002**, *9*, 758–767.
- (6) Takamiya, K.; Yamamoto, A.; Furukawa, K.; Zhao, J.; Fukumoto, S.; Yamashiro, S.; Okada, M.; Haraguchi, M.; Shin, M.; Kishikawa, M.; Shiku, H.; Aizawa, S.; Furukawa, K. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 12147–12152.
- (7) Nohara, K.; Ozawa, H.; Taic, T.; Sajib, H.; Fujimaki, H. *Biochim. Biophys. Acta* **1997**, *1345*, 207–214.
- (8) Veluraja, K.; Margulis, C. J. *J. Biomol. Struct. Dyn.* **2005**, *23*, 101–111.
- (9) Margulis, C. J. *J. Phys. Chem. B* **2005**, *109*, 3639–3647.
- (10) Almond, A.; Petersen, B. O.; Duus, J. *Biochemistry* **2004**, *43*, 5853–5863.
- (11) Woods, R. *Glycoconjugate J.* **1998**, *15*, 209–216.
- (12) Boone, M. A.; Striegel, A. M. *Macromolecules* **2006**, *39*, 4128–4131.
- (13) Striegel, A. M. *Macromolecules* **2003**, *125*, 4146–4148.
- (14) Gordon, D. B.; Hom, G. K.; Mayo, S. L.; Pierce, N. A. *J. Comput. Chem.* **2003**, *24*, 232–243.
- (15) Desmet, J.; Spriet, J.; Lasters, I. *Proteins: Struct. Funct. Genet.* **2002**, *48*, 31–43.
- (16) Kono, H.; Saven, J. G. *J. Mol. Biol.* **2001**, *306*, 607–628.
- (17) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. *Proteins: Struct. Funct. Genet.* **2000**, *40*, 389–408.
- (18) Petrella, R. J.; Lazaridis, T.; Karplus, M. *Folding Des.* **1998**, *3*, 353–377.
- (19) Huang, E. S.; Koeh, P.; Levitt, M.; Pappu, R. V.; Ponder, J. W. *Proteins: Struct. Funct. Genet.* **1998**, *33*, 204–217.
- (20) Roland, L.; Dunbrack, J.; Cohen, F. E. *Protein Sci.* **1997**, *6*, 1661–1681.
- (21) Bower, M. J.; Cohen, F. E.; Dunbrack, R. L. *J. Mol. Biol.* **1997**, *267*, 1268–1282.
- (22) DeMaeyer, M.; Desmet, J.; Lasters, I. *Folding Des.* **1997**, *2*, 53–66.
- (23) Marcusa, E.; Kellera, D. A.; Shibataa, M.; Ornsteinb, R. L.; Rein, R. *Chem. Phys.* **1996**, *204*, 157–171.
- (24) Desjarlais, J. R.; Handel, T. M. *Protein Sci.* **1995**, *4*, 2006–2018.
- (25) Roland, L.; Dunbrack, J.; Karplus, M. *J. Mol. Biol.* **1993**, *230*, 543–574.
- (26) Koehl, P.; Delarue, M. *Nat. Struct. Biol.* **1995**, *2*, 163–170.
- (27) Leach, A. R. *J. Mol. Biol.* **1994**, *235*, 345–356.
- (28) Kolinski, A.; Godzik, A.; Skolnick, J. *J. Chem. Phys.* **1993**, *98*, 7920–7433.
- (29) Balducci, R.; Pearlman, R. S. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 822–831.
- (30) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.
- (31) Jay, W.; Ponder, F. M. R. *J. Comput. Chem.* **1987**, *8*, 1016–1024.
- (32) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
- (33) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Biol.* **2001**, *7*, 306–317.
- (34) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.
- (35) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *117*, 11225–11236.
- (36) Cumming, D. A.; Carver, J. P. *Biochemistry* **1987**, *26*, 6664–6676.
- (37) Cumming, D. A.; Carver, J. P. *Biochemistry* **1987**, *26*, 6676–6683.
- (38) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4546–4559.
- (39) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4559–4570.
- (40) Ueda, N.; Aoki-Kinoshita, K. F.; Yamaguchi, A.; Akutsu, T.; Mamitsuka, H. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 1051–1064.
- (41) Yamaguchi, A.; Aoki, K. F.; Mamitsuka, H. *Inf. Process. Lett.* **2004**, *92*, 57–63.
- (42) Aoki, K. F.; Ueda, N.; Yamaguchi, A.; Akutsu, T.; Kanehisa, M.; Mamitsuka, H. *Sigmod Rec.* **2004**, *33*, 33–38.
- (43) Aoki, K. F.; Yamaguchi, A.; Ueda, N.; Akutsu, T.; Mamitsuka, H.; Goto, S.; Kanehisa, M. *Nucleic Acids Res.* **2004**, *32*, W267–W272.
- (44) McKay, B. D. *Congr. Numerantium* **1981**, *30*, 45–87.
- (45) Ullmann, J. R. *J. Assoc. Comput. Mach.* **1976**, *23*, 31–42.
- (46) Pelillo, M.; Siddiqi, K.; Zucker, S. W. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 1105–1120.
- (47) Jain, B. J.; Wysotzki, F. *Neurocomputing* **2005**, *63*, 45–67.
- (48) Xia, J.; Daly, R. P.; Chuang, F.-C.; Parker, L.; Jensen, J. H.; Margulis, C. J. **2007**, *4*, 1629–1643.

- (49) Brisson, J. R.; Carver, J. P. *Biochemistry* **1983**, *22*, 3680–3686.
- (50) Imberty, A.; Tran, V.; Pérez, S. *J. Comput. Chem.* **1989**, *11*, 205–216.
- (51) Imberty, A.; Gerber, S.; Tran, V.; Pérez, S. *Glycoconjugate J.* **1990**, *7*, 27–54.
- (52) Woods, R. J.; Pathiaseril, A.; Wormald, M. R.; Edge, C. J.; Dwek, R. A. *Eur. J. Biochem.* **1998**, *258*, 372–386.
- (53) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824–19839.
- (54) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246*, 122–129.
- (55) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.
- (56) Gordon, M. S.; Schmidt, M. W. Advances in Electronic Structure Theory: GAMESS a Decade Later. In *Theory and Applications of Computational Chemistry, the First Forty Years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005.
- (57) Cancès, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032–3041.
- (58) Li, H.; Jensen, J. H. *J. Comput. Chem.* **2004**, *25*, 1449–1462.
- (59) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981.

CT700033Y

Sugar Folding: A Novel Structural Prediction Tool for Oligosaccharides and Polysaccharides 2

Junchao Xia,^{†,§} Ryan P. Daly,^{†,§} Feng-Chuan Chuang,[†] Laura Parker,[†]
Jan H. Jensen,[‡] and Claudio J. Margulis^{*,†}

*Department of Chemistry, University of Iowa, Iowa City, Iowa 52242, and
Department of Chemistry, University of Copenhagen, Universitetsparken 5,
2100 Copenhagen, Denmark*

Received February 8, 2007

Abstract: This is the second in a set of two articles where we describe our newly developed scheme to predict conformations of complex oligosaccharides in solution. We apply our fast sugar conformation prediction tool to the case of two complex human milk oligosaccharides LNF-1 and LND-1. As described in detail in the first paper, our protocol initially delivers a set of “unique structures” corresponding to important minima on the potential-energy landscape of a complex sugar using an implicit solvent model. The nuclear Overhauser effect ranking of individual conformations provides a suitable way for comparison with available experiments. The structures obtained agree well with earlier computational predictions but are obtained at a significantly lower computational cost. Sugar conformations corresponding to stable energy minima not found by earlier molecular dynamics studies were also detected using our methodology. In order to evaluate the effects of explicit solvation and thermal fluctuations on several different predicted conformers, we also performed short-time molecular dynamics simulations in an explicit solvent.

1. Introduction

Oligosaccharides; polysaccharides; and their glycoconjugates, glycoproteins, and glycolipids play a very important role in biological phenomena such as cell–cell interaction, inflammatory processes, immunity, and fertilization.^{1,2} Conformational studies are crucial to understand biological function.¹ In most cases, the determination of an oligosaccharide's conformation involves the characterizing of the ϕ – ψ glycosidic linkages between monosaccharide residues. Glycosidic linkages of oligosaccharides can usually be quite flexible.³ Such flexibility creates difficulty on crystallization and results in considerable limitations^{4,5} when applying standard experimental techniques such as crystallography. Martin-Pastor and Bush^{6,7} pointed out that the internal motions of oligosaccharides might be classified into two kinds: the fluctuation around a single energy minimum (conformation) on the order of 15–50° and the interconversion between distinct energy minima. Depending on the free-energy difference between conformers, experimental nuclear

Overhauser effect (NOE) data may be dominated by a single conformation corresponding with a structure that is thermally fluctuating around a single minimum, or with an ensemble average over several different conformations when these are close in free energy.^{3,8} Even though conformations may be close in free energy, they are not necessarily structurally close. Due to experimental challenges in working with sugars, computational methods have become important tools to understand or predict the conformations of oligosaccharides and glycoconjugates in various environments.^{5,9–18}

Different computational methods to perform a conformation analysis of oligosaccharides in vacuo^{5,9,10,12,13} have been proposed. Some of these are based on relaxed or adiabatic maps in potential-energy surfaces for disaccharides;^{19,20} others are based on the CICADA method combined with simulated annealing to travel through conformational space.^{21–23} Monte Carlo methods²⁴ and genetic algorithms^{25,26} have also been used. Our methodology described in the first paper is fully automatic; it works on almost any oligosaccharide; it is based on a ring perception algorithm that automatically detects rotatable dihedrals, a systematic coupled dihedral space search for the whole oligosaccharide, and the use of a substructure matching algorithm that recognizes a branch within a

* Corresponding author e-mail: claudio-margulis@uiowa.edu.

[†] University of Iowa.

[‡] University of Copenhagen.

[§] Equal contributions.

complex sugar when that branch has already been studied and stored in a database. The methodology is first applied to find regions within the complex oligosaccharide dihedral space that are sterically allowed. Subsequently, minimizations are performed and structures are pooled into what we have defined in the first paper as “unique structures”. These unique structures can be sorted on the basis of different criteria such as their root-mean-square deviation against experimental NOEs, their energies in implicit solvent, or any other desired criteria.

Even though the idea of a systematic search over dihedral space for a complex oligosaccharide may appear to be an exponentially untractable problem for which Monte Carlo or other techniques could be more applicable, our experience is that, for biologically interesting sugars which are not linear, crowding severely limits the number of available conformations as the sugar becomes larger. At the same time, since dihedral angle motion becomes coupled, free-energy barriers to rotation appear to be larger and transitions between certain conformations become rare events on the typical length of a molecular dynamics (MD) or Monte Carlo simulation. In this work, we will show that the number of “unique structures” obtained for our larger sugar is smaller than that for the one with less monosaccharide units. It turns out that our systematic sampling of the whole fully coupled dihedral phase space for a complex oligosaccharide with size on the order of seven units can be easily performed within a few hours. In our approach, if further information is desired about particular structures, our “unique structures” can be used as sensible starting conformations for MD in solution. Furthermore, if part of the sugar in question has previously been studied and stored in our database, only those saved conformations and not the whole dihedral space need to be searched when further complexity is added to the molecule. The situation is different with other techniques such as molecular dynamics and Monte Carlo simulations in explicit solvent. Our experience has been that even the longest simulations currently available for complex oligosaccharides (on the order of 50 ns) only visit basins that are close to the initial MD conformation of the sugar. This is mainly because, in the case of branched sugars, torsions are strongly coupled, particularly when the branching occurs on adjacent linkages.

In this paper, we test our tool by predicting the conformations of complex oligosaccharides present in human milk. Our choice is based on the fact that these have been extensively studied via experimental NOEs and MD.

Hundreds of lactose-derived oligosaccharides exist in human milk. These oligosaccharides are the third largest component in milk and are thought to provide mechanisms of breast-feeding protection for infants against enteric pathogens.²⁷ Although conformational studies have been performed for several human milk oligosaccharides on the basis of NOEs, *J* coupling, residual dipolar couplings, and molecular dynamics simulations,^{7,15,28,29} more work remains to be done to fully determine the conformations that these oligosaccharides can take in solution and which of these are relevant to their biological function.

Our results are for oligosaccharides LNF-1 [α -L-Fucp-(1 \rightarrow 2)- β -D-Galp-(1 \rightarrow 3)- β -D-GlcpNAc-(1 \rightarrow 3)- β -D-Gal-(1 \rightarrow 4)-

β -D-Glcp] and LND-1 [α -L-Fucp-(1 \rightarrow 2)- β -D-Galp-(1 \rightarrow 3)-[α -L-Fuc-(1 \rightarrow 4)]- β -D-GlcpNAc-(1 \rightarrow 3)- β -D-Gal-(1 \rightarrow 4)- β -D-Glcp]. In a recent paper,¹⁵ Almond et al. investigated these two oligosaccharides by NMR and long (50 ns) molecular dynamics simulations in explicit water. The conclusions from their very interesting work were that these oligosaccharides possess relatively ordered structures (i.e., they fold). The authors also showed that the oligosaccharides can be easily trapped in the “wrong” free-energy minima for times as long as 50 ns, if initial structures were “incorrectly” selected. These wrong initial conditions produced trajectories that yielded incorrect NOEs. This fundamentally important result emphasizes the need to have a fast systematic way of generating all relevant sugar conformers a priori without the need to rely on MD for sampling. In this paper, we will show that our method can accomplish this in a very efficient manner. These conformers can be used to quickly predict which structure is closest to the correct experimental NOE and also to generate a family of initial structures that can be further tested via MD or other methods of choice.

2. Simulation Methods

2.1. Coarse-Graining Systematic Search. We performed conformational searches for the LNF-1 and LND-1 oligosaccharides. The scanning increment for each linkage was 10°. In order to reduce the number of conformations studied, structures have been pooled so that four adjacent points on ϕ and four adjacent points on ψ are converted into a single geometry-averaged point. This was done for each glycosidic linkage. The first dihedral angle of the longest side chain (NAc group) was also rotated with increments of 60°.

As described in the first paper, an energy minimization for each allowed conformation was carried out using the software TINKER^{30,31} with the MM3 force field³² and the generalized Born surface area (GBSA) implicit solvent model.^{33,34} For a comparison test, we also performed energy minimization using GROMACS^{35,36} with the OPLS-AA force field³⁷ in the gas phase. CPU times for a full conformational search and energy minimizations were less than a day on a single-processor (Intel Pentium 4 CPU, 2.80 GHz) computer.

2.2. Molecular Dynamics Simulation. After scoring the structures obtained on the basis of a comparison between experimental and computationally obtained NOEs (see section 2.3), short MD simulations on the order of 5 ns were carried out for selected structures. This was done in order to test the stability of these structures in the presence of an explicit solvent and in order to get better solvent-averaged NOEs. MD simulations were carried out using the software GROMACS³⁵ with the OPLS-AA force field.³⁷ In each case, the simulation box was 5 nm \times 5 nm \times 5 nm and the simple point charge (SPC)³⁸ water model was used to model water explicitly. Simulations were carried out under periodic boundary conditions. Constant pressure, temperature, and number of particles (NPT) simulations were carried out at $T = 300$ K and $P = 1$ atm. The Nose–Hoover thermostat^{39,40} and the Berendsen pressure coupling scheme⁴¹ were used for this purpose. A time step of 0.001 ps was used for integration.

2.3. Nuclear Overhauser Enhancement. In the first paper, we showed that sorting structures according to their

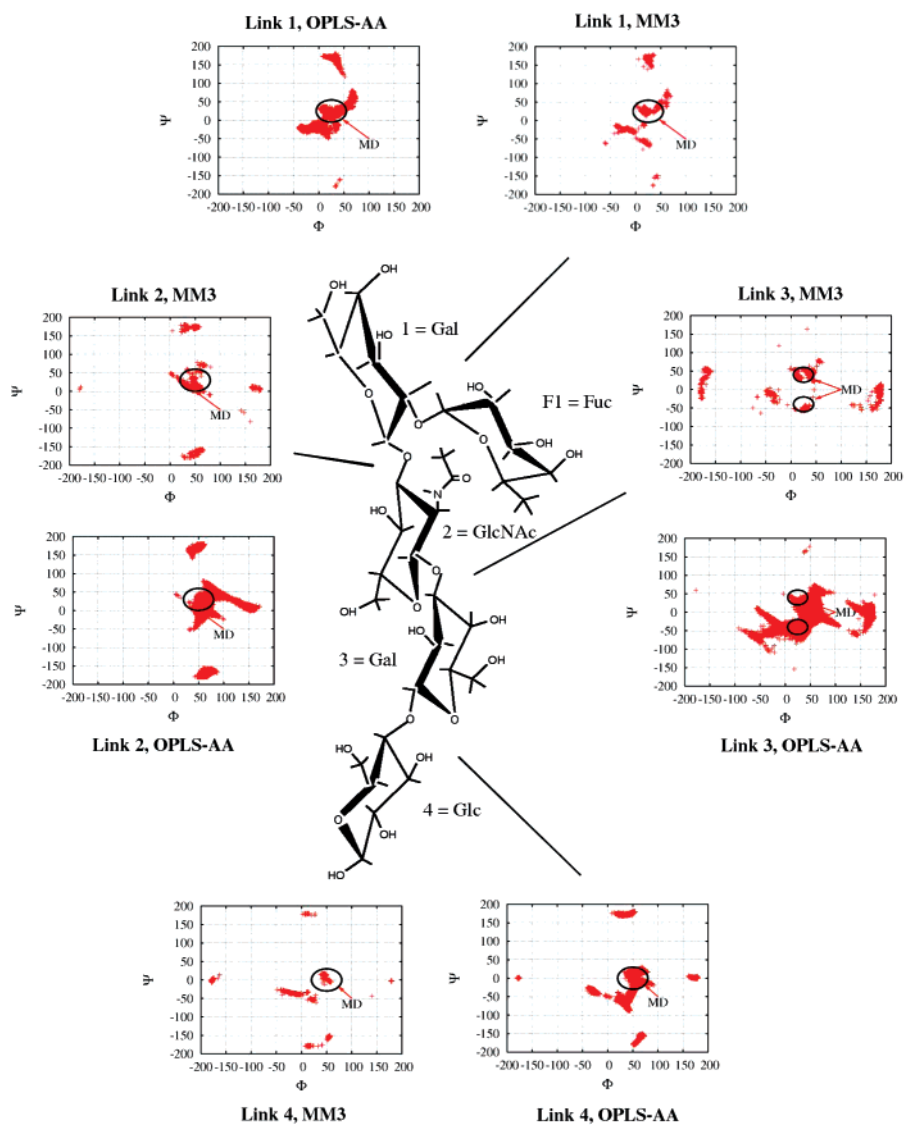


Figure 1. Distributions of all ϕ - ψ pairs from 1108 unique conformations of LNF-1 with the MM3 force field and 7030 with the OPLS-AA force field. Clearly, the distribution for each linkage is clustered around several important regions. The circled regions are those found in ref 15 by extensive 50 ns molecular dynamics simulations using the CHARMM⁴⁴ force field and the explicit TIP3P⁴⁵ water model. As it is clear from this picture, our method captures many more allowed regions than the MD simulations. MD is not able to visit most of these structures because transitions between these are rare events on a nanosecond time scale. In our database, we have linked information about these regions, a vector in $2n$ dimensional space (here, n is the number of dihedral linkages and rotatable side chains).

energies in implicit solvent or free energies computed using a harmonic approximation does not necessarily yield the most likely conformations in solution. This is because of the effect of entropy, anharmonicity, and explicit solvation on the free-energy landscape. On the other hand, a good estimator appears to be

$$\text{RMSD} = \sqrt{\frac{\sum_i^N \left(\frac{\sigma_i - \sigma_{0i}}{\sigma_{0i}} \right)^2}{N}} \quad (1)$$

where σ_i is our calculated NOE value for the i th proton pair and σ_{0i} is the corresponding experiment value, and the summation is over all available N experimental NOEs. This estimator could fail in the case in which several local free-energy minima are within a small fraction of KT from each

other. This would correspond to the case in which the sugar does not have a well-defined fold and its structure is more consistent with a random configuration. This does not appear to be the case for the sugars studied in this article.¹⁵ Such a case would be of little interest from a sugar-folding prediction perspective. As we will demonstrate later in this manuscript, our prediction method using this structure-sorting scheme agrees well with very expensive MD simulations previously published¹⁵ when these are started from the proper initial conditions. In order to obtain better thermal averages, once relevant conformations are identified, short MD simulations in explicit solvent can be used to refine the NOE results obtained from individual configurations.

Extensive literature is available on the nuclear Overhauser enhancements of oligosaccharides or glycoconjugates in solution; see, for example, refs 4 and 5. In order to create

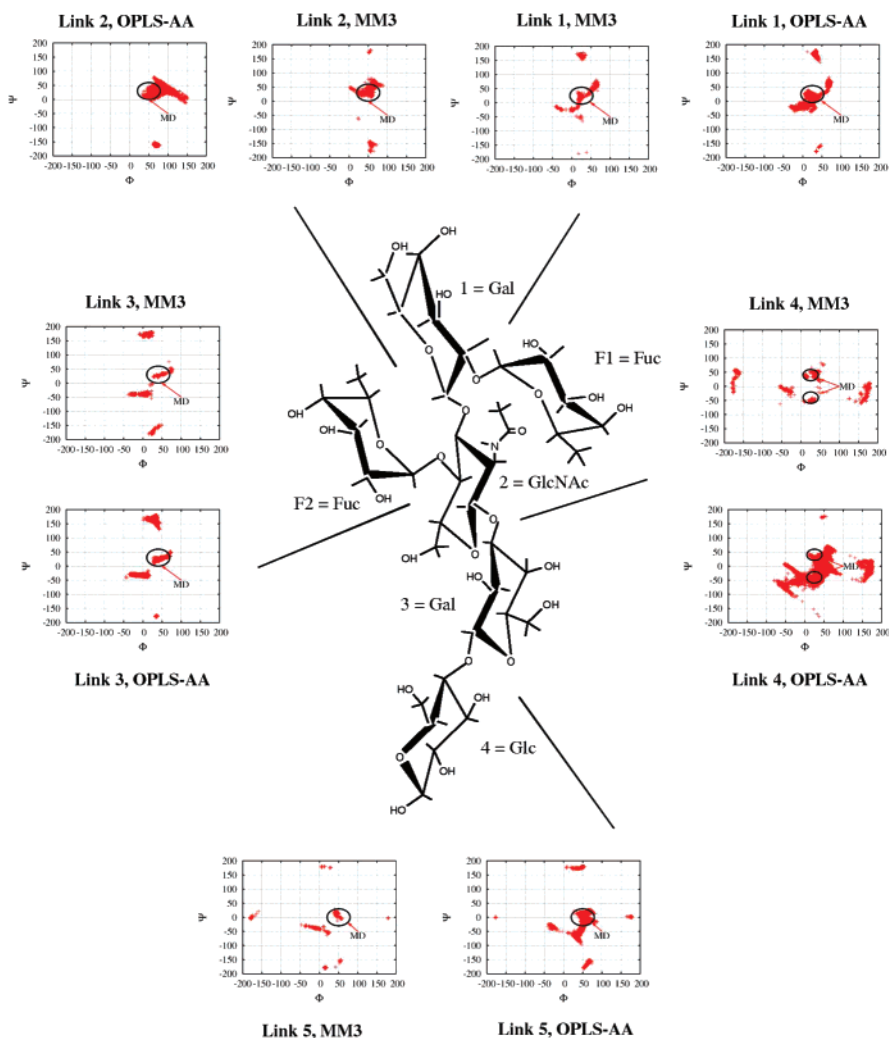


Figure 2. Distributions of all ϕ - ψ pairs from 989 “unique conformations” of LND-1 using the MM3 force field and 5220 conformations obtained from minimization using the OPLS-AA force field. The distribution of important regions is similar to that previously found for LNF-1 shown in Figure 1. The circled regions are those sampled by the 50 ns MD simulations in ref 15. Link 2 has a smaller allowed dihedral space than in the case of LNF-1 because of hindrance due to the presence of link 3.

our root-mean-square deviation sorting scheme, we have coded the model-free approach^{42,43} into our tool. NOEs from selected initial conditions obtained from MD time averages were also computed according to the scheme of Cumming and Carver.^{3,8}

3. Results and Discussions

3.1. Coarse-Graining Grid Search. We applied our method to determine the possible conformations of LNF-1 and LND-1 human milk sugars in solution. Detailed chemical structures of these molecules are shown in Figures 1 and 2. These two fucosylated oligosaccharides have similar structures. LND-1 has an additional α -L-Fuc connected to GlcNAc.

3.1.1. LNF-1 Milk Sugar. Dihedral space search and structure pooling resulted in 24 041 allowed conformations for LNF-1. After energy minimization, only 1108 of these were defined by the program as “unique conformations”. We have chosen the criteria $\Delta E < 5.0$ kcal/mol, $\Delta\psi < 10^\circ$, and $\Delta\phi < 10^\circ$ to define a “unique conformation”. According to our experience, the number of unique conformations could have been reduced ever further to less than 100 if we would

have chosen $\Delta\psi < 50^\circ$ and $\Delta\phi < 50^\circ$. Even though the grid may seem too coarse in this case, 50° is a reasonable number since it is compatible with the size of our energy basins at thermal conditions. We know this from the time evolution of ϕ - ψ for each linkage in our molecular dynamics simulations. Nonetheless, we have used the finer grid since the algorithm was fast enough that all minimizations could be carried out on a single PC in less than a day.

The insert graphs in Figure 1 show the distributions of all ϕ - ψ pairs of unique conformations for LNF-1 using the MM3 force field and the GBSA implicit solvent model as well as the OPLS-AA force field in the gas phase. As mentioned before, these structures are sterically allowed and energy-minimized. It is obvious from this figure that the ϕ - ψ distribution for each linkage is clustered into several different important regions. In particular, the circled regions are those previously reported by Almond et al.¹⁵ from one of their two very long 50 ns molecular dynamics simulations which matched the correct experimental NOE values. Clearly, our exhaustive search generated a much larger pool of allowed conformational regions. Another interesting feature of these plots is that both MM3 in implicit solvent and OPLS-AA in

Table 1. Potential Energy Differences (kcal/mol) of Four Selected Unique Structures of LNF-1 from Figure 1^a

conformation	link 1	link 2	link 3	link 4	ΔE (GR)	ΔE (TK)
conf. 1	(11.0, 26.4)	(165.7, 11.3)	(23.9, 56.2)	(-177.9, -3.1)	0.0	0.0
conf. 2	(24.1, 22.7)	(45.5, 5.5)	(28.3, -53.7)	(41.8, 4.6)	5.96	4.33
conf. 3	(-4.1, -27.6)	(46.6, 169.5)	(25.0, -55.6)	(-13.3, -37.4)	6.51	4.864
conf. 4	(65.9, 73.1)	(54.7, 12.5)	(15.5, 34.2)	(17.4, -178.4)	11.15	8.563

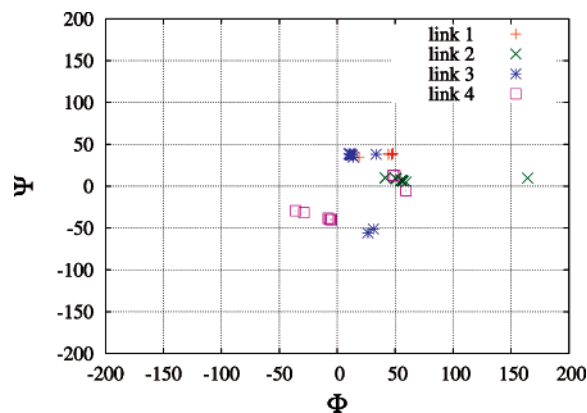
^a TK corresponds to energies calculated from TINKER using the MM3 force field and the implicit GBSA solvent model, and GR corresponds to calculations using GROMACS with OPLS-AA in the gas phase. The underlined pairs of angles are not within the circled regions corresponding to best NOE values shown in Figure 1. Only conf. 2 has all linkages within the circled regions.

Table 2. Comparison of NOEs Computed for Different Proton Pairs in Each of Our Selected Conformers of LNF-1 against Experimental and MD Values

proton pairs		NOE calculated									
		exp. ¹⁵		conf. 1		conf. 2		conf. 3		conf. 4	
				ind.	MD	ind.	MD	ind.	MD	ind.	MD
F1 H2	F1 H1	5.5	6.9	10.5	13.6	10.9	14.0	10.6	14.0	13.7	14.4
F1 H5	F1 H1	0.6	0.4	0.5	0.5	0.5	0.5	0.6	0.5	0.4	0.5
F1 H3	F1 H5	4.9	6.5	5.4	4.0	4.8	3.9	5.4	4.2	5.6	4.0
F1 H4	F1 H5	5.8	7.5	7.9	6.7	7.4	6.8	7.7	7.0	8.4	6.9
1 H5	1 H1	6.2	6.8	6.1	9.7	7.7	9.8	6.3	9.5	9.4	9.9
1 H2	1 H1	2.5	1.8	1.5	2.6	2.3	2.7	1.8	2.5	2.1	2.7
1 H3	1 H1	5.7	5.0	4.4	5.1	6.6	5.1	4.8	5.6	5.1	5.2
1 H4	1 H5	5.8	7.2	7.9	7.6	8.6	7.5	8.2	7.3	7.9	7.4
2 H2	2 H1	2.2	1.8	4.2	2.6	2.6	2.5	2.1	2.0	3.9	2.6
2 H3	2 H1	7.0	2.1	5.4	4.0	3.8	3.9	8.0	6.0	5.7	4.1
2 H5	2 H1	7.7	9.9	11.3	10.4	9.1	10.5	9.3	9.8	11.6	11.3
4 H3	4 H1	3.0	4.0	7.3	4.3	6.9	4.1	5.9	4.3	4.1	3.7
4 H5	4 H1	7.8	6.4	9.7	11.0	10.6	11.1	10.8	11.0	7.2	10.5
1 H2	F1 H1	6.8	5.8	11.2	9.0	19.9	9.0	19.2	10.8	-0.4	9.9
1 H3	F1 H1	0.5	0.7	-1.0	0.9	-1.4	0.9	-0.2	0.6	6.0	0.6
1 H2	F1 H5	1.3	2.4	0.1	2.2	0.6	2.3	0.1	1.7	11.4	2.1
2 H2	F1 H5	6.8	4.7	0.1	9.8	7.0	10.0	0.0	0.0	0.2	8.8
2 H4	F1 H5	0.6	2.5	0.1	0.0	-0.3	0.0	0.2	0.1	0.3	0.2
2 H3	1 H1	5.7	5.6	0.5	7.8	9.2	7.9	0.3	0.5	7.9	8.1
3 H1	2 H1	0.4	0.3	-0.2	-0.3	-0.4	-0.2	-0.5	-0.3	-0.3	-0.3
3 H3	2 H1	11.5	9.8	6.1	7.2	10.5	7.2	11.5	8.5	13.5	7.5
3 H4	2 H1	0.8	1.1	-0.4	-0.2	3.6	0.0	5.2	0.3	-0.9	-0.5
3 H3	2 H5	0.5	0.9	-0.2	-0.3	-0.8	-0.4	-0.9	-0.6	-0.7	-0.3
4 H4	3 H1	11.0	11.3	0.4	0.6	13.7	11.7	11.9	11.6	0.4	8.6
RMSD			0.74	1.41	0.78	1.01	0.74	1.50	0.77	2.89	0.73
RMSD rank				2	4	1	2	3	3	4	1

^a The columns labeled "MD¹⁵" and "exp.¹⁵" correspond to NOEs from the MD simulations and experimental measurements of Almond and co-workers in ref 15. The subcolumn labeled "ind." represents NOEs calculated from a single individual conformer, and that labeled "MD" corresponds to our time-averaged NOEs computed from short 5 ns MD trajectories using as starting conditions conformer 1, 2, 3, or 4.

the gas phase appear to produce similar sugar conformations. We suspect this is generally true for all available force fields even though relative energies in each case may be different. These energetic differences which we have observed with different force fields and ab initio calculations do not significantly affect our results since, as we have shown in the first paper, the energy ranking in implicit solvent does not generally coincide with the ranking of free-energy minima in solution, which is what determines the corresponding NOE values. Almond and co-workers' important study sheds light on the fact that even during very long molecular dynamics simulations the full configuration space is not readily visited. This is clear from the fact that their two trajectories produced significantly different NOE values. Only one of them being close to the correct experimental value. The reason for this is that, in the case of complex branched oligosaccharides, typical molecular dynamics time scales are not long enough to fully sample this space. Hence, our inexpensive a priori identification of conformational regions together with our ranking of structures based on their RMS deviation with respect to the corresponding experimental NOE values provides not only a good way to identify correct configurations in solution but also a way to generate

**Figure 3.** Distributions of all ϕ - ψ pairs from the 20 conformations of LNF-1 ranked with smallest RMSD from experimental NOE values. Most conformations are located within the circled regions in Figure 1.

initial conditions for further sampling with molecular dynamics in explicit solvent without having to rely on the trajectory to sample configuration space.

Table 1 exhibits the potential energies and ϕ - ψ values for four selected conformations of LNF-1 from Figure 1. Configuration 1 is the lowest-energy minimum found by the

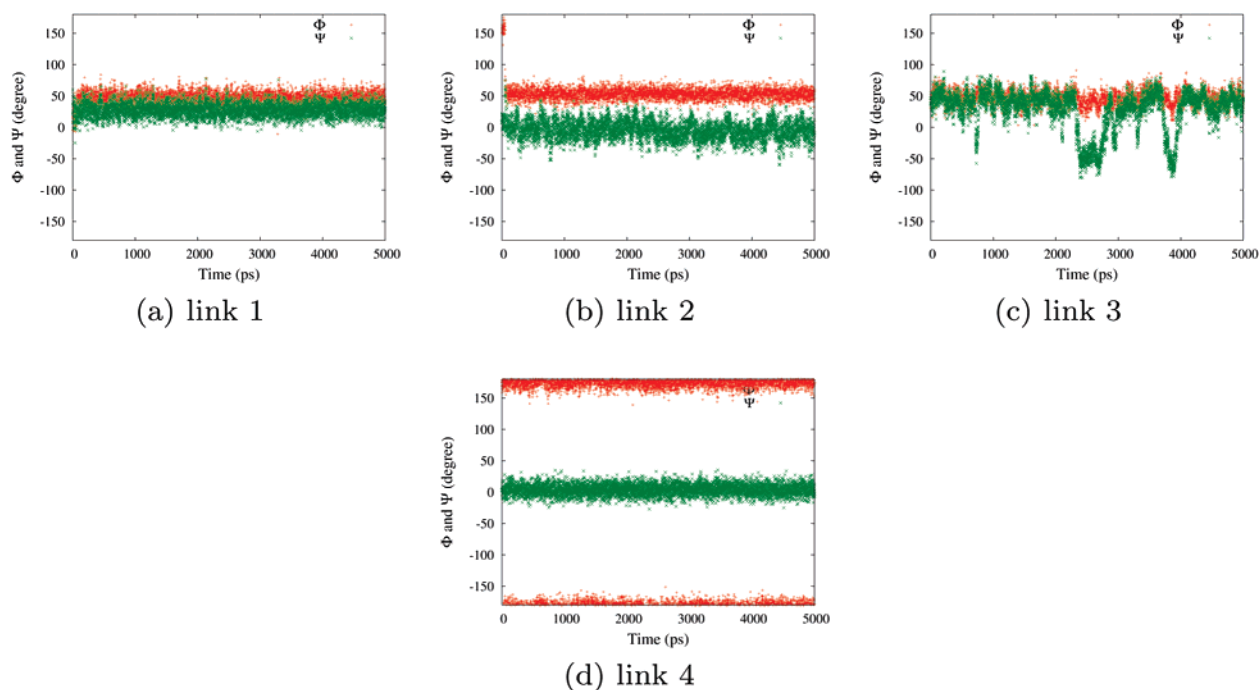


Figure 4. Time evolution of conformation 1 (global energy minimum) of LNF-1 as shown in Table 1 simulated using GROMACS with the OPLS-AA force field and the SPC explicit water model. Link 2 transfers to the circled region in Figure 1 relatively quickly (50 ps). In contrast, link 4 stays outside the circled region during our 5 ns run.

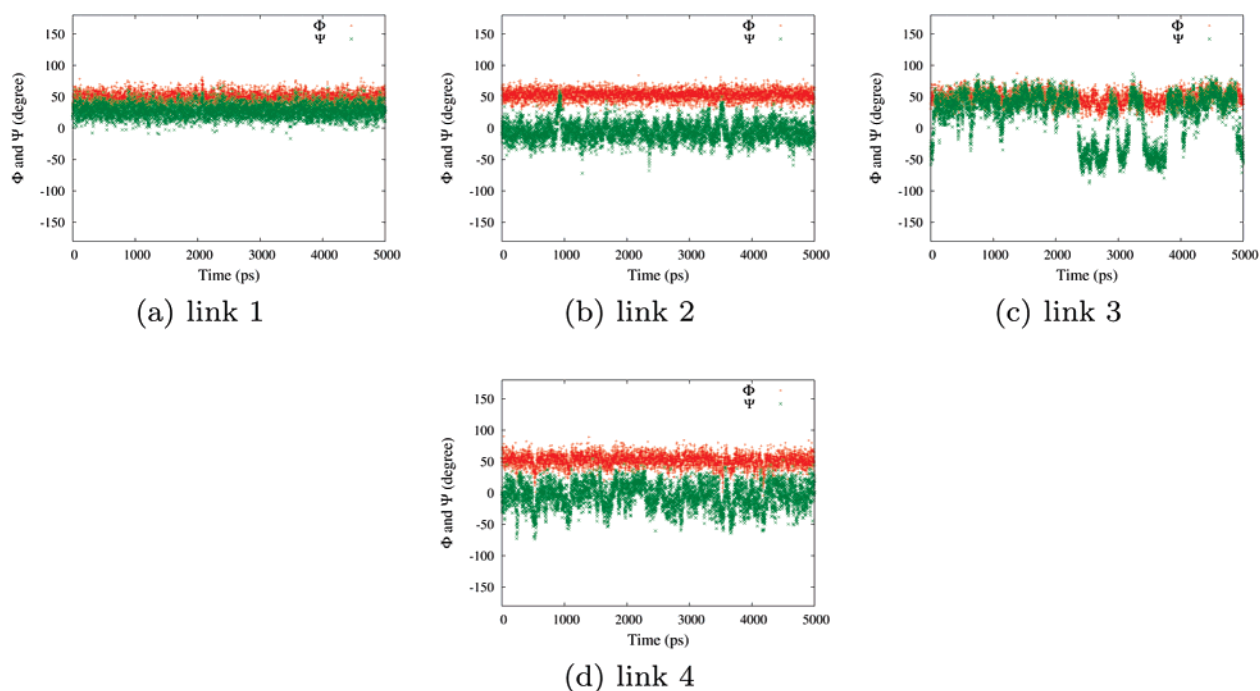


Figure 5. Time evolution of conformation 2 of LNF-1 in explicit water. All linkages fluctuate around the initial values in the circled regions as shown in Figure 1. Time-averaged NOEs in Table 2 show that this final conformation is a good candidate for the most likely structure in solution.

algorithm in implicit solvent. Conformation 2 is our candidate for best structure in solution. For conformer 2, all values of ϕ - ψ pairs are located within the circled regions in Figure 1. Conformers in this region have minimal RMS deviation from the experimental NOE values. Conformations 1, 3, and 4 have at least one linkage outside of this region. In particular, conformation 1 (our global-energy minimum in implicit solvent) has two linkages outside the circled regions.

The RMSD of its NOE values with respect to experiments is quite large as shown in Table 2. In contrast, conformation 2 is selected from Figure 3 and has the best NOE values compared to experiments,¹⁵ but its energy in implicit solvent is much higher ($\Delta E \approx 5$ kcal/mol). This result is consistent with our findings in the first paper. The effects of explicit solvent and entropy must be taken into account to obtain a good approximation of the free energy of these systems.

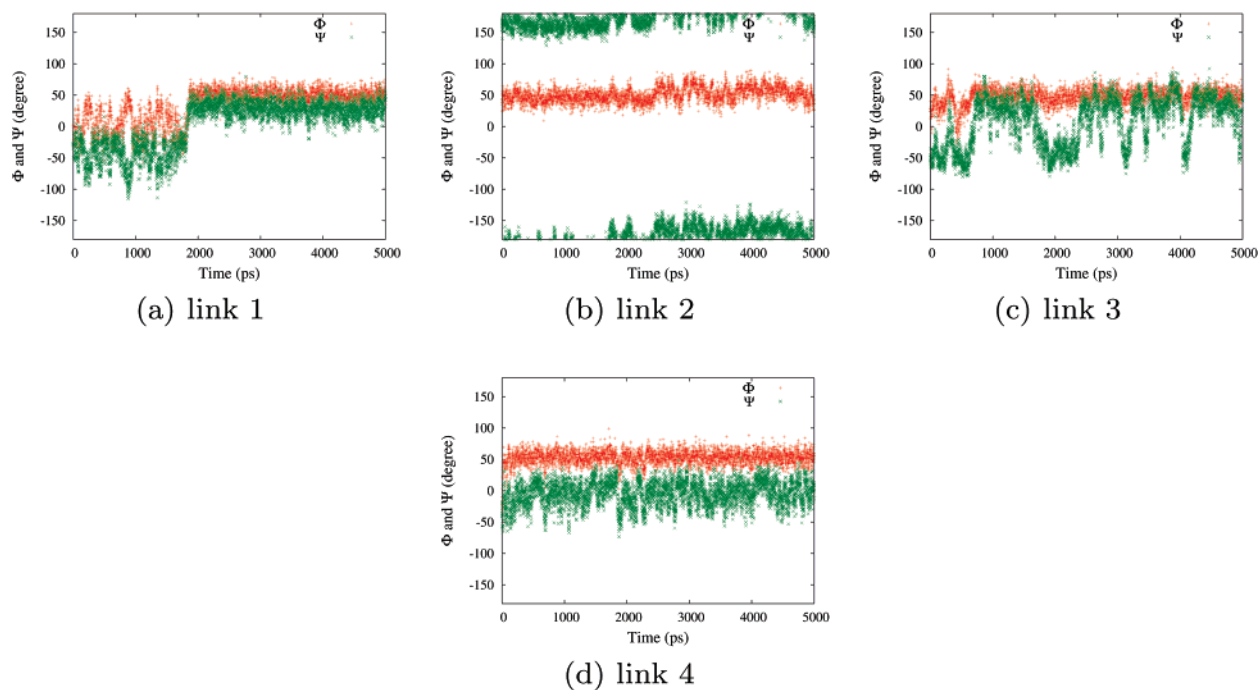


Figure 6. Time evolution of conformation 3 of LNF-1 as shown in Table 1 in explicit water. Link 1 shifts to the circled region in Figure 1 within 2 ns. Link 2 fluctuates outside the circled region.

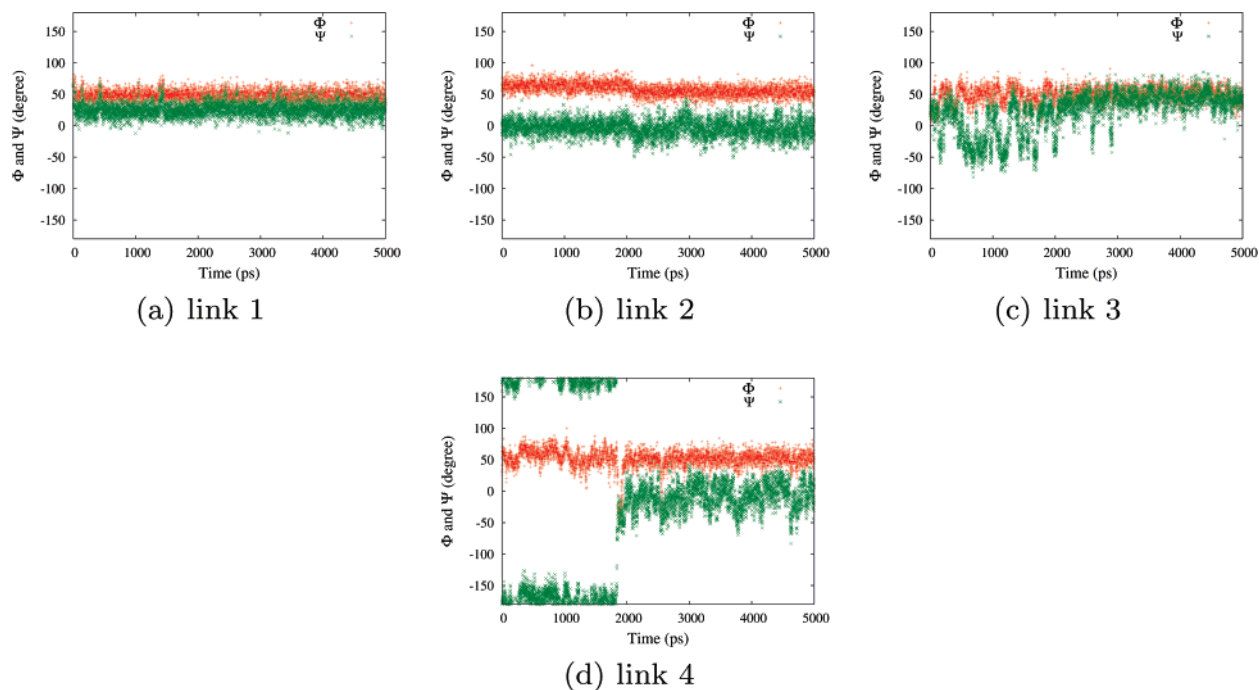


Figure 7. Time evolution of conformation 4 of LNF-1 as shown in Table 1 in explicit water. After the transition of link 4 around 2000 ps, the final conformation is the same as that of conformation 2 in Figure 5.

Conformations 3 and 4 have NOE RMSDs larger than that of conformation 2 and were chosen for comparison as initial conditions for molecular dynamics simulations.

It is clear from Table 2 that conformer 2 is the best candidate on the basis of NOEs. It is also clear from the same table that, for all initial conformers, solvent-averaged NOEs are closer to the experimental values than those resulting from single initial configurations. This is due to significant changes in conformation during MD that bring

one or more glycosidic angles closer to the values of conformer 2.

We performed relatively short 5 ns MD simulations using the software GROMACS³⁵ with the OPLS-AA force field³⁷ in SPC water³⁸ for the four selected conformations in Table 1. We are particularly interested in understanding whether the regions that our algorithm singled out as most likely in solution on the basis of NOE RMSDs are stable during explicit solvent simulations or if structures in these regions

Table 3. Final Structures from 5 ns MD Simulations with Explicit Solvent for the Four Unique Conformations in Tables 1 and 4

	LNF-1				
	link 1	link 2	link 3	link 4	
conf. 1	(50, 25)	(50, 0)	(50, -50/50)	(180, 0)	
conf. 2	(50, 25)	(50, 0)	(50, -50/50)	(50, 0)	
conf. 3	(50, 25)	(50, 180)	(50, -50/50)	(50, 0)	
conf. 4	(50, 25)	(50, 0)	(50, -50/50)	(50, 0)	
	LND-1				
	link 1	link 2	link 3	link 4	link 5
conf. 1	(50, 25)	(50, 0)	(50, 25)	(50, -50/50)	(50, 0)
conf. 2	(50, 25)	(50, 0)	(50, 25)	(50, -50/50)	(50, 0)
conf. 3	(50, 25)	(50, 180)	(50, 180)	(50, -50/50)	(50, 180)
conf. 4	(50, 25)	(50, 0)	(50, 25)	(50, -50/50)	(180, 0)

^a The ϕ - ψ values are in degrees. For LNF-1, initial conf. 2 and initial conf. 4 result in identical final conformations; the final conformations for initial conf. 1 and initial conf. 3 have only one linkage (link 4 or link 2) that is different from conf. 2. For LND-1, confs. 1 and 2 share identical final conformations after MD. Initial confs. 3 and 4 have different final conformations.

undergo significant configurational modifications. Other structures that are also local energy minima but possess several linkages outside this selected configuration space region were studied in order to gauge whether barriers to interconversion were readily crossed.

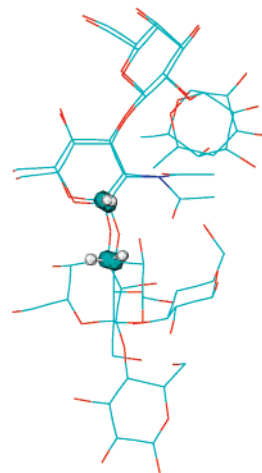
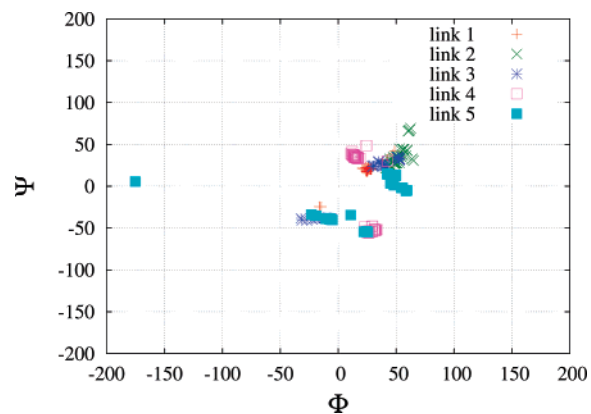
Figures 4–7 show the time evolution of the dihedral angles for the four selected conformations in explicit water. Final conformations for these four runs are listed in Table 3. Initial conformation 2 (Figure 5) has all (ϕ, ψ) values within the dihedral regions being the best NOE values as compared with those of experiments. Throughout our 5 ns simulation, the trajectory corresponding to initial conformation 2 does not depart from the angular areas circled in Figure 1. It is interesting to notice that these areas correspond to two clearly different conformations that share identical link angles 1, 2, and 4, but link 3 transitions between $\psi = -50^\circ$ and $\psi = +50^\circ$. Both of these two structures have good NOEs for the protons considered on each side of linkage 3 since at $\psi = -50^\circ$ and $\psi = +50^\circ$ the proton distances involved are very similar as can be appreciated from Figure 8. This is a clear example that shows how experimental NOEs may correspond to a linear combination of structures in different local basins instead of an average over structures in a single free-energy minimum. One should therefore be careful when experimentally assigning a structure simply on the basis of NOE constraints since these may not correspond to a single structure, but instead to a combination of several different structures.

Results from our simulations with configuration 1 (the global-energy minimum in implicit solvent) as the initial

Table 4. The Potential Energy Differences of Four Sterically Allowed Minimized Structures of LND-1 Selected from Figure 2^a

conformation	link 1	link 2	link 3	link 4	link 5	ΔE (GR)	ΔE (TK)
conf. 1	(27.6, 20.2)	(62.7, 22.1)	(72.3, 51.6)	(14.9, 38.1)	(-175.1, 5.4)	0.0	0.0
conf. 2	(28.4, 20.7)	(48.6, 27.2)	(51.5, 33.9)	(11.7, 37.8)	(-5.5, -40.0)	1.62	1.14
conf. 3	(65.9, 64.4)	(61.9, -155.8)	(18.0, 176.5)	(12.4, 39.1)	(-175.1, 4.8)	5.14	7.55
conf. 4	(27.9, 20.2)	(47.2, 28.0)	(52.9, 34.3)	(159.2, -37.7)	(-178.3, 0.5)	14.46	8.40

^a Conf. 1 is the global energy minimum in implicit solvent. Conf. 2 has all linkages within the circled regions in Figure 2 and corresponds to the one with the closest NOE values to experimental values. TK corresponds to energies calculated from Tinker using the MM3 force field and the implicit GBSA solvent model, and GR corresponds to calculations using GROMACS with OPLS-AA in the gas phase. The underlined pairs of angles are not within the circled regions corresponding to the best NOE values shown in Figure 2. Only conf. 2 has all linkages within the circled regions.

**Figure 8.** Two conformations of LNF-1 that share identical link 1, 2, and 4 angles but have different link 3 angles ($\psi = -50^\circ$ and $\psi = +50^\circ$). Both of these two structures have NOEs close to the experimental values for the proton pair (3H3–2H1) considered on each side of linkage 3. This is because the distance in each case between the protons considered is very similar, 2.36 and 2.51 Å, respectively.**Figure 9.** Distribution of ϕ - ψ angle pairs corresponding to the 20 conformations of LND-1 with smallest NOE RMSD. Most conformations fall inside the circled regions in Figure 2.

condition are shown in Figure 4. In this case, link 2 transfers to dihedral angles similar to those of configuration 2 within 50 ps. Just as in the case of starting structure 2, link 3 fluctuates between $\psi = +50^\circ$ and $\psi = -50^\circ$. The angles corresponding to link 4 remain almost constant throughout our 5 ns simulations and are different from those in configuration 2. The resulting time-averaged NOEs for trajectories with configuration 1 as the initial condition are tabulated in Table 2. Since most linkages undergo rotations

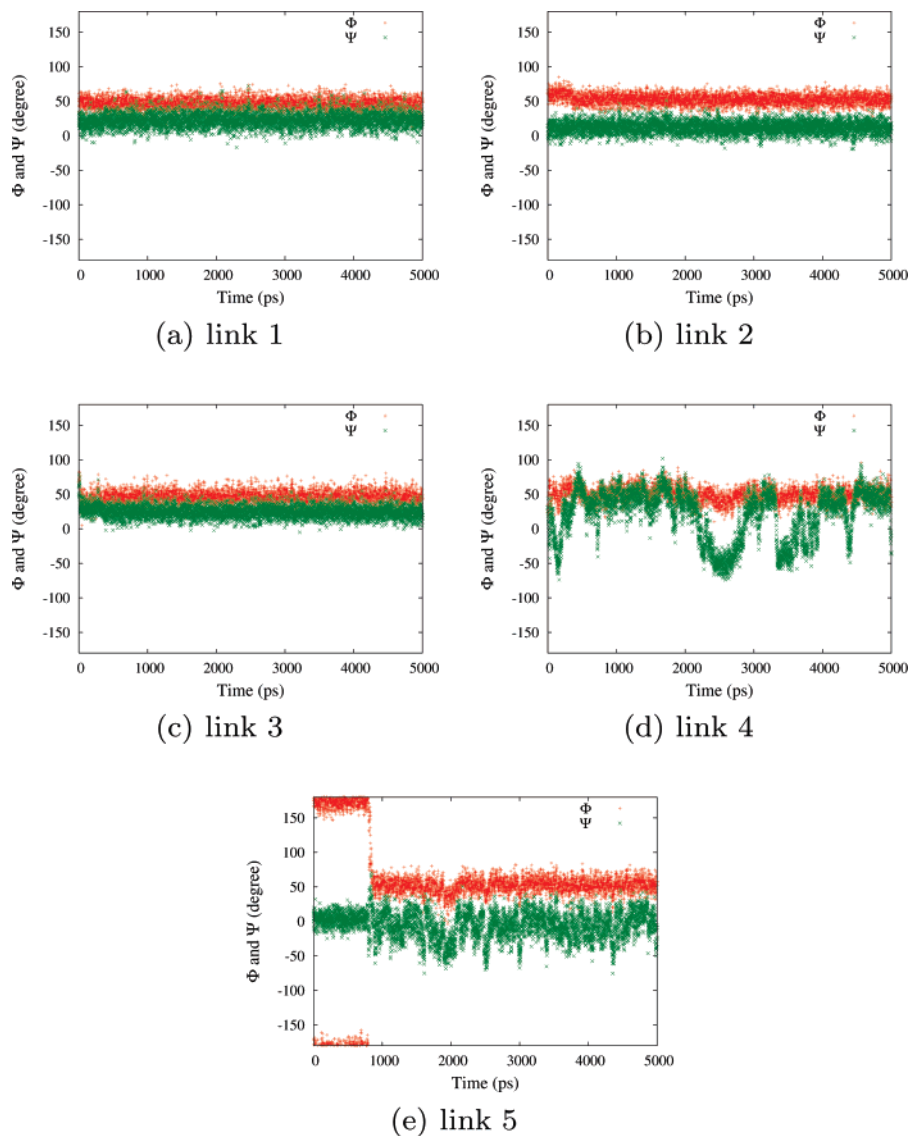


Figure 10. Time evolution of conformation 1 (global energy minimum in implicit solvent) of LND-1 as shown in Table 4 simulated using GROMACS with the OPLS-AA force field and the explicit SPC water model. Initially, link 5 is outside the circled region in Figure 2, but it switches into this region after 1000 ps. The small RMSD (Table 5) indicates that this final conformation is a good candidate for the most likely structure in solution. This structure is the one previously identified by long 50 ns MD simulations.¹⁵

to final configurations analogous to that of structure 2, it is not surprising that the value of the NOE RMSD is much smaller than that from initial structure 1. Since link 4 is a terminal residue and is far from the crowded linkage, it is likely that on a longer time scale it will undergo conformational changes. The case of conformation 3 is quite different from the previous two. After 2 ns, link 1 (Figure 6) shifts to the same angles as in conformer 2, but link 2 does not. Therefore, the NOEs corresponding to 2 H3 to 1 H1 in Table 2 are quite different from those experimentally observed. In the case of conformer 4, after link 4 of Figure 7 undergoes a transformation at around 2 ns, the conformation of the molecule is identical to that of conformer 2.

3.1.2. LND-1 Milk Sugar. In order to study LND-1, we follow two different approaches. The first one is analogous to our procedure in the case of LNF-1, and it involves the search of the whole dihedral space. The second approach,

which we use for comparison and benchmarking, makes use of our substructure recognition algorithm and database. This approach generates “unique conformations” for LND-1 on the basis of a database entry previously stored for the conformations of its subfragment LNF-1.

In the first case, after the coarse-grained grid search, we obtained only 9071 sterically allowed conformations as opposed to the case of LNF-1 (the smaller oligosaccharide) in which our algorithm found 24 041 structures. This is interesting since adding degrees of freedom to the system appears to reduce instead of increase the number of accessible regions in dihedral space. The additional branch α -L-Fuc in LND-1 is the cause for this reduction in number of allowed conformations. We expect this to be a general trend in sugars that are branched, particularly those with adjacent linkages. An extreme case in which torsional degrees of freedom are reduced to a minimum is that of cyclodextrins. When

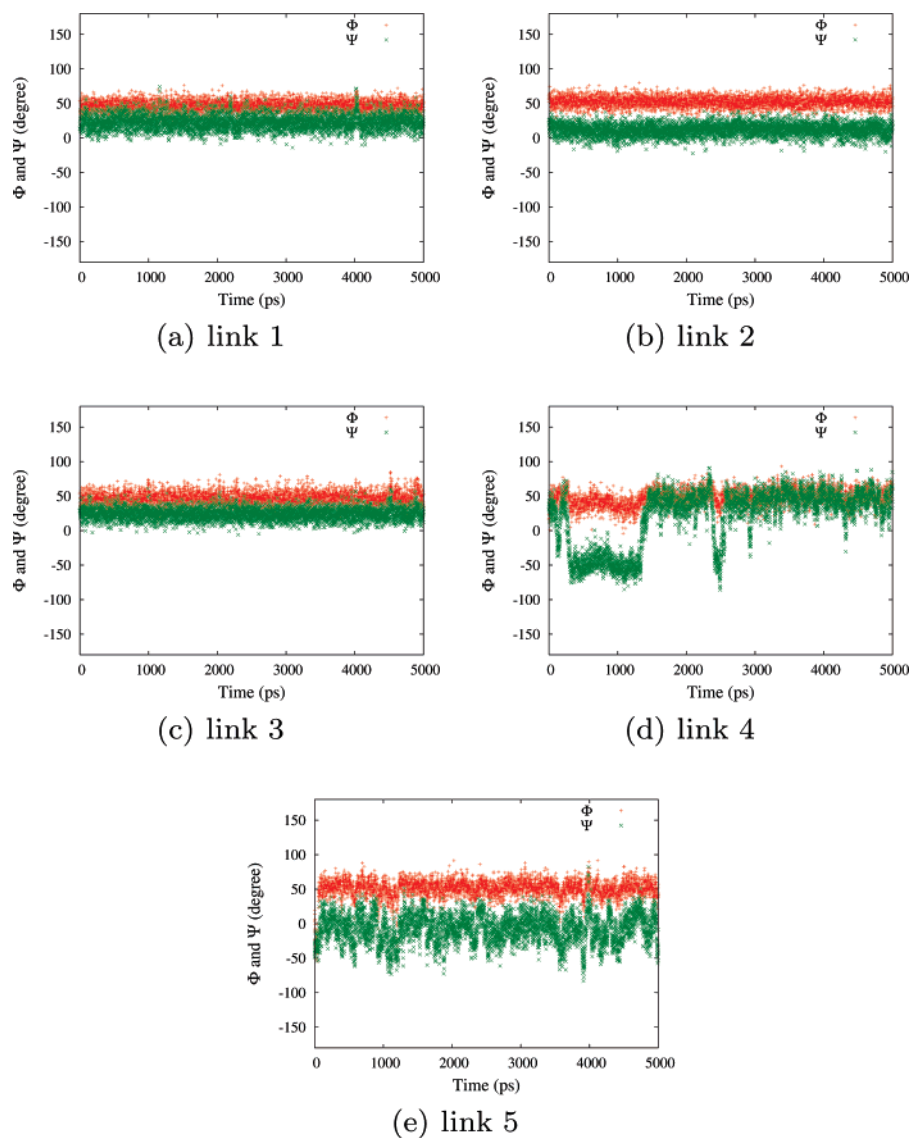


Figure 11. Time evolution of conformation 2 of LND-1 as shown in Table 4 in explicit water. All linkages fluctuate in the circled regions depicted in Figure 2. The final conformation is the same as that resulting from the MD simulation of conformation 1 (Figure 10).

identical angular and energetic criteria are used as previously described in the case of LNF-1, energy minimizations using the MM3 force field and the GBSA implicit solvent model produced 989 “unique conformations”.

The set of insert graphs in Figure 2 show the distributions of all ϕ – ψ pairs corresponding to the 989 unique conformations of LND-1. For comparison, we also show the 5220 minima obtained using OPLS-AA in the gas phase. It is clear that several energy-minimized regions in ϕ – ψ space are present in addition to those previously found during MD simulations¹⁵ (circled regions in Figure 2). Results using OPLS-AA and MM3 are qualitatively similar. When we compare LND-1 with LNF-1, we notice that in the case of LND-1 the width of certain allowed regions is narrowed due to the presence of the additional α -L-Fuc branch. Furthermore, some regions in dihedral space completely disappear in the case of LND-1. As an example, the region around (180°, 0°) for link 2 (β -D-Galp-(1→3)- β -D-GlcpNAc) is absent in the case of LND-1.

We have ranked the LND-1 conformations on the basis of their energy in implicit solvent and also on the basis of their RMSD with respect to experimental NOEs. Figure 9 exhibits the distributions of all ϕ – ψ pairs of the best 20 LND-1 conformations on the basis of RMSD. Most conformations appear to be located in the regions circled in Figure 2.

In Table 4, we compare four conformations from Figure 2. Just as in the case of LNF-1, we have chosen these four structures because one is the global-energy minimum in implicit solvent; the second one is the structure ranked best on the RMSD scale, while the third and fourth are allowed minimum-energy structures that have not been previously reported computationally but are not in the correct regions according to our NOE calculations. Table 4 displays potential-energy values and corresponding ϕ – ψ values.

Just as in the case of LNF-1, without the need of expensive MD simulations in explicit solvent, a simple NOE ranking based on our exhaustive search algorithm was able to

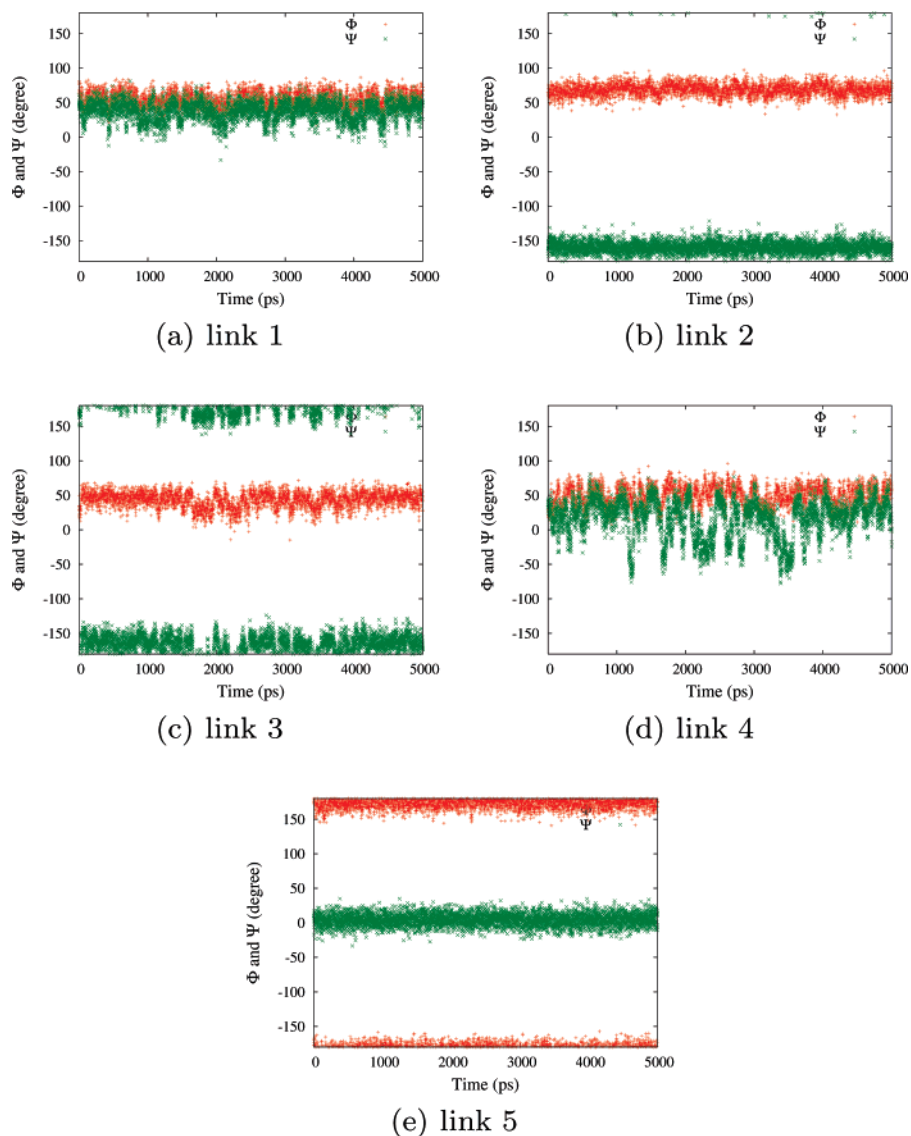


Figure 12. Time evolution of LND-1 conformation 3 as shown in Table 4 in explicit water. Links 2, 3, and 5 fluctuate outside the circled regions in Figure 2. The RMSD with respect to experimental values is large.

efficiently identify the correct regions of configuration space in which conformation 2 is located. This region coincides with that proposed by Almond and co-workers¹⁵ from one of their trajectories that was initiated from an appropriate initial conformation. Our calculation was carried out in less than 1 day on a single-processor desktop PC.

In order to quantify the advantage of using a rotameric substructure database and a substructure matching algorithm, we also analyzed LND-1 using LNF-1 as a database entry. When a new sugar is added to the database, our program adds an entry with the following information: the residue names and topology (residue connectivity, chirality of atoms, etc.) and a unique name for a file in which vectors of allowed dihedral conformations are stored representing points in dihedral phase space from which the whole oligosaccharide can be reconstructed. When in search mode, the program checks all entries in the database and calls our substructure mapping algorithm in order to determine whether there is any molecule in the database that could potentially be a substructure of the new molecule. If several substructures

are available, only the largest one is used in order to build a model for the new molecule. These models take as starting points the vectors of dihedral angles stored for the subfragment and only do full searches on the parts of the molecule not originally stored as a substructure in the database. All vectors in the database for that particular substructure are used as starting points to obtain the full dihedral phase space for the new molecule. Every time a new vector from the subfragment is retrieved, the new molecule is reassembled by adding the remaining residues and side-chain linkages. In the case of LND-1, after the conformational search and database storage for LNF-1 was performed, the search for LND-1 was carried out by simply adding one residue, which provides a branching point. The CPU time for a full search of the dihedral space of LND-1 previously described in this paper was 2665 s. In contrast, it only took 1114 s to search using the database.

For the LND-1 milk sugar, Figures 10–13 exhibit the dynamics in explicit solvent of the four initial conformations shown in Table 4. All final conformations are displayed in

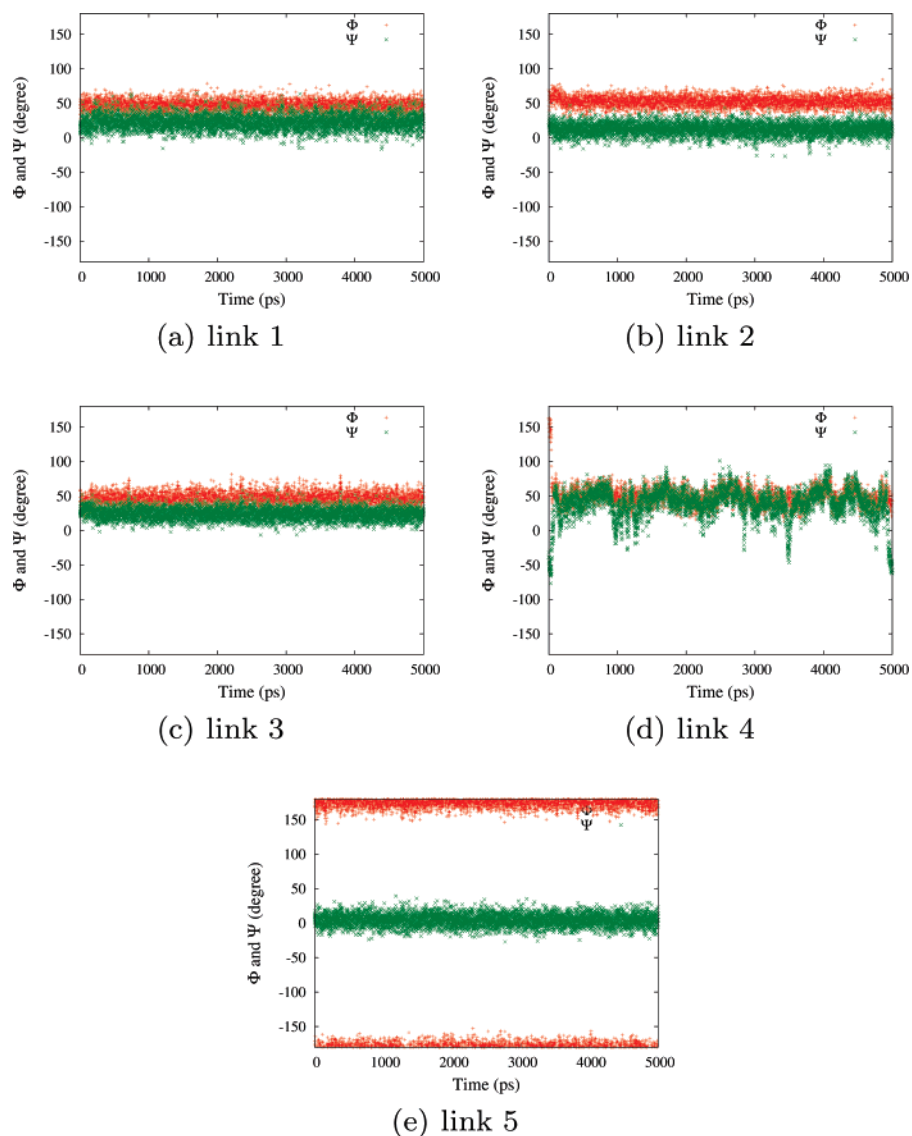


Figure 13. Time evolution of dihedral angles in LND-1 conformation 4 (Table 4) in explicit water. Initially, links 4 and 5 are not within the circled regions in Figure 2. Link 4 transfers into the circled region within a few picoseconds.

Table 3. Similar results to those obtained in the case of LNF-1 are observed. Only conformation 2 (see Figure 11) has all initial dihedral angle values within the circled NOE regions shown in Figure 2. Figure 10 shows that link 5 of conformation 1 (the global-energy minimum in implicit solvent) evolves toward the circled region in Figure 2 within 1000 ps. This final conformation is the one identified by Almond and co-workers¹⁵ and by our prediction algorithm as being the most likely in solution. As is to be expected, the time-averaged RMSD of NOE values is small, as can be appreciated in Table 5. Our time-averaged NOEs appear to be slightly worse than those in ref 15. This is reasonable since their simulations are much longer (50 ns) and their methodology involves obtaining a time correlation function, while in our case, for economy of time in our prediction procedure, we simply average over NOE values of individual snapshots along a 5 ns trajectory.

Figure 11 shows the results from our MD simulations with initial conditions corresponding to conformation 2 (the structure a priori predicted to have the best NOEs). In all

cases, dihedral angles fluctuate within the circled regions in Figure 2. The small time-averaged RMSD with respect to experiments (Table 5) indicates that this conformation is indeed a stable structure and the best candidate for solution conformation.

In the case of conformation 3 in Figure 12, links 2, 3, and 5 fluctuate outside the circled regions, and this results in a large RMSD with respect to experimental values. Figure 13 shows that, in the case of conformation 4, link 4 transfers to the corresponding circled region while link 5 remains outside the corresponding circled areas. Because of the small overall deviation of the time-averaged RMSD values (Table 5) with respect to experiments, it is possible that the final conformations from this trajectory correspond to actual structures often visited in solution. This is likely since the only difference between these structures and those in the free-energy basin corresponding to conformer 2 is the terminal unit far from the set of crowded linkages. This situation was also observed in the case of the MD simulation of conformer 1 of LNF-1.

Table 5. Table NOE Values for Different Proton Pairs for the Four Selected Structures of LND-1 from Table 4^a

proton pairs		NOE calculated											
		exp. ¹⁵		MD ¹⁵		conf. 1		conf. 2		conf. 3		conf. 4	
						ind.	MD	ind.	MD	ind.	MD	ind.	MD
1 H5	1 H1	6.1	8.9	6.6	8.4	7.3	8.5	7.0	9.0	7.2	8.5		
1 H3	1 H1	4.1	6.8	5.7	3.8	7.0	3.7	5.3	4.9	6.9	3.9		
2 H3	2 H1	9.9	5.6	7.4	4.2	6.6	4.2	5.7	5.4	5.2	4.3		
2 H5	2 H1	13.2	10.3	10.4	8.6	10.6	8.9	7.3	8.0	10.9	9.1		
F1 H2	F1 H1	8.8	9.9	9.8	12.6	9.9	12.6	12.9	13.2	9.9	12.4		
F1 H3	F1 H5	7.9	9.3	4.7	3.8	4.5	3.8	4.9	3.2	4.6	3.7		
F1 H4	F1 H5	10.7	11.0	7.2	6.3	7.1	6.1	7.3	6.2	5.4	6.3		
F2 H2	F2 H1	13.3	12.0	9.0	12.4	9.6	11.9	11.3	11.0	9.6	12.6		
F2 H3	F2 H5	8.4	9.3	3.0	3.2	4.0	3.3	5.0	4.5	4.1	3.2		
F2 H4	F2 H5	14.5	11.2	6.8	6.0	6.5	6.0	7.6	6.7	6.5	5.9		
2 H3	1 H1	4.4	8.9	5.2	6.6	6.8	6.5	0.2	0.4	6.3	6.5		
3 H3	2 H1	16.5	14.7	11.5	6.8	12.4	6.5	11.5	7.9	1.0	5.6		
4 H4	3 H1	12.7	14.9	0.3	8.9	10.9	10.8	0.3	0.6	0.4	0.6		
1 H2	F1 H1	8.5	9.7	11.3	6.7	12.0	6.8	1.8	5.1	12.1	6.7		
2 H2	F1 H5	9.3	10.0	5.5	10.2	9.2	10.3	0.0	0.1	5.2	10.3		
2 H4	F1 H5	2.0	1.5	0.2	0.1	0.2	0.0	0.0	0.0	0.5	0.1		
2 H3	F2 H1	0.6	0.4	0.3	0.0	0.1	0.0	15.9	14.1	0.1	0.0		
2 H4	F2 H1	9.8	8.5	0.6	8.4	5.0	8.0	0.3	0.4	4.6	8.1		
2 H5	F2 H1	1.7	1.4	1.0	0.2	0.3	0.1	8.1	2.1	0.2	0.2		
1 H2	F2 H5	8.4	10.3	2.1	7.1	7.8	6.9	0.0	0.0	7.8	6.9		
2 H3	F2 H5	1.1	0.9	0.2	0.5	0.6	0.5	0.0	0.0	0.6	0.5		
2 H4	F2 H5	2.4	1.9	9.5	1.4	3.3	1.5	0.5	0.3	3.3	1.4		
	RMSD		0.33	0.83	0.52	0.48	0.52	5.53	4.83	0.56	0.56		
	RMSD rank			3	1	1	1	4	4	2	3		

^a Similar to the case of LNF-1 in Table 2, the global energy minimum in an implicit solvent (conf. 1) does not have the best NOE values. Conf. 2 has the best NOEs but is ranked higher on an implicit solvent potential-energy scale. Except in the case of conf. 2, MD-averaged NOEs appear to be closer to experimental data.

By comparing each linkage of the best solution conformation for LND-1 with corresponding linkage for LNF-1, it is easy to see that ϕ – ψ values are similar, as displayed in Table 3. The extra linkage prevents certain configurations but otherwise preserves the oligosaccharide fold.

4. Conclusions

We have developed a sugar structure prediction tool based on a ring perception algorithm, automatic recognition of rotatable dihedrals, Euler rotations, implicit solvent minimizations, NOE calculations, and molecular dynamics in explicit solvent. We have also implemented a subtree recognition algorithm for finding an oligosaccharide fragment within a more complex molecule and a database for storing structural and rotameric information. Oligosaccharides are complex topological molecules with multiple possible branching points. Since dihedral rotations are strongly coupled, particularly in the case of adjacent linkages or when branching is present, the use of a simple rotameric library to study conformations of these systems is many times not feasible. Our database and subtree recognition algorithm overcomes this problem by storing all coupled ψ – ϕ regions of oligosaccharide fragments as vectorial quantities that can be queried when a larger sugar is presented to the program.

This automatic tool for sugar structure prediction was applied to the case of LNF-1 and LND-1, two related oligosaccharides present in milk. Our tool identified and pooled all important “unique conformations” for these oligosaccharides. The distribution of these unique conformations is much wider than previously reported from MD simulations. Structures appear to be clustered around distinct important regions. Previous MD studies show that even very long (50 ns) molecular dynamics studies do not reproduce the correct experimental NOEs unless initial conditions are

carefully chosen. This is because in a complex oligosaccharide the sampling of angular space is slow on a molecular dynamics time scale and the molecule remains trapped for very long times in local minima that do not necessarily correspond to the solution structure. Our algorithm overcomes this problem by brute-force sampling of the whole dihedral angular space. Since sugars are bulky and can be branched, this search does not exponentially explode and in fact is much faster than sampling allowed conformations with molecular dynamics or Monte Carlo techniques. Once a full dihedral space search is accomplished, structures are pooled by an implicit solvent minimization procedure. Simple NOE calculations and ranking against experimental data reveals in a very short time which of all allowed minimized conformations are most likely to exist in solution. Short-time MD simulations (5 ns) for different initial structures sampled according to our algorithm allow us to test whether these are stable in explicit solvent and provide a good strategy to obtain “local basin averaged” NOE values. In this article and in the first paper,⁴⁶ we have shown that other ranking criteria such as implicit solvent energies are poor estimators of the free-energy difference of oligosaccharides in solution. In general, the lowest-energy structure in implicit solvent is not the overall free-energy minimum in solution and does not correspond to a structure that has correct NOE values as compared to experimental data.

Our algorithm was successful in finding the best possible candidate structures in solution for LNF-1 and LND-1 in a very short time and without the need to resort to MD simulations. Our MD simulations confirm the fact that these structures are indeed stable in solution because when initial conditions were given in these regions of dihedral angular space we did not observe departure from the corresponding basin throughout our simulations. This is not the case for other

studied initial structures with linkages that although allowed did not match the experimental NOEs. An exception to this is terminal residues far apart from crowded linkages for which rotations are most likely decoupled from other glycosidic torsions.

An interesting question that arises in the case of oligosaccharides is whether as in the case of small proteins a clear fold exists. We find that these two sugars have particularly ordered structures. The additional branch of α -L-Fuc in LND-1 has some influence on the conformations of other linkages, but most allowed conformations for LNF-1 are also allowed in the case of LND-1. It is interesting that we have identified fewer allowed conformations in the case of the larger sugar than in that of the smaller one, pointing to the fact that branching and crowded linkages can indeed shrink the conformational space of larger sugars. The question of whether small sugars have well-defined folded structures in general is interesting and should be the focus of future studies.

Acknowledgment. This research was funded by Grant #05-2182 from the Roy J. Carver Charitable Trust awarded to C.J.M. and by the Skou Fellowship from the Danish Natural Sciences Research Council awarded to J.H.J.

References

- (1) Dwek, R. A. *Chem. Rev.* **1996**, *96*, 683–720.
- (2) Rudd, P. M.; Dwek, R. A. *Crit. Rev. Biochem. Mol. Biol.* **1997**, *32*, 1–100.
- (3) Cumming, D. A.; Carver, J. P. *Biochemistry* **1987**, *26*, 6664–6676.
- (4) Duus, J. Ø.; Gotfredsen, C. H.; Bock, K. *Chem. Rev.* **2000**, *100*, 4589–4614.
- (5) Wormald, M. R.; Petrescu, A. J.; Pao, Y.-L.; Glithero, A.; Elliott, T.; Dwek, R. A. *Chem. Rev.* **2002**, *102*, 371–386.
- (6) Martin-Pastor, M.; Bush, C. A. *Biochemistry* **1999**, *38*, 8045–8055.
- (7) Martin-Pastor, M.; Bush, C. A. *Carbohydr. Res.* **2000**, *323*, 147–155.
- (8) Cumming, D. A.; Carver, J. P. *Biochemistry* **1987**, *26*, 6676–6683.
- (9) Imberty, A.; Pérez, S. *Chem. Rev.* **2000**, *100*, 4567–4588.
- (10) Bush, C. A.; Martin-Pastor, M.; Imberty, A. *Annu. Rev. Biophys. Biomolec. Struct.* **1999**, *28*, 269–293.
- (11) Woods, R. *Glycoconjugate J.* **1998**, *15*, 209–216.
- (12) Woods, R. J. *Curr. Opin. Struct. Biol.* **1995**, *5*, 591–598.
- (13) Woods, R. J. The Application of Molecular Modeling Techniques to the Determination of Oligosaccharide Solution Conformations. In *Reviews of Computational Chemistry*; Lipkowitz, K., Boyd, D. B., Eds.; VCH Publishers: New York, 1996; Vol. 9, pp 129–165.
- (14) Kirschner, K. N.; Woods, R. J. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10541–10545.
- (15) Almond, A.; Petersen, B. O.; Duus, J. *Biochemistry* **2004**, *43*, 5853–5863.
- (16) Almond, A.; Sheehan, J. K. *Glycobiology* **2003**, *13*, 255–264.
- (17) Almond, A.; Bunkenborg, J.; Franch, T.; Gotfredsen, C. H.; Duus, J. Ø. *J. Am. Chem. Soc.* **2001**, *123*, 4792–4802.
- (18) Woods, R. J.; Pathiaseril, A.; Wormald, M. R.; Edge, C. J.; Dwek, R. A. *Eur. J. Biochem.* **1998**, *258*, 372–386.
- (19) Imberty, A.; Tran, V.; Pérez, S. *J. Comput. Chem.* **1989**, *11*, 205–216.
- (20) Imberty, A.; Gerber, S.; Tran, V.; Pérez, S. *Glycoconjugate J.* **1990**, *7*, 27–54.
- (21) Kocã, J. *THEOCHEM* **1994**, *308*, 13–24.
- (22) Kocã, J. *Prog. Biophys. Mol. Biol.* **1998**, *70*, 137–173.
- (23) Engelsena, S. B.; Kocab, J.; Braccinic, I.; Penhoatc, C. H.; Pérez, S. *Carbohydr. Res.* **1995**, *271*, 1–29.
- (24) Peters, T.; Meyer, B.; Stuike-Prill, R.; Somorjai, R.; Brisson, J.-R. *Carbohydr. Res.* **1993**, *238*, 49–73.
- (25) Nahmany, A.; Strino, F.; Rosen, J.; Kemp, G. J.; Nyholm, P.-G. *Carbohydr. Res.* **2005**, *340*, 1059–1064.
- (26) Strino, F.; Nahmany, A.; Rosen, J.; Kemp, G. J.; Sá-correia, I.; Nyholm, P.-G. *Carbohydr. Res.* **2005**, *340*, 1019–1024.
- (27) Newburg, D. S.; Ruiz-Palacios, G. M.; Morrow, A. L. *Annu. Rev. Nutr.* **2005**, *25*, 37–58.
- (28) Martin-Pastor, M.; Bush, C. A. *Biochemistry* **2000**, *39*, 4674–4683.
- (29) Landerjö, C.; Jansson, J. L. M.; Maliniak, A.; Widmalm, G. *J. Phys. Chem. B* **2005**, *109*, 17320–17326.
- (30) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.
- (31) Ponder, J. W. F. M. R. *J. Comput. Chem.* **1987**, *8*, 1016–1024.
- (32) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.
- (33) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824–19839.
- (34) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246*, 122–129.
- (35) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
- (36) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Biol.* **2001**, *7*, 306–317.
- (37) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *117*, 11225–11236.
- (38) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981.
- (39) Nosé, S. *Mol. Phys.* **1984**, *52*, 255–268.
- (40) Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695–1697.
- (41) Berendsen, H. J. C.; Postma, J. P. M.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

- (42) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4546–4559.
- (43) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4559–4570.
- (44) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (45) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (46) Xia, J.; Daly, R. P.; Chuang, F.-C.; Parker, L.; Jensen, J. H.; Margulis, C. J. update: **2007**, *4*, 1620–1628.

CT700034Q