JCTC Journal of Chemical Theory and Computation

# Replica-Exchange Accelerated Molecular Dynamics (REXAMD) Applied to Thermodynamic Integration

Mikolai Fajer,*[,†,‡] Donald Hamelberg,[§] and J. Andrew McCammon[†,‡,||,⊥]

*Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, California 92093-0365, Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, California 92039-0365, Department of Chemistry, Georgia State University, Atlanta, Georgia 30302-4098, Department of Pharmacology, University of California at San Diego, La Jolla, California 92093-0365, and Howard Hughes Medical Institute, University of California at San Diego, La Jolla, California 92093-0365*

Received June 26, 2008

**Abstract:** Accelerated molecular dynamics (AMD) is an efficient strategy for accelerating the sampling of molecular dynamics simulations, and observable quantities such as free energies derived on the biased AMD potential can be reweighted to yield results consistent with the original, unmodified potential. In conventional AMD the reweighting procedure has an inherent statistical problem in systems with large acceleration, where the points with the largest biases will dominate the reweighted result and reduce the effective number of data points. We propose a replica exchange of various degrees of acceleration (REXAMD) to retain good statistics while achieving enhanced sampling. The REXAMD method is validated and benchmarked on two simple gas-phase model systems, and two different strategies for computing reweighted averages over a simulation are compared.

## Introduction

Free energy is one of the most important quantities in biophysics. The calculation of free energy using molecular dynamics simulations is complicated by the dependence on the amount of the relevant phase space sampled. The complication is more pronounced when two alchemical free energy end points differ by more than a few trivial moieties. The use of restraints to restrict the phase space has proven useful in the convergence of thermodynamic integration, umbrella sampling, and the Bennett acceptance ratio techniques.[1−3] Another approach is to enhance phase space sampling instead of restricting the phase space and often relies on the modification of the original Hamiltonian during molecular dynamics simulations.[4,5] Accelerated molecular dynamics (AMD), which conventionally modifies the energy landscape by adding a bias to states below an energy threshold, $E_{cut}$ (eq 1), is an example of the Hamiltonian modification approach and has proven capable of efficiently generating canonical ensembles consistent with experiments on the millisecond time scale.[6,7]

$$V^*(r, E_{cut}, \alpha) = V(r) + \begin{cases} 0 & V(r) \geq E_{cut} \\ \Delta V(r, E_{cut}, \alpha) & V(r) < E_{cut} \end{cases}$$

$$\Delta V(r, E_{cut}, \alpha) = \frac{(E_{cut} - V(r))^2}{\alpha + (E_{cut} - V(r))} \quad (1)$$

A potential problem with modifying the Hamiltonian occurs when reweighting an observable $O^*$ from the accelerated simulation to find $O$ on the original potential (eq 2 for AMD). If the simulation is highly accelerated and involves a large range of boost factors $\Delta V$, the reweighted average will be dominated by the relatively few points/structures with large values of $\Delta V$ in the limit of finite sampling. This statistical problem has recently been quantified as a reduction in the effective number of data points in the simulation.[8] Thus there is a tradeoff between the degree of acceleration and the statistical precision in AMD simulations. The calculation of free energies using thermodynamic integration computes $\langle dV/d\lambda \rangle_\lambda$ over the course of a simulation, and the calculation of free energy is very sensitive to the statistical accuracy of the computed averages.

$$\langle O \rangle = \frac{\langle O^* \exp[\beta \Delta V(r)] \rangle}{\langle \exp[\beta \Delta V(r)] \rangle} \quad (2)$$

In order to take advantage of the sampling efficiency of the AMD method as well as maintain the statistical relevance of every data point, we propose using a replica-exchange framework to couple varying degrees of acceleration. The low degrees of acceleration will not be prone to the reweighting problem and can still take advantage of the high acceleration through replica exchanges. This replica-exchange accelerated molecular dynamics (REXAMD) is a member of the Hamiltonian replica-exchange (HREM) class of simulations, varying from other

* Corresponding author e-mail: mfajer@gmail.com.

† Department of Chemistry and Biochemistry, University of California at San Diego.

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

§ Georgia State University.

|| Department of Pharmacology, University of California at San Diego.

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

10.1021/ct800250m CCC: $40.75 
Published on Web 09/03/2008

HREM techniques in the specific Hamiltonian modification scheme. A similar REXAMD approach has recently been applied to studying the effects of neighboring side chains on peptide backbone conformations in short peptides.[9] We demonstrate the REXAMD approach by increasing the convergence rate of thermodynamic integration (TI) for two simple gas-phase model systems, although the method could utilize other free energy calculation methods instead of TI.

## Computational Detail

First some terms should be defined. "State" is used to denote a specific level in the replica-exchange scheme. For example, in temperature replica-exchange each state corresponds to a specific temperature, and in REXAMD each state is a modified Hamiltonian described by a set of boost parameters. The term "replica" is used to denote the individual structures that are exchanged between the various REXAMD states. The term "simulation" refers to a specific setup of REXAMD, and the term "run" refers to an instance of a simulation. Simulation is also used to identify the average and standard error computed from multiple runs.

The current replica-exchange framework is a Python program that launches a modified AMBER8 accelerated molecular dynamics simulation[6] for each replica in between Metropolis Monte Carlo exchanges (eq 3). The Monte Carlo (MC) exchanges occur every 1000 molecular dynamics (MD) steps, and the pairs that attempt exchanges alternate every other MC period. For example, in a simulation with four states (labeled s0-s3) the simulation would execute 1000 MD steps, attempt MC exchanges between the s0-s1 states and the s2-s3 states, execute 1000 MD steps, attempt a MC exchange between the s1-s2 states, and repeat. The molecular dynamics simulations used a 1 $fs^{-1}$ time step and were coupled to a 300 K Langevin thermostat with a collision frequency of 10.0 $ps^{-1}$. The Python program reset the seed number for the AMBER random number generator after every MC exchange.

$$p_{ex}(i,j) = \begin{cases} 1 & \Delta(i,j) \le 0 \\ \exp[-\beta\Delta(i,j)] & \Delta(i,j) > 0 \end{cases}$$

$$\Delta(i,j) = \Delta V(r_j, \alpha_i) + \Delta V(r_i, \alpha_j) - \Delta V(r_i, \alpha_i) - \Delta V(r_j, \alpha_j) \quad (3)$$

The boosting scheme is identified as a suffix added to the REXAMD acronym as follows: REXAMDt denotes a boost only to the torsional potential, and REXAMDtT denotes a dual boost scheme applied to the torsional and total potentials.[10] The "-rw" suffix indicates the reported results are from the reweighting of the most accelerated state in a specific simulation. When the "-rw" suffix is not present, the result is coming from the least accelerated state, which in this paper is always no acceleration.

In order to separate the effect of acceleration from the effect of using M replicas, the REXREG control simulations are a replica-exchange between identical regular dynamics potentials. Note that this makes the acceptance probability of MC exchange in eq 3 identically equal to one. The REXREG simulations are analogous to M independent runs from the same starting point with different initial velocities and taking an average result from the M runs.

The replica-exchange efficiency will be monitored based on two criteria. The first criterion is the average acceptance ratio of the replica-exchanges over the course of a run and gives a rough idea of how capable the given replica-exchange scheme is at mixing replicas. The second criterion is the observed relative frequency rmsd metric.[11] This metric compares the observed population frequency of the replicas against the idealized case where each of M replicas spends $1/M$ of the total time in any given state of the system. The rmsd metric varies from zero for the ideal mixing to $\sqrt{M-1}/M$ for no mixing. The observed relative frequency metric is more detailed than the average acceptance ratio in monitoring the mixing efficiency of the replica-exchange simulation.

The thermodynamic integration of the model systems was computed using a linear scaling of an all-atom potential (eq 4). Gaussian quadrature integration was used to evaluate the thermodynamic integral from a finite number of $\langle dV/d\lambda \rangle_\lambda$ calculated at specific $\lambda$ values (eq 5). The Gaussian quadrature points and weights were taken from the AMBER8 manual.[12] Two strategies were used to calculate $\langle dV/d\lambda \rangle_\lambda$ at each $\lambda$. The first strategy, *reweighted periods*, calculated the reweighted average of each block of 1000 MD steps in between MC exchanges. These reweighted averages were then averaged together over a complete run to yield $\langle dV/d\lambda \rangle_\lambda$ for a specific $\lambda$. The assumption behind this approach is that the $dV/d\lambda$ values sampled during 1 ps give rise to $\langle dV/d\lambda \rangle_\lambda$ for a local region of the conformational space. This strategy becomes exact when the period is longer than the potential energy correlation time of the system. The replica exchange will then balance the occurrence of the local regions. The second strategy, *reweighted run*, takes an instantaneous $dV/d\lambda$ and its corresponding $\Delta V$ from the MD step immediately prior to a MC exchange. These values are then used to compute a reweighted $\langle dV/d\lambda \rangle_\lambda$ for the entire simulation. This approach virtually guarantees uncorrelated $dV/d\lambda$ values at the expense of the number of points being considered in the average. In both strategies each $\lambda$ was simulated ten times with different random seeds and velocities. An average and standard error for each $\langle dV/d\lambda \rangle_\lambda$ is then determined and combined into the overall $\Delta G$. The average $\Delta G$ is only reported to the first significant digit of the standard error.

$$V(\lambda) = (1-\lambda)V_0 + \lambda V_1 \quad (4)$$

$$\Delta G = \int_0^1 \langle dV/d\lambda \rangle_\lambda d\lambda \approx \sum_i w_i \langle dV/d\lambda \rangle_i \quad (5)$$

Two model systems were studied to validate and benchmark the REXAMD method. Both model systems are symmetric alchemical mutations where the product has an identical structure to the reactant, and thus the $\Delta G$ is zero and independent of the force field. Model system A (MSA) is a gas-phase alchemical mutation from ethane-to-ethane (Figure 1A). This system will serve as a positive control to show that REXAMD can reproduce the results of an ergodic regular molecular dynamics simulation. The relative simplicity of the system and the low transition barriers guarantees that the regular molecular dynamics (REXREG) is able to sample the entire conformational space in a short time scale. The thermodynamic integration for MSA uses a 9-point Gaussian quadrature. The MSA REXAMDt simulations used only two replicas: an unmodified potential and an accelerated potential with a torsional boost ($E_{cut}$ of 5.0 kcal

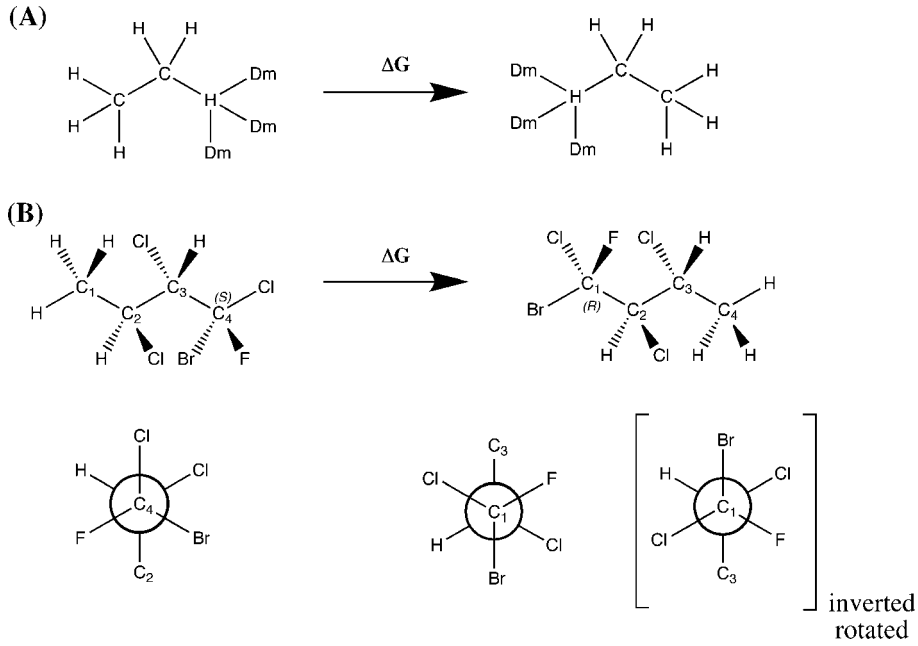



***Figure 1.*** Structure of the model systems (A) and (B). The Dm atoms indicate a dummy atom with no nonbonded interactions.

mol$^{-1}$, α of 2.0 kcal mol$^{-1}$). Each run was simulated for 8 million MD steps or the equivalent of 8 ns for an unmodified potential.

Model System B (MSB) is a highly halogenated butane (Figure 1B). The initial conformation of the system is in a different rotameric state for the two $\lambda$ end points, as seen in the Newman projections in Figure 1B, and thus requires proper conformational sampling to yield the correct $\Delta G$. The chlorine atoms attached to $C_2$ and $C_3$ were added to make the rotameric sampling more difficult, requiring acceleration to achieve the correct answer within the current time scale of 20 ns. The dual boosting scheme was used for this model system in order to accelerate the large van der Waals interactions experienced in this system. In order to increase the difficulty of converging to the correct result, we are only using a 3-point Gaussian quadrature. The boost parameters for the eight replicas in the MSB REXAMDtT simulations are shown in Table S-I (Supporting Information) and are labeled from s0 to s7 in terms of increasing boost.

## Results and Discussion

**Model System A.** In MSA both the REXREG and REXAMDt simulations were able to efficiently and exhaustively explore the conformational space (data not shown), and the replica mixing was quite efficient (Table 1) within the 8 ns runs. The exhaustive sampling resulted in converged $\Delta G$ values within the first ns of the REXREG and REXAMDt simulations (Figure 2). The $\Delta G$ results from the entire 8 ns are summarized in Table 2. Recall that "MSA REXAMDt" refers to results taken from the nonaccelerated state and "MSA REXAMDt-rw" refers to the reweighted results of the accelerated state.

The statistical precision can be monitored in terms of the number of values that were used in computing $\langle dV/d\lambda \rangle_\lambda$. For example, applying the reweighted run strategy to the REXAMDt simulation yields a total of 80,000 data points for each $\langle dV/d\lambda \rangle_\lambda$ (ten 8 ns trajectories). This strategy resulted in a $\Delta G$ of

***Table 1.*** Summary of the Replica-Exchange Efficiency

| simulation | acceptance ratio[a] | observed relative frequency rmsd |
|---|---|---|
| MSA REXAMDt (8 ns, 2 states) | 39.3 ± 2.2% | 0.00565 ± 0.00420 |
| MSB REXAMDtT (20 ns, 8 states) | 40.2 ± 16.0% | 0.00827 ± 0.00231 |

[a] The average and standard deviation of the acceptance ratios are from the ten runs and the M states. The average and standard deviation of the rmsd of the relative occupancy of the M replicas over the M states, as defined by Abraham et al.,[11] are reported.

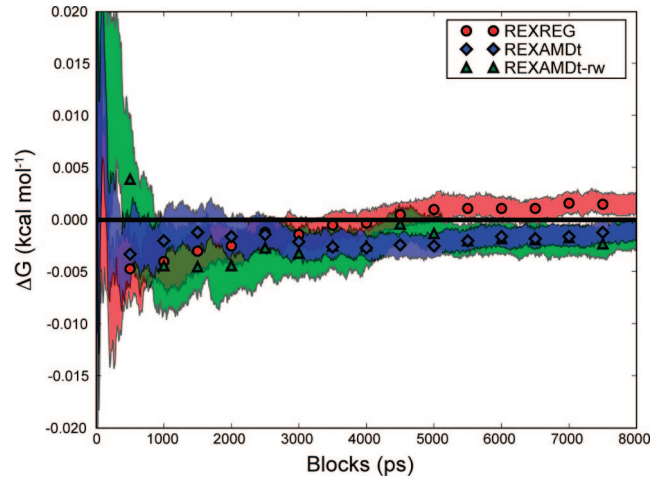



***Figure 2.*** Block average of the MSA thermodynamic integration results when using the reweighted periods strategy. The symbols show the average value of each simulation type, and the shaded region shows the standard error from the ten duplicate runs.

0.02 ± 0.02 kcal mol$^{-1}$. In order to produce the same number of points when using the reweighted periods strategy we consider only the first 8 ps of the ten duplicate runs for each $\lambda$, which yields a $\Delta G$ of 0.02 ± 0.03 kcal mol$^{-1}$. Note the

***Table 2.*** $\Delta G$ Summary of MSA Thermodynamic Integration Results[a]

| reweighting strategy | REXREG | REXAMDt | REXAMDt-rw |
|---|---|---|---|
| periods | $+0.002 \pm 0.001$ | $-0.001 \pm 0.001$ | $-0.001 \pm 0.001$ |
| runs | $-0.04 \pm 0.01$ | $+0.02 \pm 0.02$ | $-0.01 \pm 0.03$ |

[a] The units are in kcal mol$^{-1}$. The average and standard error from ten simulations are reported for each simulation type.

***Table 3.*** $\Delta G$ Summary of MSB Thermodynamic Integration Results[a]

| reweighting strategy | REXREG | REXAMDtT | REXAMDtT-rw |
|---|---|---|---|
| periods | $+0.12 \pm 0.08$ | $+0.04 \pm 0.01$ | $+0.08 \pm 0.06$ |
| runs | $+0.16 \pm 0.07$ | $+0.03 \pm 0.04$ | $-9 \pm 7$ |

[a] Units are in kcal mol$^{-1}$. The average and standard error from ten simulations are reported for each simulation type.

similarity in both the accuracy and precision of these two results, indicating that exhaustive sampling occurs below the picosecond time scale. The slower $\Delta G$ convergence of the reweighted run strategy versus the reweighted periods strategy is due to the slower rate of data collection for the reweighted run strategy.

The REXAMDt-rw simulations also exhibit high accuracy and precision (Figure 2 and Table 2). The average boost applied over the MSA REXAMDt simulations from all of the $\lambda$ values was $2.0 \pm 0.9$ kcal mol$^{-1}$. The small range of boosts (standard deviation of 0.9 kcal mol$^{-1}$) is predicted to have a relatively small effect on the reweighted precision as predicted by Shen and Hamelberg.[8] The reweighted periods strategy reduces the effective number of instantaneous $dV/d\lambda$ values from 80 million to 16 million for each $\langle dV/d\lambda \rangle_\lambda$, and the REXAMDt-rw simulations exhibit marginally worse precision than the REXAMDt (Table 2). A similar effect is observed in the reweighted runs strategy (a reduction from 80,000 to approximately 15,000).

**Model System B.** The 20 ns MSB REXAMDtT simulations are well mixed (Table 1, Figures S-I and S-II (Supporting Information)). The regular molecular dynamics (REXREG) was unable to efficiently sample the conformational space (Figure S-III in the Supporting Information) and still shows a substantially nonzero $\Delta G$ after the 20 ns for both the reweighted periods and reweighted runs strategies (Table 3). The slow convergence of the REXREG result can also be seen in the block averaging of $\Delta G$ in Figure 3. In contrast, the REXAMDtT simulations were able to efficiently sample the conformational space Figure S-IV in the Supporting Information. The $\Delta G$ was consistently within 0.1 kcal mol$^{-1}$ of zero after 2.9 and 5.5 ns for the reweighted periods strategy and the reweighted runs strategy, respectively.

The reweighting procedure was applied to the state with the highest degree of acceleration, s7, because this state is the most independent of the other states in terms of convergence. The most accelerated state is also expected to have the highest range of $\Delta V$ boost factors and therefore exhibit the largest reweighting problem.[8] This prediction can be seen in the poor accuracy and precision of the $\Delta G$ of reweighted runs for REXAMDtT-rw (Table 3, Figure 4). The effective numbers of data points for the s7 states are shown in Table S-I (Supporting Information) and demonstrate the source of the poor statistics. For example,

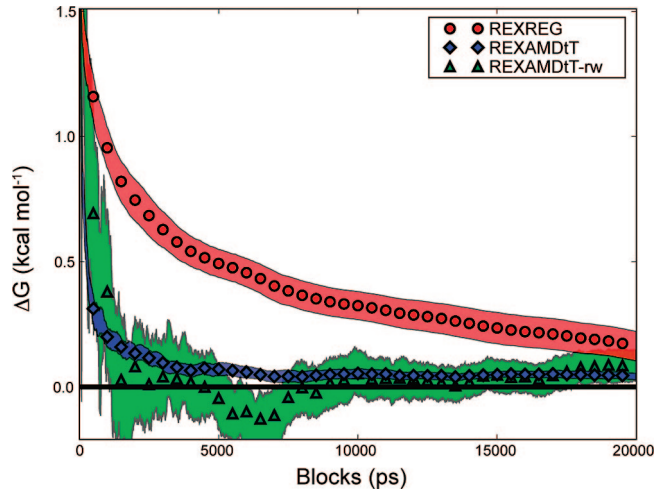



***Figure 3.*** Block average of the MSB thermodynamic integration results from the reweighted periods strategy. The symbols show the average value of each simulation type, and the shaded region shows the standard error for each simulation type.

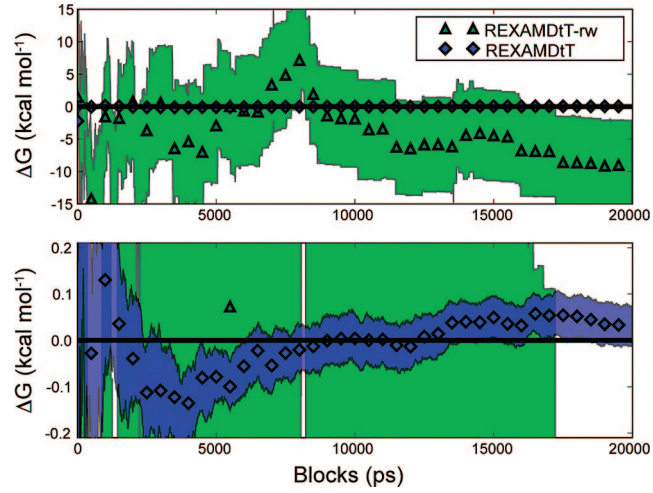



***Figure 4.*** Block average of the MSB thermodynamic integration results from the reweighted runs strategy shown on two different scales. The symbols show the average value of each simulation type, and the shaded region shows the standard error for each simulation type. The top plot shows the REXAMDtT-rw results on scale and shows how poor the statistics are after reweighting. The bottom plot shows the REXAMDtT results on scale and shows how quickly the REXAMD technique converges to within statistical accuracy.

the $\lambda$ of 0.5 simulations had a standard deviation of boost values of 13 kcal/mol, and only 30 of the 200,000 data points from the ten duplicate runs contributed to $\langle dV/d\lambda \rangle_{\lambda=0.5}$.

The reweighted periods strategy for REXAMDtT-rw has at least one effective point in each 1 ps period and therefore at least 200,000 data points for each $\langle dV/d\lambda \rangle_\lambda$ when the ten duplicate runs are considered. Compared to the reweighted runs strategy, the increase of the effective number of points results in the increase of the accuracy and precision of the computed $\Delta G$ by 2 orders of magnitude (Table 3). The effective number of points is still less than that of REXAMDtT, which has 200 million data points, and the accuracy and precision of REX-

AMDtT are still better than those of REXAMDtT-rw when using the same averaging strategy (Table 3, Figure 3).

## Conclusion

The REXAMD method has been shown to efficiently accelerate conformational sampling while avoiding the statistical reweighting problem inherent in AMD. The REXAMD method was validated on the simple model system A. In the more complex model system B the dual boost REXAMD scheme showed marked improvement over the regular molecular dynamics approach as well as better statistical accuracy and precision in comparison to the reweighted results of the accelerated replicas. We are currently researching the application of this method to more complicated systems.

**Acknowledgment**. We would like to acknowledge Dr. Xiaolin Cheng for insightful discussions and Robert Swift, Arneh Babakhani, and Morgan Lawrenz for manuscript editing. M.F. was supported in part by an NIH Molecular Biophysics Training Grant (GM-08326) and subsequently by an NSF Graduate Research Fellowship. Funding by NIH GM31749, NSF MCB-0506593, and MCA93S013 (to J.A.M.) also supports this work. Additional support from the Howard Hughes Medical Institute, San Diego Supercomputing Center, the National Biomedical Computational Resource, and the Center for Theoretical Biological Physics is gratefully acknowledged.

**Supporting Information Available:** Additional figures and tables with more detailed information about the replica-exchange efficiency and acceleration parameters. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C. D.; Shirts, M. R.; Sorin, E. J.; Pande, V. S. *J. Chem. Phys.* **2005**, *123*, 084108.

(2) Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2004**, *126*, 7683–7689.

(3) Wang, J.; Deng, Y.; Roux, B. *Biophys. J.* **2006**, *91*, 2798–2814.

(4) Li, H.; Fajer, M.; Yang, W. *J. Chem. Phys.* **2007**, *126*, 024106.

(5) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13703–13710.

(6) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919–11929.

(7) Markwick, P. R. L.; Bouvignies, G.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 4724–4730.

(8) Shen, T.; Hamelberg, D. *J. Chem. Phys.* **2008**, *129*, 034103.

(9) Xu, C.; Wang, J.; Liu, H. *J. Chem. Theory Comput.* **2008**, *4*, 1348–1359.

(10) Hamelberg, D.; de Oliveira, C. A. F.; McCammon, J. A. *J. Chem. Phys.* **2007**, *127*, 155102.

(11) Abraham, M. J.; Gready, J. E. *J. Chem. Theory Comput.* **2008**, *4*, 1119–1128.

(12) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. *Comput. Phys. Commun.* **1995**, *91*, 1–41.

CT800250M

JCTC Journal of Chemical Theory and Computation

# Toward Large Scale Parallelization for Molecular Dynamics of Small Chemical Systems: A Combined Parallel Tempering and Domain Decomposition Approach

Henk A. Slim and Mark R. Wilson*

*Department of Chemistry, University of Durham, South Road, Durham, DH1 3LE, U.K.*

Received June 27, 2008

**Abstract:** A combined parallel tempering (replica exchange) and domain decomposition approach is presented, which allows for the effective use of large numbers of processor cores (>256) on modest sized simulations of chemical systems (~5000 sites). The approach is implemented in the **gbmoldd** molecular dynamics program for the simulation of coarse-grained molecular systems composed of combinations of isotropic and/or anisotropic particles. Benchmark results are presented for two test systems: a $C_{24}$ united atom chain and a coarse-grained system of spherocylinders.

## 1. Introduction

In recent years, parallel tempering has become a powerful technique to improve sampling in molecular simulation. It has already shown to be useful in the simulation of a range of systems, including improved conformational sampling in peptides,[1,2] proteins,[3] and polymers,[4,5] studies of phase transitions in water clusters,[6] simulations of lattice models,[7,8] and in biased Monte Carlo (MC) for crystal structure determination.[9] A useful review of parallel tempering methods, with particular reference to many molecular simulation problems, has been produced recently by Earl and Deem.[10]

The key to the success of parallel tempering methods is a good overlap of the configurational energy between successive replica systems. Without this, many attempted relica swaps will be rejected. For many molecular applications, a useful guide is that approximately 20−25% of moves should be accepted. This figure can be arrived at by using an iterative scheme for the optimal allocation of temperatures[11] or by tuning temperature intervals to maximize the mean squared displacement of a system.[12] However, as system size grows, the ratio of the width of the configurational energy distribution over the number of particles, $N$, effectively becomes more sharply peaked. This is because the width of the energy distribution increases as $\sqrt{N}$, but the average energy increases as $N^9$. This means that as the system size increases, the number of replicas must grow as $\sqrt{N}$ to keep comparable sampling over similar temperature ranges.

While replica exchange favors a small system size, parallel molecular dynamics (MD) simulation favors large systems. In a typical MD simulation, as the number of sites increases, the CPU time required for computing the pair interactions increases more quickly than the communication time required between processors. However, while parallelization is often very efficient when $N > 50\,000$, for small systems of a few thousand particles, communication costs can easily win out. This is a particularly serious limitation in replicated data algorithms,[13,14] which typically require a global sum (all reduce) operation to sum and distribute forces over all processing cores on every time step. However, even for more efficient domain decomposition strategies,[13,14] where communication is limited to a minimum and global sum operations are avoided, parallelization for a few thousand particles quickly leads to the algorithm becoming communications bound. Moreover, in a normal domain decomposition for a few thousand particles, the number of parallel subdomains physically possible is severely restricted by the size of the simulation box. While parallelization is not necesssary for some small systems, increasingly there is a desire to carry out molecular simulation to observe events that normally occur rarely (barrier crossing, crystallization, structural transformations, and transitions). Coupled with the relative cheapness and high availability of arrays of com-

* Author for correspondence. Tel.: +44 191 334 2144. Fax: +44 191 386 1127. E-mail: mark.wilson@durham.ac.uk.

10.1021/ct800255r CCC: $40.75 
Published on Web 09/19/2008

modity processors, it would be extremely useful to be able to apply large parallel compute arrays to study relatively small systems.

The current work presents a simple simulation approach, which combines parallel domain decomposition with parallel tempering methods into a single efficient code, **gbmoldd**, which is suitable for the study of small molecular systems composed of atomistic or coarse-grained potentials. The combination of $r$ parallel tempering replicas and an efficient domain decomposition of a small system over $N_p$ processors cores, means that $rN_p$ commodity processor cores can be concentrated on the simulation of a small chemical system. Typical values of $N_p = 8$ ($2 \times 2 \times 2$ domain decomposition) and $r = 32$, which are often appropriate for a few thousand sites, already provide for the possibility of using 256 processors efficiently. For systems where domain decomposition can be extended to a $4 \times 4 \times 4$ box, it is already possible to exceed the number of processors which are available on all but the most specialist parallel machines.

## 2. Computational Method

The program **gbmoldd** has been designed as a simulation code to be used primarily for coarse-grained simulations containing mixtures of anisotropic and isotropic coarse-grained sites within molecular systems, e.g. in liquid crystalline and/or macromolecular systems,[15−17] but it can also be used for modeling of atomistic systems based on Lennard-Jones sites. Message passing in **gbmoldd** is implemented with the message passing interface (MPI) communications protocol. The basic processor arrangement for the algorithm is shown in Figure 1. At the start of the molecular dynamics run, $N_t$ identical copies of the program are started on $N_t$ identical CPU cores using a MIMD parallel approach.[18] Here, for a system of $r$ replicas, $N_t = r \times N_p = r \times N_{px} \times N_{py} \times N_{pz}$, and an initial domain decomposition over $N_p$ processor cores is made for $x$, $y$, and $z$ spatial directions. In practice, the processor cores are grouped into groups of size $N_p$ (ideally a group of processors should be physical neighbors with the fastest available interconnects available between group members) and the processor map is distributed to each processor in turn. All MPI communication within the processor group of each replica takes place in the context of MPI intracommunicators; the exchange of information between replicas is facilitated with MPI intercommunicators.

An outline of the algorithm used in this work is shown in Figure 2. The main part of the algorithm is dominated by a domain decomposition approach carried out for each replica in parallel. This makes use of a division of each replica domain into subdomains, and each subdomain into cells which are equal to (or slightly larger than) the short-range cutoff. Particles are initially sorted into the subdomains, and a single CPU core is responsible for the coordinates, velocities, orientations, and orientational derivatives of the particles within its subdomain. This has the usual advantage that adjacent processors only require information about particles in their own subdomain and in cells on the boundary with adjacent nodes. In domain decomposition, there are two main choices of how to carry out the core message passing.[14,19] An initial communication of the subdomain

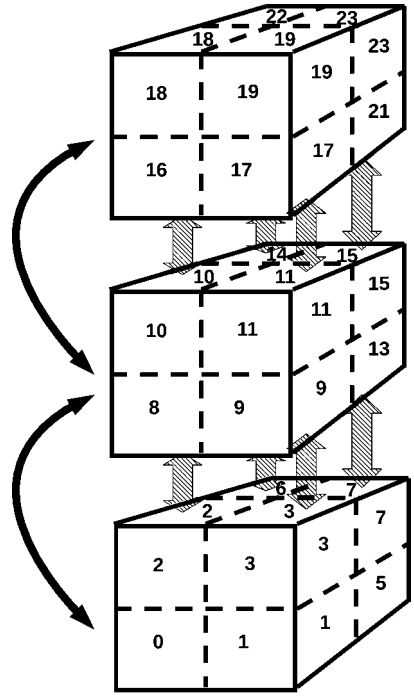



***Figure 1.*** Schematic diagram showing the domain decomposition approach used in this work. MC parallel tempering moves take place between separate parallel systems as indicated by the solid arrows on the left of the diagram. Each separate replica system uses a domain decomposition approach. The shaded arrows between domains correspond to data transfers, which in this case involve a transfer of ensemble averages (rather than coordinates) from each domain.

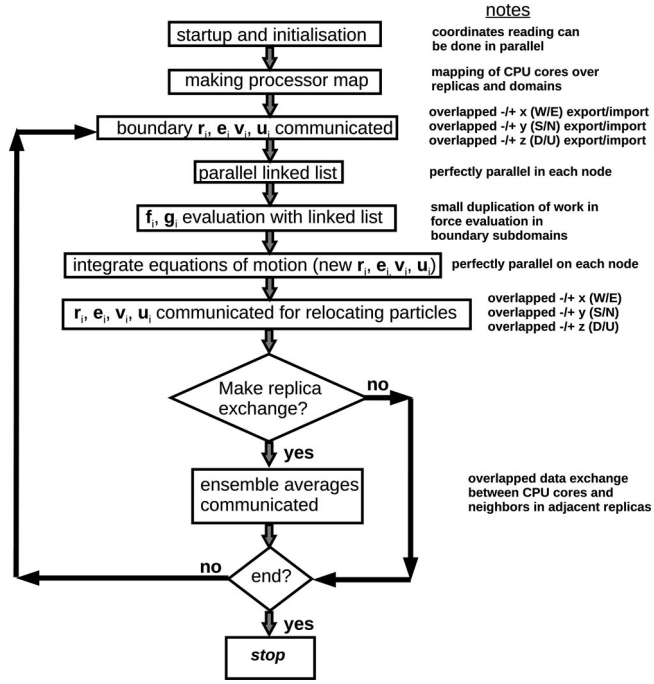



***Figure 2.*** Flowchart diagram showing the structure of the combined parallel tempering−domain decomposition program **gbmoldd**.

coordinates can be carried out, with message passing for the positive and negative directions executed at the same time (which is particularly useful in cases where processors have

more than one network connection). Then if force calculations are carried out separately on each node, there is no need to carry out message passing for the forces (as shown in Figure 2). This leads to some duplication of the force calculation for particles in the boundary subdomains but helps reduce communication costs.[14,19] The alternative involves passing coordinate information for the boundary subdomains, in one direction only, prior to the force calculation, and then passing the force information back in the opposite direction. This approach avoids the duplication of force calculations for boundary particles but involves slightly more message passing. For **gbmoldd**, we adopt the former approach (Figure 2) to reduce communication, because anisotropic particles already involve additional communication costs associated with the transfer of particle orientations as well as positions.

Although different cutoffs can be used for different types of interaction sites, in **gbmoldd**, the subdomains are always set up to correspond to the largest cutoff in the system (usually between two anisotropic particles) to avoid multiple layers of message passing for different types of sites. For molecular systems involving multiple bonded sites, for the force decomposition to be successful, all 1−2, 1−3, and 1−4 based potentials must have a maximum interaction distance, which is less than or equal to the width of a subdomain. This is always the case for atomistic systems but is usually also the case for most coarse-grained systems. Within the simple (nonreallocation) force scheme described above, when atoms taking place in bonded potentials are split across subdomains, the potential is evaluated in each subdomain separately to obtain the correct forces but the energy is only counted once. This tends to be a very minor cost in terms of total CPU time.

The rest of the molecular dynamics algorithm follows a conventional domain decomposition strategy. For anisotropic particles, we use an anisotropic form of the Velocity Verlet algorithm[20,21] (an alternative anisotropic leapfrog algorithm is also available). For anisotropic particles, the simulation keeps track of the position vectors, $\mathbf{r}(t)$, orientations, $\mathbf{e}(t)$, velocities, $\mathbf{v}(t)$, and orientational derivatives, $\mathbf{u}(t)$, and the total kinetic energy is given by

$$K = \sum_{i=1}^{N} \frac{m_i v_i^2}{2} + \sum_{i=1}^{N} \frac{I_i u_i^2}{2}$$

for masses, $m_i$, and moments of inertia, $I_i$. The integrator used forces, $\mathbf{f}(t)$, and gorques, $\mathbf{g}(t)$, and integration uses the following steps, which make use of a variable, $\xi$ (with $\dot{\eta} = \xi$), to control the thermostat and are executed in parallel making use of a central difference Stoermer−Verlet integrator[22] (NB: it is also possible to carry out the integration in other ensembles but the best results for replica exchange are obtained for constant *NVT*):

1. Evaluate $\mathbf{f}(t)$, $\mathbf{g}(t)$, and virial.
2. Advance $\mathbf{v}(t)$ and $\mathbf{u}(t)$ by a half-step time-step.
3. Evaluate kinetic energy, $K$.
4. Advance $\mathbf{r}(t)$, $\mathbf{v}(t)$, $\xi(t)$, and $\eta(t)$ by a time-step (using half-step quantities from above).
5. Reallocation (see below).
6. Advance $\mathbf{v}(t)$ and $\mathbf{u}(t)$ by a second half-step time-step.

At the end of the positions/orientations integration step, particles which leave a subdomain are reallocated to neighboring processors. As usual in domain decomposition, both the initial coordinate pass (prior to the force calculation) and the particle reallocation have to take place by passing in the $x$, $y$, and then $z$ directions in turn, with a coordinate sort in between to ensure that nodes connected via edges or vertices (as well as faces) are able to interchange particles.[19]

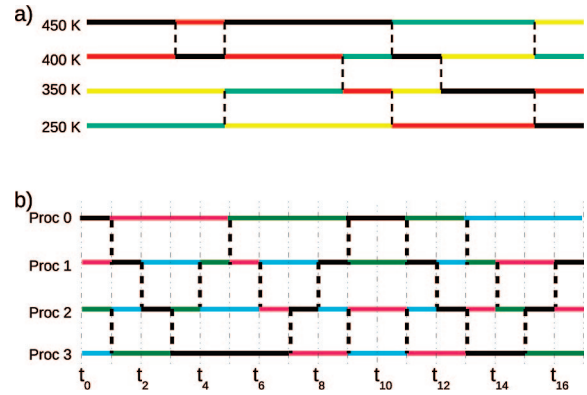



***Figure 3.*** Schematic diagram showing replica exchange approaches. (a) Traditional approach. Coordinates are exchanged between replicas with each processor handling a single temperature with the color coding indicating the "time evolution" of individual coordinate sets. (b) Approach used in this work. Temperatures are exchanged between processors with the color coding indicating the path of temperatures through the processors; ensemble averages must be calculated over individual temperatures by data exchange.

At the end of $N_{\text{steps}}$ molecular dynamics steps, the system attempts to undertake a "parallel tempering swap". Here, a swap between two replicas $i$ and $j$ is accepted with the probability

$$\min\{1, \exp[+(\beta_i - \beta_j)(U_i - U_j)]\}$$

where $U_i$ and $U_j$ are the potential energies of each separate replica. In practice, this also involves a scaling of the velocities to ensure that the average kinetic energy obeys equipartition.[23]

As with the communications in domain decompostion, there are two alternative approaches for communications in replica exchange (Figure 3). Coordinates can be swapped between replicas or the ensemble conditions of the replica can be swapped. The former involves a coordinate swap for each accepted move, the latter involves a swap of ensemble information, i.e. in the case of temperature parallel tempering, a swap of the temperatures between replicas takes place. To avoid swapping coordinates, we have implemented the latter. The advantage of doing this is that the communication costs are greatly reduced by avoiding coordinate swaps. However, additional care must be taken in computing thermodynamic averages and other simulation output data. Unlike a normal simulation (and unlike the situation where coordinate swapping takes place), these can not be computed directly by each processor separately and written at the end of the simulation because the average must be made over the individual ensembles. Each replica therefore has to swap any thermodynamic (or other) averages being calculated during the simulation run, at the same time as the temperature is

swapped. To do this efficiently, each subdomain within a replica needs to swap information with its corresponding subdomain in the other replica. This pairwise swap can occur simultaneously across all swapping domains (Figure 1) using a single send and receive operation on each processor involved in a swap. In cases where simulation data (e.g., coordinates and velocities for postprocessing) are written to files, care must also be taken to ensure that the correct coordinate data belonging to a single temperature is used. In practice, this can be carried out most efficiently in a postprocessing step. Each separate CPU simply writes its own coordinate/velocity data independently to a separate file, with each set tagged by simulation step and temperature flag. Postprocessing information, such as the velocity autocorrelation for each temperature, is then calculated by a single program, which sorts the multiple data files combining data from different domains belonging to the same temperature.

It should be noted that it is also possible to swap the potential used for molecular dynamics within this approach (Hamiltonian replica exchange).[24] Doing so allows for potential softening approaches, whereby one can soften selected parts of the Hamiltonian. Here, each parallel copy of the program running on each processor core contains an alternative set of potentials. Message passing simply involves the transfer of an integer flag to swap which potential is used within each replica. As with temperature parallel tempering, the running averages are swapped at the same time as the replicas, and the data within files saved for postprocessing are identified by the time-step and an unique integer tag for each of the potentials.

## 3. Simulations

Calculations for this work are based on two models: a united atom *n*-tetracosane ($C_{24}$) system simulated in the constant *NVT* ensemble, at a density of 776.2 kg $m^{-3}$ using using the Trappe force field[25] and a 1 fs time-step and a ($L/D$ = 5) soft repulsive spherocylinder model (corresponding to a cut-and-shifted 12:6 potential for the minimum distance of two line-segments[26,27]). The latter was simulated in reduced units in the *NVT* ensemble, with $D = \sigma = \varepsilon = 1$, using a reduced density of $\rho^* = N\sigma^3/V$, a reduced temperature of $T^* = kT/\varepsilon = 1$, and the methodology described by Earl et al.[26] Parallel tempering Monte Carlo moves were attempted every 200 time-steps. For the purposes of benchmarks, coordinate data was saved every 10 000 time-steps (corresponding to 10 ps for the united atom model), which is typical of that required for many molecular or coarse-grained simulations. The algorithm was tested on two parallel cluster systems. The CLX system at the CINECA supercomputer center in Italy is based around 512 2-way IBM X335 nodes (Xeon Pentium IV at 3.06 GHz) with Myrinet interconnects, while the newer BCX system used the next generation of cluster technology with Opteron Dual Core 2.6 GHz processors and Infiniband (5 Gb $s^{-1}$) interconnects for each two (dual-core) processor node.

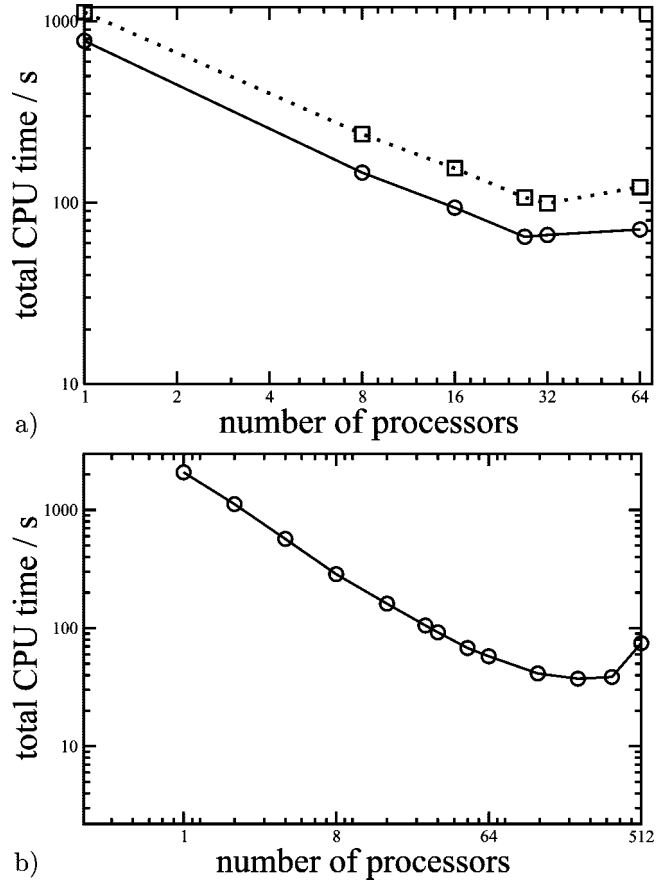



***Figure 4.*** (a) Timings for a system of 216 *n*-tetracosane molecules (5184 united atom sites) on the CINECA CLX (dotted line) and BCX (bold line) clusters as a function of the number of CPU cores. (b) Timings for 64 000 spherocylinder sites on the BCX cluster.

## 4. Results and Discussion

Figure 4 shows the typical parallel scaling behavior of a domain decomposition program. We show two systems, a small system of 216 *n*-tetracosane molecules (5184 sites) and a larger system of 64 000 soft repulsive spherocylinders. The scaling shown is extremely system size dependent as expected. The maximum speed for the small united atom system is achieved for a 3 × 3 × 3 domain decomposition on the BCX system, and thereafter, addition of parallel cores leads to a degredation in performance. It is interesting to note that changing to the next generation of parallel cluster (BCX instead of CLX) does not help the scaling signifcantly (in fact parallel scaling is slightly worse). This is because for the Lennard-Jones united atom system used, the CPU time required per step is rather small. So although parallel communications are improved in going from Myrinet to Infiniband, this is more than compensated by the increases in processor speed for the newer system. As with many molecular dynamics models of this size, eight cores provides a good balance between speedup, CPU cost, and performance and provide an absolute speedup on the BCX system by a factor of 5.3× compared to a single processor. Clearly with more expensive potentials, or if long-range electrostatic forces, etc. were included, increasing the ratio of computation/communication leads to even better scaling.

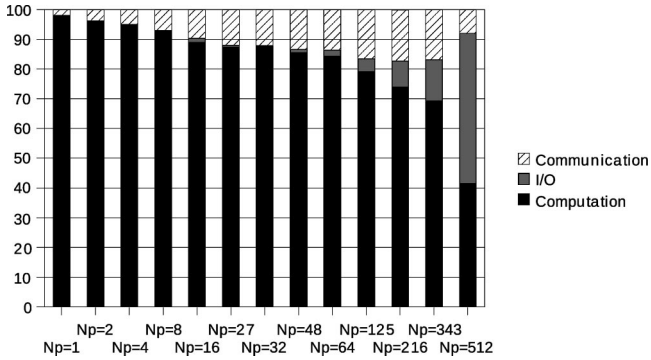



***Figure 5.*** (a) Breakdown of total processing time in terms of computation, I/O, and communication for 64 000 spherocylinder sites on the BCX cluster.

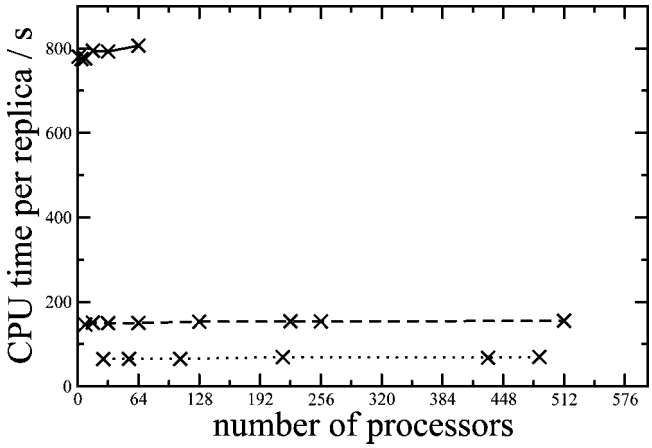



***Figure 6.*** CPU time per replica showing almost perfect scaling for a 1 × 1 × 1 (bold line), 2 × 2 × 2 (dashed line), and 3 × 3 × 3 domain decomposition (dotted line) for 216 *n*-tetracosane molecules.

As expected, an increased computation/comunication ratio is reflected in the better parallel scaling seen for the 64 000 site spherocylinder system (bottom of Figure 4), which shows increases in performance up to 216 processors (a 6 × 6 × 6 domain decomposition). For this system, the relative costs of computation, commmunications, and input/output (I/O) are shown in Figure 5. As the processor number increases, the percentage of time spent in computation drops significantly. Interestingly, at very large numbers of processors, the I/O starts to become a significant cost. Although I/O is carried out in parallel for **gbmoldd**, the problems of many processor cores writing to the same filesystem in parallel is clearly seen. At 125 cores, I/O is becoming a major contribution to the total time required for a run. Alongside communication and I/O costs, the effects of some force duplication for boundary atoms, imperfect load balancing due to density fluctuations, and additional sorting all limit parallel scaling (a full discussion of these features is given in the recent work of Hess et al.).[28]

In contrast to domain decomposition, the parallel tempering communication costs are extremely small. In Figure 6, we show the CPU time per replica for 1 × 1 × 1, 2 × 2 × 2, and 3 × 3 × 3 domain decompositions of the (5184 atom site) *n*-tetracosane system. The time per replica is virtually unchanged as the number of replica systems grow in each

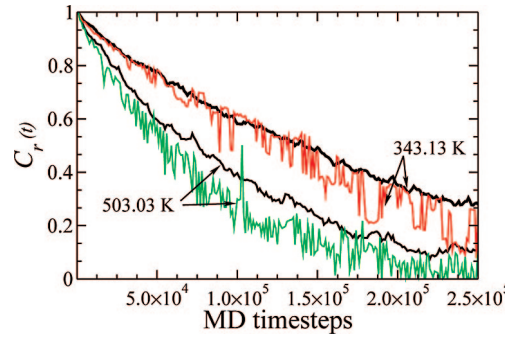



***Figure 7.*** Time correlation function, $C_r(t)$, for the normalized molecular end-to-end vector calculated for $C_{24}$ united atom chains in the liquid phase. Results are shown for two temperatures from conventional molecular dynamics simulations (bold black lines) superimposed on the results from parallel tempering (red and green lines). The latter are taken from a simulation with 32 replicas with each replica started from the same initial coordinates.

case. In the algorithm presented above, this is helped by avoiding coordinate swaps. The swapping of ensemble information (temperatures and averages) can be achieved by overlapping communications between a set of processor cores within a replica and the corresponding cores in the neighboring replica. Even for processor numbers of over 256, there is only a very small increase in CPU time, caused principally by parallel I/O degredation.

The real benefits of parallel tempering combined with domain decomposition become apparent in equilibrating molecular systems with slow relaxation times at a range of temperatures. In Figure 7, we compare the slow decay of the time correlation function of the molecular end-to-end vector in the $C_{24}$ system, defined by

$$C_r(t) = \left\langle \frac{(\mathbf{r}_1(0) - \mathbf{r}_{24}(0))\cdot(\mathbf{r}_1(t) - \mathbf{r}_{24}(t))}{|\mathbf{r}_1(0) - \mathbf{r}_{24}(0)||\mathbf{r}_1(t) - \mathbf{r}_{24}(t)|} \right\rangle$$

The decay of this function is influenced by both internal conformational changes and (to a lesser extent) by molecular rotation. However, for chain molecules, the decay of $C_r(t)$ is rather slow as illustrated in the figure. In fact, for longer (polymer) chains, the decay of $C_r(t)$ becomes particularly slow. This occurs for chain lengths greater than the entanglement length of the polymer. Here, reptation becomes the dominant mechanism for a polymer chain to relax. This provides a major limitation in the use of MD simulation to equilibrate polymer melts. The plots in Figure 7 result from starting each simulation (including each replica) from the same configuration. Here, for the parallel tempering simulations, 32 replicas are used to span the temperature range of 343.13−735.66 K. For both temperatures plotted in Figure 7 (and also all temperatures in the PT simulation), parallel tempering is seen to improve the decay of $C_r(t)$ by facilitating the movement of the system through phase space.

The results of this paper suggest that parallel tempering (replica exchange MD) and domain decomposition combine well together within a single molecular dynamics code. As a consequence, it becomes cost-effective to use a relatively small number of processors in an efficient domain decomposition and to then add a further level of parallelization

through replica exchange to improve sampling of phase space. In this way, large parallel clusters may be applied to a single small system problem. While domain decompostion is limited for a small number of particles (both in efficiency and in the physical limits imposed by the number of domains used), replica exchange is most efficient with relatively small systems, because the distribution of energy states in a smaller system allows for larger pertubations of the temperature or the potential between replicas.

Finally, it is worth noting that the approach adopted in this work is particularly useful for many of today's commodity clusters. Typical commodity systems can have two quad-core processors sharing RAM on a single board. Here, it becomes very cost-effective to use the fast shared memory communications for the communication intensive domain decomposition within a single replica and use slower off-board communications (gigabit ethernet, Myrinet, or Infiniband) for the less communications bound parallel tempering.

## 5. Conclusions

A combined domain decomposition−parallel tempering algorithm is described and implemented for the coarse-grained molecular simulation program **gbmoldd**. Results are presented for benchmarks on two systems: a $C_{24}$ united atom alkane and a coarse-grained spherocylinder system. It is shown that this combined approach allows for very large number of processors (>256) to be applied efficiently to a relatively small system (∼5000 sites) with minimal communication costs.

**Acknowledgment.** The authors thank the UK research council, EPSRC, for providing funding through grant GR/S43054/01 and the CINECA supercomputer centre in Italy for providing computer time on its CLX and BCX systems. M.R.W. thanks the HPC Europa++ programme for providing funding for a visit to CINECA and the research group of Prof. C. Zannoni (University of Bologna).

### References

(1) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281* (1−3), 140.

(2) Wu, M. G.; Deem, M. W. *Mol. Phys.* **1999**, *97* (4), 559.

(3) Lin, C. Y.; Hu, C. K.; Hansmann, U. H. E. *Proteins* **2003**, *52* (3), 436.

(4) Bedrov, D.; Smith, G. D. *J. Chem. Phys.* **2001**, *115* (3), 1121.

(5) Yan, Q. L.; de Pablo, J. J. *J. Chem. Phys.* **2000**, *113* (3), 1276.

(6) Nigra, P.; Carignano, M. A.; Kais, S. *J. Chem. Phys.* **2001**, *115* (6), 2621.

(7) Swensen, R. H.; Wang, J. S. *Phys. Rev. Lett.* **1986**, *57* (21), 2607.

(8) Rathore, N.; de Pablo, J. J. *J. Chem. Phys.* **2002**, *116* (16), 7225.

(9) Falcioni, M.; Deem, M. W. *J. Chem. Phys.* **1999**, *110* (3), 1754.

(10) Earl, D. J.; Deem, M. W. *Phys. Chem. Chem. Phys.* **2005**, *7* (23), 3910.

(11) Rathore, N.; Chopra, M.; de Pablo, J. *J. Chem. Phys.* **2005**, *122* (2), 024111.

(12) Kone, A.; Kofke, D. *J. Chem. Phys.* **2005**, *122* (20), 206101.

(13) Smith, W. *Comput. Phys. Commun.* **1991**, *62*, 229.

(14) Wilson, M. R.; Allen, M. P.; Warren, M. A.; Sauron, A.; Smith, W. *J. Comput. Chem.* **1997**, *18* (4), 478.

(15) Hughes, Z. E.; Wilson, M. R.; Stimson, L. M. *Soft Matter* **2005**, *1* (6), 436.

(16) Stimson, L. M.; Wilson, M. R. *J. Chem. Phys.* **2005**, *123*, 034908.

(17) Wilson, M. R.; Ilnytskyi, J. M.; Stimson, L. M. *J. Chem. Phys.* **2003**, *119* (6), 3509.

(18) Hwang, K.; Briggs, F. A. In *Computer Architecture and Parallel Processing*; McGraw-Hill Book Company: New York, 1985; pp 32−35.

(19) Wilson, M. In *Advances in the Computer Simulations of Liquid Crystals*; Pasini, P., Zannoni, C., Eds.; Kluwer Academic Publishers: Norwell, MA, 2000; chapter 13, pp 389−412.

(20) Ilnytskyi, J. M.; Wilson, M. R. *Comput. Phys. Commun.* **2001**, *134*, 23.

(21) Ilnytskyi, J. M.; Wilson, M. R. *Comput. Phys. Commun.* **2002**, *148*, 43.

(22) Cuetos, A.; Ilnytskyi, J. M.; Wilson, M. R. *Mol. Phys.* **2002**, *100* (24), 3839.

(23) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314* (1−2), 141.

(24) Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, *116* (20), 9058.

(25) Martin, M. G.; Siepmann, J. I. *J. Phys. Chem. B* **1998**, *102* (14), 2569.

(26) Earl, D. J.; Ilnytskyi, J.; Wilson, M. R. *Mol. Phys.* **2001**, *99*, 1719.

(27) Cuetos, A.; Martinez-Haya, B.; Rull, L. F.; Lago, S. *J. Chem. Phys.* **2002**, *117* (6), 2934.

(28) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4* (3), 435.

CT800255R

JCTC Journal of Chemical Theory and Computation

# Bonding in Low-Coordinate Environments: Electronic Structure of Distorted Square-Planar Iron-Imido Complexes With Pincer-Type Ligands

Jeanet Conradie[†,‡] and Abhik Ghosh*[,†]

*Department of Chemistry and Center for Theoretical and Computational Chemistry, University of Tromsø, N-9037 Tromsø, Norway, and Department of Chemistry, University of the Free State, 9300 Bloemfontein, Republic of South Africa*

Received April 6, 2008

**Abstract:** Low-coordinate architectures sustain unusual chemistry for middle and late transition metals, of which imido complexes are an excellent example. Recent DFT studies have uncovered a number of unusual features in the bonding in trigonal-planar and pseudotetrahedral imido complexes. Herein, we have extended these studies to a unique, distorted square-planar iron-imido complex with a pincer-type pyridine-2,6-diimine (PDI) supporting ligand. DFT calculations indicate that the iron center in the formally Fe(II) complex Fe(PDI)(NPh) is better described as intermediate-spin Fe(III), antiferromagnetically coupled to a $b_2$-symmetry PDI $\pi$-anion radical. A comparative analysis of the major classes of low-coordinate imido complexes has uncovered a certain similarity between Fe(PDI)(NPh) and a trigonal-planar Fe(III)-nacnac-imido complex. Both ligand architectures afford a total of four energetically accessible d orbitals, resulting in intermediate-spin Fe(III) centers.

## Introduction

Middle and late transition metal-imido complexes have remained rather elusive until quite recently. Undoubtedly, the strong metal($d_\pi$)-$N_{imido}$($p_\pi$) antibonding interactions that many such complexes would entail have stood in the way of their synthesis and characterization. The use of low-coordinate architectures, however, has recently led to the synthesis of a number of trigonal-planar and pseudotetrahedral iron-,[1,2] cobalt-,[3,4] and nickel-[5] imido complexes. The bonding in these complexes has unusual features, which we have analyzed with density functional theory calculations in the course of a number of papers.[6−8] One middle transition metal-imido complex that has so far escaped a quantum chemical analysis is an iron-imido complex with a pincer-type pyridine-2,6-diimine (PDI) supporting ligand.[9] We undertook a theoretical study of this complex, not only on account of its distorted square-planar geometry, which is unique for an iron-imido complex, but also to examine the proposal that this formally $Fe^{II}$-imido species is better described as an $Fe^{III}$-imido $PDI^{\bullet-}$ radical.[9] Our goal, first and foremost, was to arrive at a qualitative molecular orbital description for this complex. Accordingly, we adopted a spin-unrestricted (broken-symmetry) density functional theory (DFT) approach in this study, even though some might have preferred considerably more demanding multiconfigurational *ab initio* methods (such as the popular CASPT2 method) for this problem.[10,11] DFT has been extensively and successfully applied to a variety of noninnocent[12,13] ligands (and metalloradicals), and we will see that our calculations nicely confirm the Fe(III) PDI-anion-radical description proposed by the experimentalists.

## Methods

In general, all calculations used the OLYP[14,15] generalized gradient approximation (GGA), triple-$\zeta$ plus polarization Slater-type orbital basis sets, and a fine mesh for numerical integration of the matrix elements, all as implemented in the ADF2006 program system.[16] The choice of OLYP as the default functional is based on a number of recent studies from our laboratory, where OLYP proved to be one of the

* Corresponding author e-mail: abhik@chem.uit.no.
† University of Tromsø.
‡ University of the Free State.

10.1021/ct800118c CCC: $40.75 
Published on Web 09/13/2008

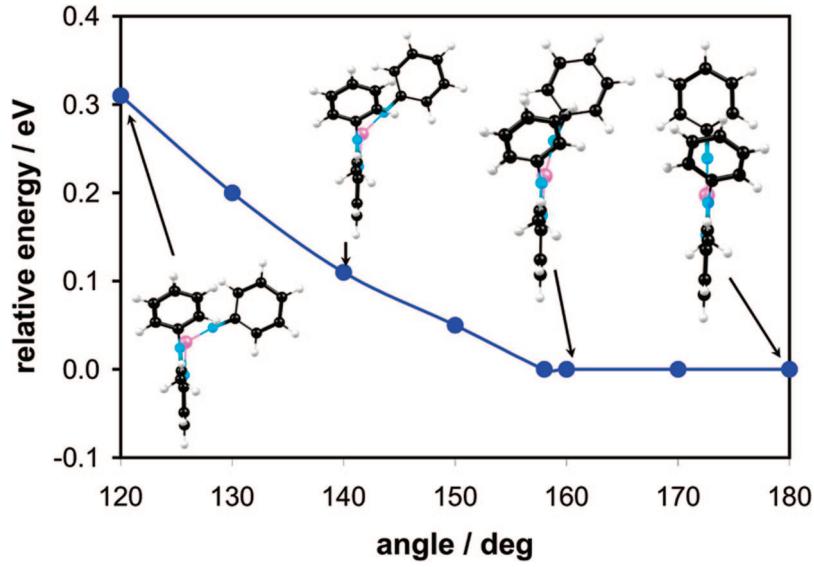



***Figure 1.*** Potential energy of Fe(PDI)(NPh) as a function of out-of-plane bending of the Fe−N−$C_{Ph}$ angle.

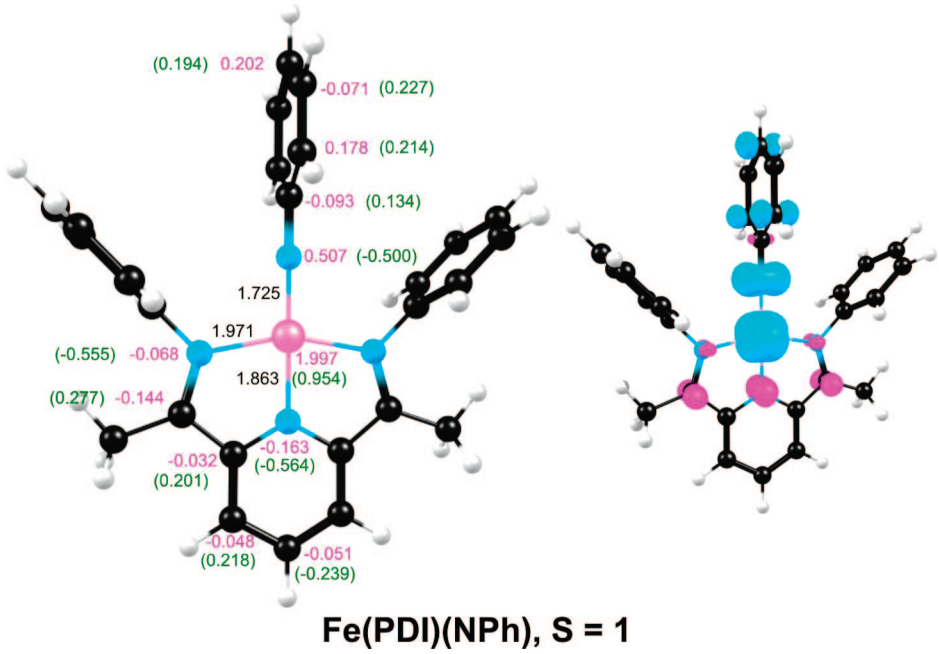



***Figure 2.*** Ground-state ($M_S = 1$) OLYP/TZP results of Fe(PDI)(NPh) ($C_{2v}$). The diagrams to the left depict bond distances (Å, in black), Mulliken spin populations (magenta), and charges (green). Spin density plots are shown to the right (majority spin in cyan, minority spin in magenta). Color code for atoms: C (black), H (ivory), N (cyan), and Fe (pink).

better functionals for transition metal systems.[6−8] Most calculations were also repeated with other functionals, including PW91,[17] BP86,[18] BLYP,[19,15] PBE,[20] B3LYP,[21,15] and B3LYP*.[22,15]

The experimentally characterized iron-imido complex that we sought to model is based on the $^{i\mathrm{Pr}}$PDI (= 2,6-$^{i}$Pr$_2$C$_6$H$_3$-N=CMe)$_2$C$_5$H$_3$N) ligand.[9] In our calculations, we chose a simpler version of the ligand, viz. PDI (=PhN=CMe)$_2$-C$_5$H$_3$N). In the same spirit, we chose the simple phenylimido ligand as the fourth ligand, whereas a more sterically hindered arylimido ligand was employed experimentally. Although the experimentally studied complex exhibits a bent Fe−$N_{imido}$−$C_{Ar}$ angle, our calculations showed that deformation of this angle is exceedingly soft, as shown in Figure 1. Much of our analysis of the Fe(PDI)(NPh) model complex is therefore based on a $C_{2v}$ geometry. In general, we found that the NPh group prefers to orient itself perpendicular to the plane of the pyridine group.

Finally, to better appreciate the neutral Fe(PDI)(NPh) complex, we have also studied the low-lying electronic states of the $[Fe(PDI)(NPh)]^+$ cation.

## Results

**(a). Basic Electronic-Structural Description of Ground-State Fe(PDI)(NPh).** Figure 2 depicts selected OLYP/TZP results for Fe(PDI)(NPh), optimized under a $C_{2v}$ symmetry constraint. The bond distances to iron in the S = 1 optimized

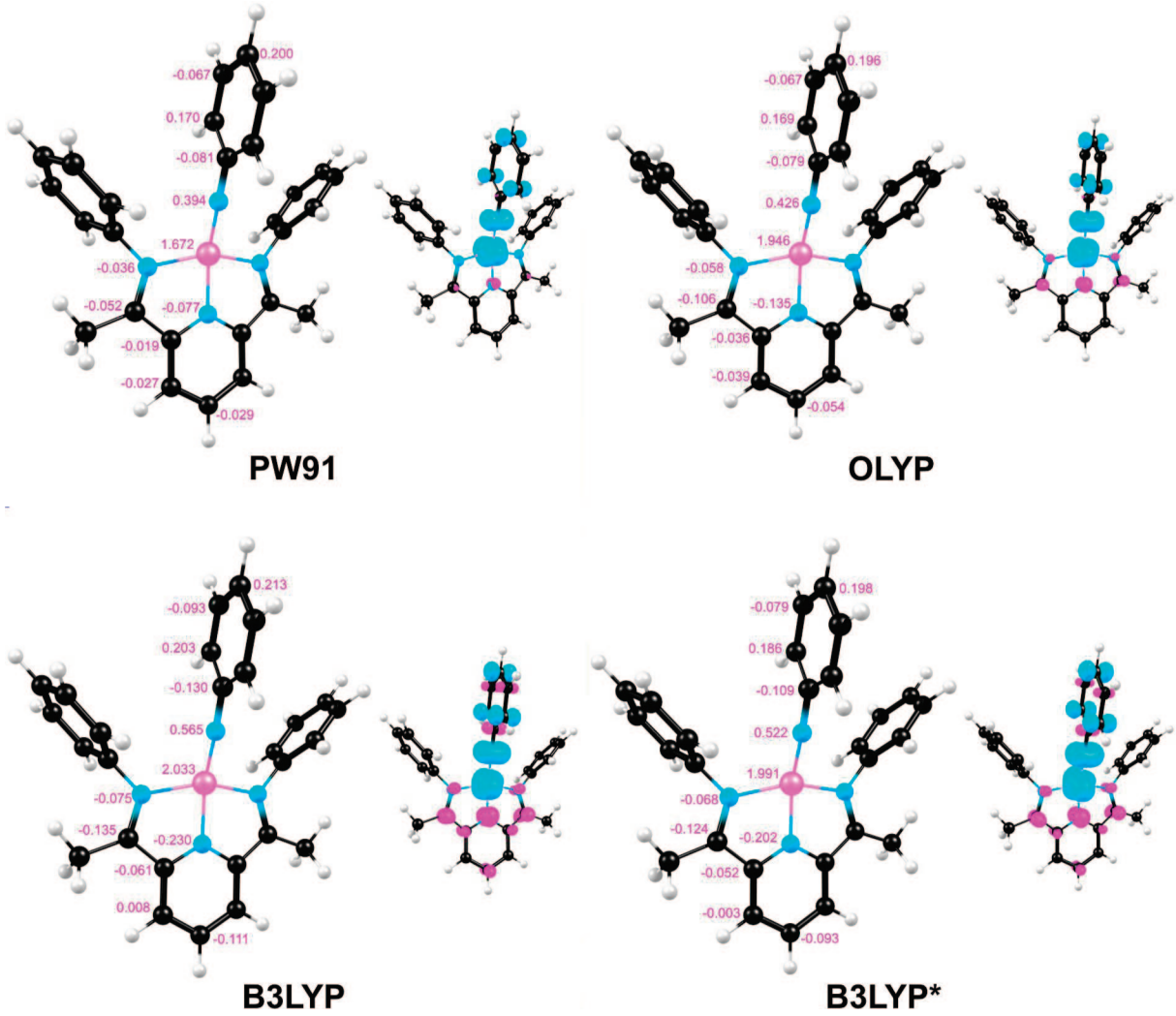



***Figure 3.*** Mulliken spin populations and spin density plots (majority spin in cyan, minority spin in magenta) for the S = 1 ground-state of Fe(PDI)(NPh) for different functionals. Color code for atoms: same as in Figure 2.

structure of Fe(PDI)(NPh) (Figure 2) agree quite well with experiment:[9] Fe−$N_{imido}$ 1.725 Å (expt. avg. 1.711 Å), Fe−$N_{pyridine}$ 1.863 Å (expt. avg. 1.856 Å), and Fe−$N_{imino}$ 1.971 Å (expt. avg. 2.0224 Å). A major difference between the optimized $C_{2v}$ structure and that observed experimentally is that the imido linkage in the latter is strongly bent out of the PDI plane. However, as mentioned above, a potential energy curve as a function of out-of-plane Fe−$N_{imido}$−C bending (under a $C_s$ symmetry constraint) revealed an exceedingly flat potential; hence the use of a $C_{2v}$ constraint appeared both convenient and justifiable.

The ground-state OLYP spin density profile of Fe(PDI)(NPh) shown in Figure 2 is of considerable interest. The majority spin density, largely localized on the Fe═NPh moiety, adds up to about 3 electrons; the minority spin density is localized largely on the PDI ligand and adds up to about 1 electron. In other words, the PDI ligand appears to be noninnocent, and the complex, overall, appears to be describable as an S = 3/2 Fe(III) $PDI^{\bullet-}$ anion radical. (Of course, broken-symmetry solutions such as this one do not correspond to a specific spin state or S but rather only to $M_S = 1$. However, the solution is *predominantly* S = 1, and, accordingly, the spin density profile indicated in Figure 2 is certainly a qualitative approximation to the true triplet spin density.) Figure 3 depicts the same (i.e., $M_S = 1$) spin density profile, obtained with three additional functionals. Observe that the classic pure functional PW91 provides a more covalent spin density relative to OLYP; i.e. there is less spatial separation of majority and minority spin density with PW91. The hybrid functional B3LYP behaves oppositely, producing the largest separation of majority and minority spin densities. In contrast, B3LYP*, which has a reduced amount of Hartree−Fock exchange (15%) relative to B3LYP (20%), yields a more or less OLYP-like spin density profile. The $\langle S^2 \rangle$ value (ideally 2.0) obtained with the different functionals may be correlated with these findings: PW91 2.137, OLYP 2.345, B3LYP* 2.534, and B3LYP 2.655.

To place the above results in context, we also carried out OLYP/TZP calculations on Fe(PDI)(O) and Fe(PDI)$Cl_2$; highlights of the results are shown in Figure 4. Not surprisingly, the spin density profile of the S = 1 ferryl species is rather similar to that described above for Fe(PDI)(NPh), implicating a similar S = 3/2 Fe(III) $PDI^{\bullet-}$ electronic description. In contrast, Fe(PDI)$Cl_2$ exhibits an S = 2 high-spin Fe(II) ground state, as found for a similar

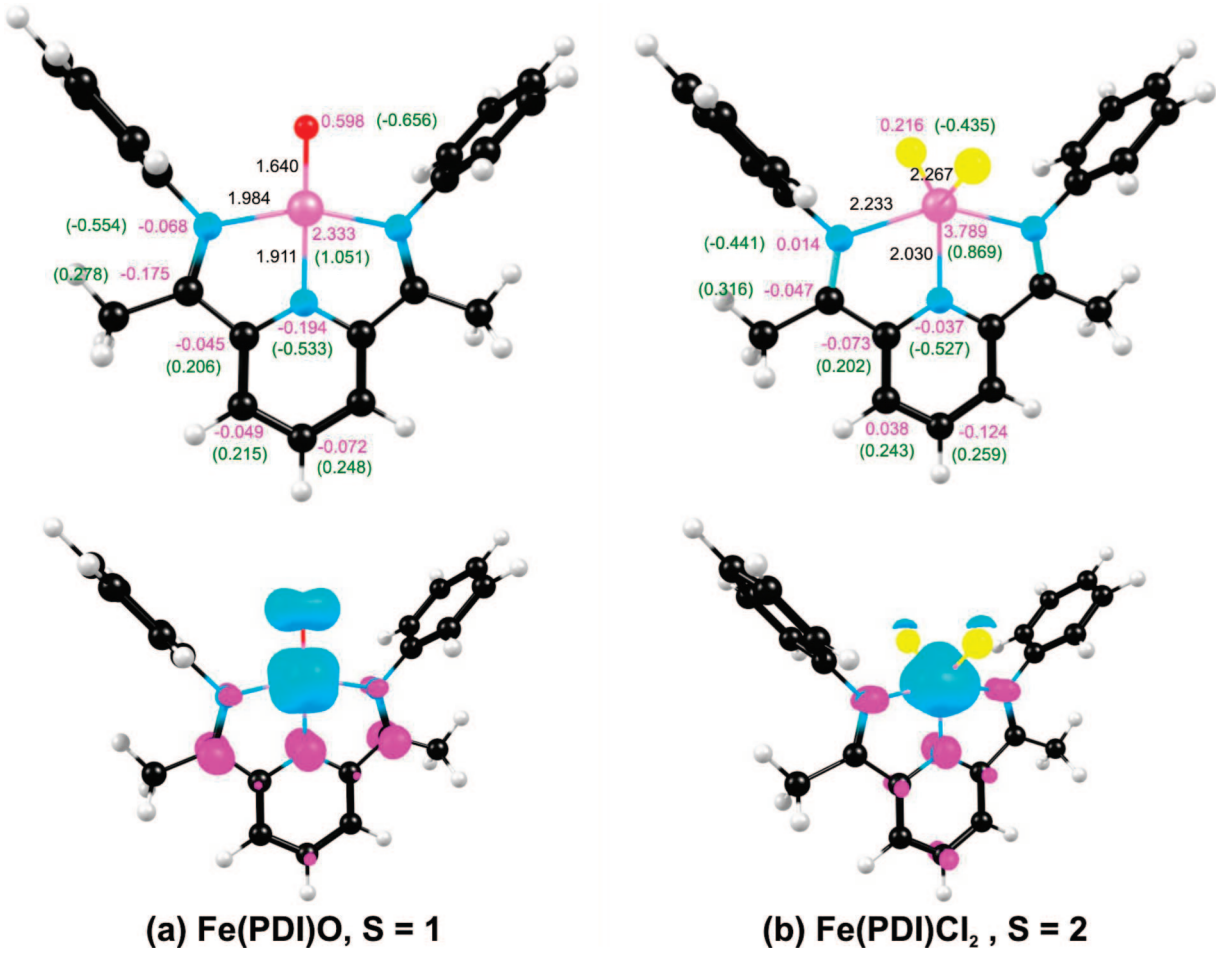



***Figure 4.*** OLYP/TZP results for the lowest-energy states of Fe(PDI)(O) ($C_{2v}$) and Fe(PDI)$Cl_2$ ($C_{2v}$). The diagrams at the top depict bond distances (Å, in black), Mulliken spin populations (magenta), and charges (green). Spin density plots are shown in the lower row (majority spin in cyan, minority spin in magenta). Color code for atoms: same as in Figure 2, O (red) and Cl (yellow).

complex in the literature, although it too exhibits small amounts of minority spin density based on the PDI.[23]

An examination of the Kohn−Sham orbitals of Fe(PDI)(NPh), as shown in Figure 5, now allows a more detailed description of its electronic configuration. Aligning the Fe-NPh vector along the $z$ axis and the mean plane of the PDI ligand along the xz plane, we may describe the symmetries of the five d orbitals as follows:

$a_1$ : $d_{y_2}(a_{1\text{-}2})$, $d_{x2\text{-}z2}(a_{1\text{-}1})$

$a_2$: $d_{xy}$

$b_1$: $d_{xz}$(also describable as $\pi_{xz}$*)

$b_2$: $d_{yz}$(also describable as $\pi_{yz}$*)

The electronic configuration of the iron center may now be described as $(d_{xy})^2(d_{y2})^1(\pi_{xz}^*)^1(\pi_{yz}^*)^1$. However, an examination of the $b_2$ minority-spin HOMO, which has some $d_{yz}$ character, indicates that it is primarily a PDI-based MO. Thus, the MO energy level diagram shown in Figure 5 provides a rather straightforward confirmation of the S = 3/2 Fe(III) $PDI^{\bullet-}$ electronic description alluded to above. Simple group-theoretic arguments now indicate that the electronic ground state of Fe(PDI)(NPh) is $^3B_1$, where the $b_1$ irrep is symmetric with respect to reflection across the PDI plane.

**(b) Spin-State Energetics.** For DFT calculations on transition metal compounds, the question of relative energetics of different spin states is a significant one. Exploiting $C_{2v}$ symmetry of Fe(PDI)(NPh), we have been able to optimize a number of different occupations as well as cationic states. Table 1 lists the relative energetics of the various charge-neutral states of Fe(PDI)(NPh) for different functionals, while Figure 6 depicts selected geometry parameters, Mulliken charges, spin populations, and spin density plots.

All the functionals yield the same S = 1 ground state. An open-shell singlet low-spin $(d_{xy})^2(d_{y2})^2(\pi_{yz}^*)^1$ Fe(III) state with an antiferromagnetically coupled $b_2$ $PDI^{\bullet-}$ radical is over an eV higher in energy, regardless of the functional. In contrast, a closed-shell singlet $(d_{xy})^2(d_{y2})^2(\pi_{yz}^*)^2$ "Fe(II)" state is significantly lower in energy, about 0.3−0.6 eV for pure functionals and about 1 eV for hybrid functionals.

All the functionals also predict a fairly low-energy quintet state, at about half an eV above the ground state; this state is best described as intermediate-spin $(d_{xy})^2(d_{y2})^1(\pi_{xz}^*)^1$-$(\pi_{yz}^*)^1$ Fe(III) ferromagnetically coupled to a $b_2$ $PDI^{\bullet-}$ radical. With a spin population of 2.222 (as shown in Figure 6), the iron center is clearly not high-spin, i.e. not S = 5/2 Fe(III).

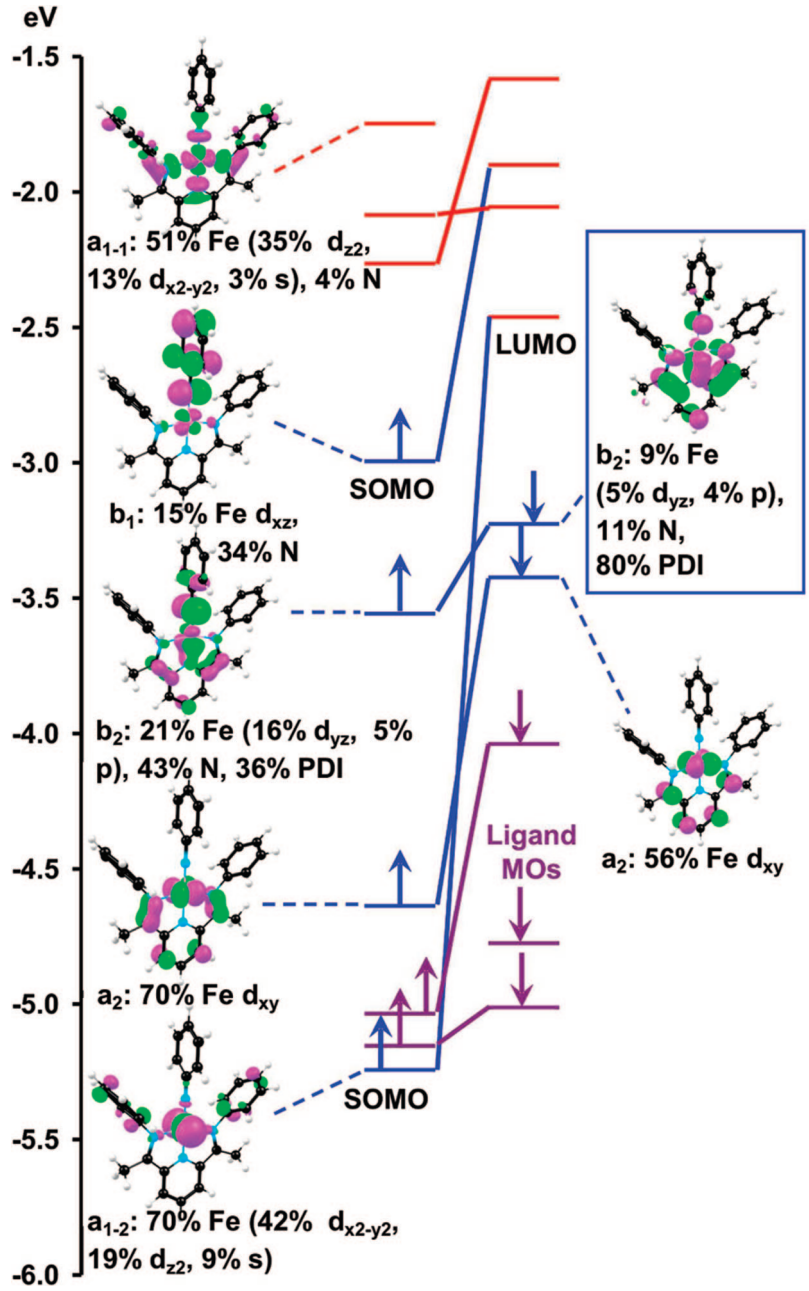



***Figure 5.*** OLYP/TZP MO energy level diagram for the S = 1 ground state of Fe(PDI)(NPh). Note that whereas the α-spin $b_2$ HOMO has 64% Fe=NPh character and only 36% PDI character, a reverse situation applies for the β-spin $b_2$ HOMO (shown in the blue box), which is overwhelmingly (80%) PDI-based. Metal d-based occupied MOs are indicated in blue, occupied ligand-based MOs in maroon, and unoccupied MOs in red.

For the septet state, involving a high-spin Fe(III) center ferromagnetically coupled to a $b_2$ $PDI^{\bullet -}$ radical, the energy is strongly functional-dependent. Thus, whereas with pure functionals the energy hovers in the range 1.1 ± 0.3 eV, B3LYP predicts an energy of only 0.33 eV relative to the ground state. As elsewhere, B3LYP* (with a reduced amount of Hartree−Fock exchange, 15%, relative to B3LYP) yields an intermediate energy of about 0.6 eV.[24]

Recently, Chirik and co-workers have reported detailed spectroscopic and theoretical studies of reduced Fe(PDI) complexes, including many with neutral ligands. Theory and experiment concur that a number of these species contain $PDI^{\bullet\bullet 2-}$ dianion-diradical ligands.[23] However, such an electronic-structural description does not appear to be relevant for the species examined in this study.

**(c) The $[Fe(PDI)(NPh)]^+$ Cation.** We have briefly studied the $[Fe(PDI)(NPh)]^+$ cation with OLYP/TZP optimizations of the lowest S = 1/2, 3/2, and 5/2 states, in case it is generated and characterized in the future. Unfortunately, the calculations do not afford a reliable energy ordering of the three states. The doublet and the quartet are approximately equienergetic, whereas the sextet is several tenths of an electronvolt higher in energy. These relative energies could be easily upset with other functionals as well as with more accurate quantum chemical methods such as CASPT2. Figure 7 presents selected results for the $[Fe(PDI)(NPh)]^+$ cation.

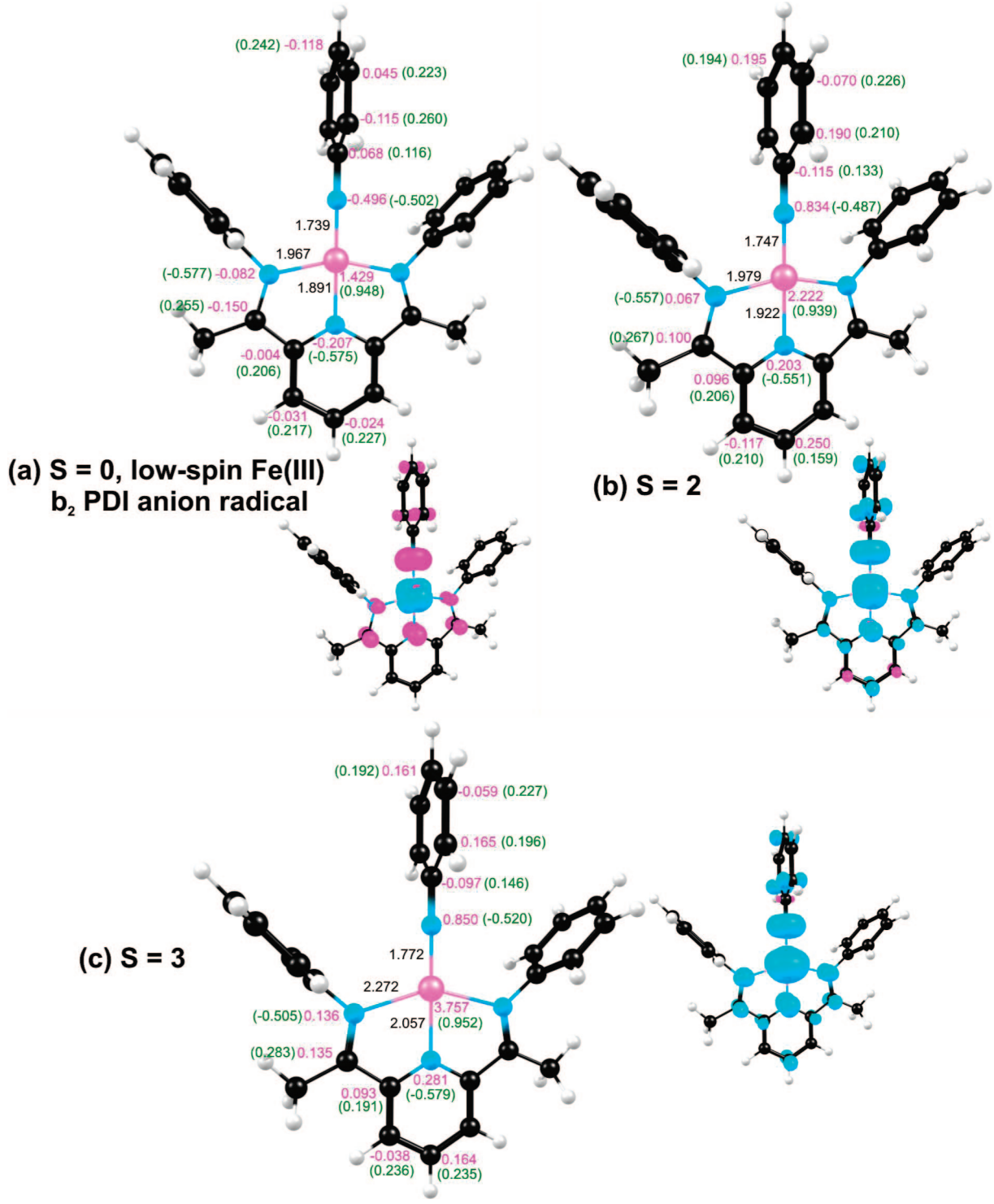



***Figure 6.*** OLYP/TZP results for the selected excited states of Fe(PDI)(NPh), optimized under a $C_{2v}$ symmetry constraint. The diagrams to the left depict bond distances (Å, in black), Mulliken spin populations (magenta), and charges (green). Spin density plots are shown to the right (majority spin in cyan, minority spin in magenta). Color code for atoms: same as in Figure 2.

The cationic S = 3/2 state is best described as an intermediate-spin Fe(III) state with a $(d_{xy})^2(d_{y2})^1(\pi_{xz}^*)^1(\pi_{yz}^*)^1$ configuration, just like the ground-state neutral species (see above). (In other words, the $b_2$ $\pi$-radical electron has been lost.) Occupancy of both $\pi^*$ orbitals results in a substantial spin density on the imido nitrogen, exactly as in the case of the ground-state neutral. In contrast, the cationic S = 1/2 state is best described as a $(d_{xy})^2(d_{y2})^1(\pi_{yz}^*)^1$ Fe(IV) center antiferromagnetically coupled to a $b_2$ $PDI^{\bullet-}$ anion radical. The low energy of this state once again demonstrates the strong tendency of the Fe-imido unit to adopt a higher oxidation state and that of the PDI ligand to become an anion radical.

## Discussion

Armed with a basic description of the electronic structure of the complex of interest in this study, we can now attempt

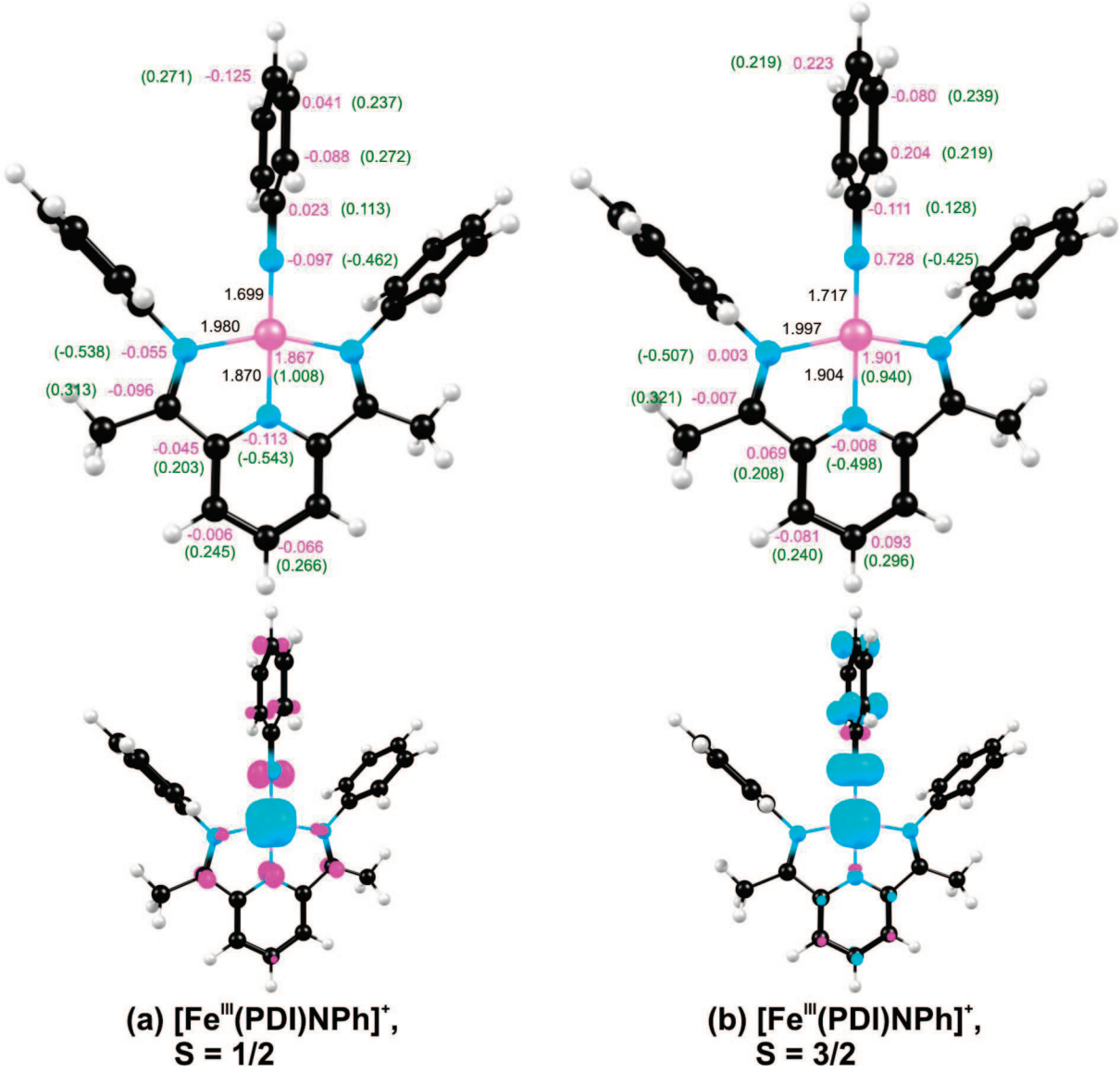



***Figure 7.*** OLYP/TZP results for the lowest-energy S = 1/2 and 3/2 states of [Fe(PDI)(NPh)]$^+$ ($C_{2v}$). The diagrams at the top depict bond distances (Å, in black), Mulliken spin populations (magenta), and charges (green). Spin density plots are shown below (majority spin in cyan, minority spin in magenta). Color code for atoms: same as in Figure 2.

to contextualize it vis-à-vis recent findings on low-coordinate middle and late transition metal-imido complexes. The optimized Fe−$N_{imido}$ distance of 1.725 Å is similar to that we calculated for a three-coordinate $Fe^{III}$(nacnac)(NPh) model complex[7] (the Fe−$N_{imido}$ distances for the relevant experimentally studied complexes are also similar). The similarity is not unexpected, given that both species feature intermediate-spin $Fe^{III}$-imido units, with a single Fe($d_\pi$)−$N_{imido}$($p_\pi$) antibonding interaction. In contrast, low-spin (S = 1/2) tris(phosphine)-supported $Fe^{III}$-imido complexes exhibit shorter Fe−N distances of about 1.65 Å, consistent with the lack of any Fe($d_\pi$)−$N_{imido}$($p_\pi$) antibonding interactions.[1,6]

A recurring theme in the low-coordinate imido literature seems to be that many of the complexes are low- or intermediate-spin, rather than high-spin.[1,6−8] In other words, the d electrons tend to avoid MOs with Fe($d_\pi$)−$N_{imido}$($p_\pi$) antibonding interactions.[25] One feature of many low-coordinate complexes that helps in this is that typically there is no ligand *trans* to the imido ligand, which results in a dramatic lowering of the energy of the $d_{z2}$-like orbital, i.e. the $d_\sigma$ orbital pointing directly at the imido group. Thus, despite its formal $\sigma$-antibonding designation, this orbital often provides a remarkably low-energy "home" for one or two electrons, which would otherwise be forced to occupy high-energy $\pi^*$ MOs.

Clearly, the above bonding paradigm does not hold for Fe(PDI)(NPh). As shown in Figure 5 and re-emphasized in Figure 8, the Fe $d_\sigma$-based orbital that points directly toward the imido group−the $d_{x2-z2}$-based $a_{1-1}$ MO−is not a low-energy orbital but a high-energy, unoccupied MO. As may be seen from Figure 5, this is the orbital in a typical square-planar complex that is destabilized by four relatively head-on $\sigma$-antibonding interactions involving the four ligands. Note from Figure 8, however, that except for this one orbital, the d orbital ordering in Fe(PDI)(NPh)

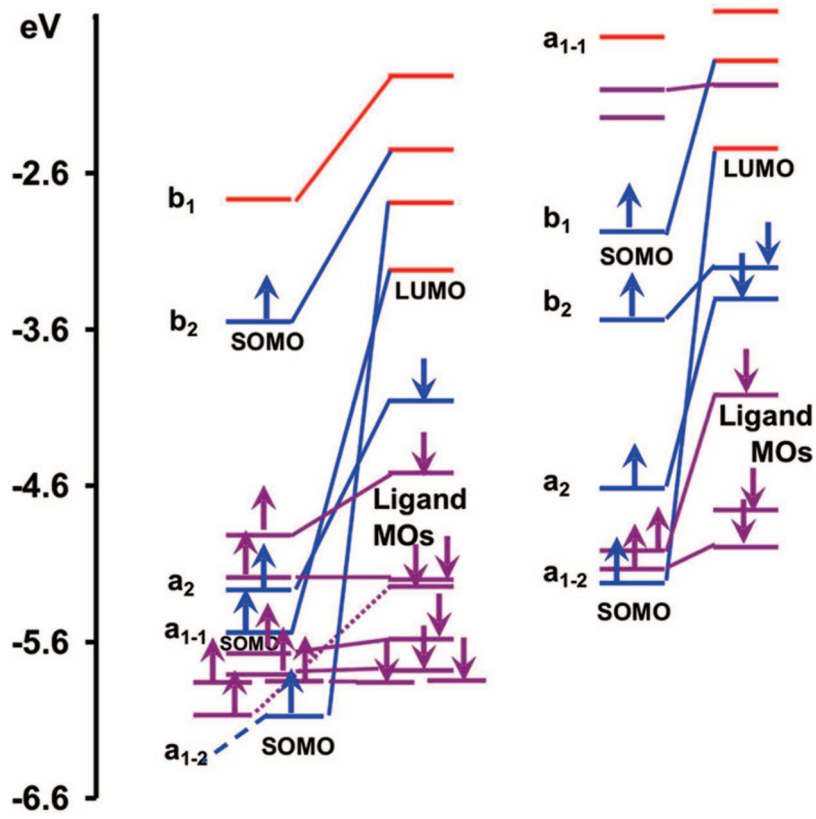



***Figure 8.*** A comparison of the OLYP Kohn−Sham metal d-based MO energy levels of Fe(PDI)(NPh) ($C_{2v}$, S = 1, right) with $Fe^{III}$(nacnac)(NPh) ($C_{2v}$, S = 3/2, left). Metal d-based occupied MOs are indicated in blue, occupied ligand-based MOs in maroon, and unoccupied MOs in red.

***Table 1.*** Relative Energies (eV) of Different Spin States of Fe(PDI)(NPh), Optimized under a $C_{2v}$ Symmetry Constraint[a]

| state | occupation | PW91 | BP86 | BLYP | OLYP | PBE | B3LYP | B3LYP* |
|---|---|---|---|---|---|---|---|---|
| S = 1 | A1 51//50<br>A2 13//13<br>B1 34//33<br>B2 23//23 | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |
| S = 0, open-shell | A1 51//50<br>A2 13//13<br>B1 33//33<br>B2 23//24 | 1.18 | 1.18 | 1.14 | 1.27 | 1.18 | 1.08 | 1.12 |
| S = 0, spin-restricted | A1 51//51<br>A2 13//13<br>B1 33//33<br>B2 23//23 | 0.32 | 0.31 | 3.12 | 0.49 | 0.29 | 1.10 | 0.85 |
| S = 2 | A1 51//50<br>A2 13//13<br>B1 34//33<br>B2 24//22 | 0.72 | 0.69 | 0.67 | 0.61 | 0.69 | 0.46 | 0.52 |
| S = 3 | A1 52//50<br>A2 13//12<br>B1 34//33<br>B2 24//22 | 1.39 | 1.32 | 1.26 | 0.80 | 1.34 | 0.33 | 0.59 |

[a] The S = 1 state has been chosen as the energy zero level.

is identical to that in $Fe^{III}$(nacnac)(NPh),[2] a molecule whose metal d-based MOs are topologically quite similar to those of Fe(PDI)(NPh).[7]

Another key distinction between the MO energy levels of the two complexes involves the "in-plane" Fe($d_{xz}$)−$N_{imido}$($p_x$) $\pi$-antibonding MO ($b_1$). In the case of trigonal-planar $Fe^{III}$(nacnac)(NPh), this MO is additionally destabilized by $\sigma$-antibonding interactions with the nacnac nitrogens and is therefore unoccupied. By contrast, no analogous $\sigma$-antibonding interaction with the PDI ligand is expected (or found) for either of the Fe−$N_{imido}$ $\pi$* MOs of square-planar Fe(PDI)(NPh). Thus, both these $\pi$* MOs are singly occupied for Fe(PDI)(NPh), whereas only one such MO is (singly) occupied for $Fe^{III}$(nacnac)(NPh). This distinction translates into a significant difference vis-à-vis the spin density profiles of the two complexes: in

essence, the former exhibits a much higher $N_{imido}$ spin population (0.51, see Figure 2) than the latter (0.13).[7]

Despite the above differences, there are intriguing similarities between the PDI and nacnac complexes. Both architectures result in a single energetically inaccessible d orbital, viz. the $a_{1-1}$ orbital in the case of Fe(PDI)(NPh) and the (above-mentioned) $b_1$ orbital for the nacnac complex. Both complexes thus afford four energetically accessible d orbitals, resulting in intermediate-spin Fe(III) centers.

## Conclusions

In conclusion, the iron center in the formally Fe(II) complex Fe(PDI)(NPh) is better described as intermediate-spin Fe(III), antiferromagnetically coupled to a $b_2$-symmetry PDI $\pi$-anion radical. With this study completed, we now have a basic theoretical survey of the major known classes of low-coordinate transition metal imido complexes. Intriguing similarities and differences have emerged between the orbital structure of Fe(PDI)(NPh) and that of a trigonal-planar Fe(III)-nacnac-imido complex. Despite key differences, both architectures, it turns out, afford a total of four energetically accessible d orbitals, thereby engendering intermediate-spin Fe(III) centers.

**Acknowledgment.** This work was supported by the Research Council of Norway (AG) and the National Research Fund of the Republic of South Africa (grant no. 61093).

**Supporting Information Available:** Table of optimized Cartesian coordinates. This material is available free of charge via the Internet at http://pubs.acs.org.

### References

(1) For a review, see: Mehn, M. P.; Peters, J. C. *J. Inorg. Biochem.* **2006**, *100*, 634–643.

(2) Eckart, N. A.; Vaddadi, S.; Stoian, S.; Lachicotte, R. J.; Cundari, T. R.; Holland, P. L. *Angew. Chem., Int. Ed.* **2006**, *45*, 6868–6871.

(3) In addition to references cited in ref 1, see: Cowley, R. E.; Bontchev, R. P.; Sorrell, J.; Sarracino, O.; Feng, Y.; Wang, H.; Smith, J. M. *J. Am. Chem. Soc.* **2007**, *129*, 2424–2425.

(4) Dai, X.; Kapoor, P.; Warren, T. H. *J. Am. Chem. Soc.* **2004**, *126*, 4798–4799.

(5) Kogut, E.; Wiencko, H. L.; Zhang, L. B.; Cordeau, D. E.; Warren, T. H. *J. Am. Chem. Soc.* **2005**, *127*, 11248–11249.

(6) Tangen, E.; Conradie, J.; Ghosh, A. *J. Chem. Theory Comput.* **2007**, *3*, 448–457.

(7) Conradie, J.; Ghosh, A. *J. Chem. Theory Comput.* **2007**, *3*, 689–702.

(8) Wasbotten, I. H.; Ghosh, A. *Inorg. Chem.* **2007**, *46*, 7890–7898.

(9) Bart, S. C.; Lobkovsky, E.; Bill, E.; Chirik, P. J. *J. Am. Chem. Soc.* **2006**, *128*, 5302–5303.

(10) Aquilante, F.; Malmqvist, P.-Å.; Pedersen, T. B.; Ghosh, A.; Roos, B. O. *J. Chem. Theory Comput.* **2008**, *4*, 694–702.

(11) Ghosh, A.; Gonzalez, E.; Tangen, E.; Roos, B. O. *J. Phys. Chem. A***2008**; ASAP Article; DOI: 10.1021/jp711159h (http://pubs.acs.org/cgi-bin/abstract.cgi/jpcafh/asap/abs/jp711159h.html).

(12) For a review, see: Ghosh, A. *J. Biol. Inorg. Chem.* **2006**, *11*, 712–724.

(13) For another DFT study of noninnocent ligands, see:(a) Ghosh, P.; Bill, E.; Weyhermuller, T.; Neese, F.; Wieghardt, K. *J. Am. Chem. Soc.* **2003**, *125*, 1293–1308.

(14) The OPTX exchange functional: Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403–412.

(15) The LYP correlation functional: Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

(16) The ADF program system was obtained from Scientific Computing and Modeling, Amsterdam (http://www.scm-.com/). For a description of the methods used in ADF, see: (a) Velde, G. T.; Bickelhaupt, F. M.; Baerends, E. J.; Guerra, C. F.; Van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 2001.

(17) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Perderson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*, 6671–6687. Erratum: Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Perderson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1993**, *48*, 4978.

(18) Becke, A. D. *Phys. Rev.* **1988**, *A38*, 3098. Perdew, J. P. *Phys. Rev.* **1986**, *B33*, 8822. Erratum: Perdew, J. P. *Phys. Rev.* **1986**, *B34*, 7406.

(19) Becke, A. D. *Phys. Rev.* **1988**, *A38*, 3098.

(20) Perdew, N. J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868. Perdew, N. J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.

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

(22) Reiher, M.; Salomon, O.; Hess, B. A. *Theor. Chem. Acc.* **2001**, *107*, 48.

(23) Bart, S. C.; Chlopek, K.; Bill, E.; Bouwkamp, M. W.; Lobkovsky, E.; Neese, F.; Wieghardt, K.; Chirik, P. J. *J. Am. Chem. Soc.* **2006**, *128*, 13901–13912.

(24) Selected studies comparing the performance of different functionals vis-à-vis transition metal spin state energetics: (a) Swart, M.; Groenhof, A. R.; Ehlers, A. W.; Lammertsma, K. *J. Phys. Chem. A* **2004**, *108*, 5479–5483. (b) Swart, M.; Ehlers, A. W.; Lammertsma, K. *Mol. Phys.* **2004**, *102*, 2467–2474. (c) Deeth, R. J.; Fey, N. *J. Comput. Chem.* **2004**, *25*, 1840–1848. (d) Groenhof, A. R.; Swart, M.; Ehlers, A. W.; Lammertsma, K. *J. Phys. Chem. A* **2005**, *109*, 3411–3417. (e) Daku, L. M. L.; Vargas, A.; Hauser, A.; Fouqueau, A.; Casida, M. E. *ChemPhysChem* **2005**, *6*, 1393–1410. (f) Ganzenmuller, G.; Berkaine, N.; Fouqueau, A.; Casida, M. E.; Reiher, M. *J. Chem. Phys.* **2005**, *122*, 234321. (g) De Angelis, F.; Jin, N.; Car, R.; Groves, J. T. *Inorg. Chem.* **2006**, *45*, 4268–4276. (h) Vargas, A.; Zerara, M.; Krausz, E.; Hauser, A.; Daku, L. M. L. *J. Chem. Theory Comput.* **2006**, *2*, 1342–1359. (i) Rong, C. Y.; Lian, S. X.; Yin, D. L.; Shen, B.; Zhong, A. G.; Bartolotti, L.; Liu, S. B. *J. Chem. Phys.* **2006**, *125*, 174102. (j) Strickland, N.; Harvey, J. N. *J. Phys. Chem. B* **2007**, *111*, 841–852. (k) Conradie, J.; Ghosh, A. *J. Phys. Chem. B* **2007**, *111*, 12621–12624.

(25) It should be noted that the ancillary ligand also plays a key role in tuning the energies of these MOs, and, with sufficiently weak-field supporting ligands, the ground-state may indeed turn out to be high-spin, as we predicted for an $Fe^{III}$(Tp)(NMe) complex (Tp = hydrotris(pyrazolyl)borate).[8]

CT800118C

JCTC Journal of Chemical Theory and Computation

# Efficient Parallel Implementation of the CCSD External Exchange Operator and the Perturbative Triples (T) Energy Calculation

Tomasz Janowski* and Peter Pulay

*Department of Chemistry and Biochemistry, Fulbright College of Arts and Sciences, University of Arkansas, Fayetteville, Arkansas 72701*

Received April 24, 2008

**Abstract:** A new, efficient parallel algorithm is presented for the most expensive step in coupled cluster singles and doubles (CCSD) energy calculations, the external exchange operator (EEO). The new implementation requires much less input/output than our previous algorithm and takes better advantage of integral screening. It is formulated as a series of matrix multiplications. Both the atomic orbital integrals and the corresponding CC coefficients are broken up into smaller blocks to diminish the memory requirement. Integrals are presorted to make their sparsity pattern more regular. This allows the simultaneous use of two normally conflicting techniques for speeding up the CCSD procedure: the use of highly efficient dense matrix multiplication routines and the efficient utilization of sparsity. We also describe an efficient parallel implementation of the perturbative triples correction to CCSD and related methods. Using the Array Files tool for distributed filesystems, parallelization is straightforward and does not compromise efficiency. Representative timings are shown for calculations with 282−1528 atomic orbitals, 68−228 correlated electrons, and various symmetries, $C_1$ to $C_{2h}$.

## 1. Introduction

Since their initial implementations in the late 1970s,[1−3] coupled cluster (CC) methods[4,5] have became one of the most successful tools of quantum chemistry. Although the CC equations are more complicated than the configuration interaction (CI) equations, the benefits they offer over CI methods are significant. In particular CC methods are size consistent (size extensive). This is very important when calculating interaction energies, particularly for weak interactions, or reaction enthalpies. The CC method with single and double substitutions (CCSD) still has significant residual errors. However, when triple substitutions are included perturbatively,[6−8] the accuracy of CC becomes almost quantitative, provided that large basis sets are used and the wave function is dominated by a single configuration. Both the size consistency and the improved accuracy relative to a truncated CI expansion are caused by the implicit inclusion of all higher order substitutions that are products of those of lower order. The quadratic CI (QCI) method is a simplified version of the CC method;[9] in QCISD, some higher order terms involving single substitutions are omitted.

We have implemented the calculation of CCSD energies in parallel for a closed-shell reference in the PQS program package.[10,11] Recently, we added the most widely used perturbative triples correction (T),[7] allowing the calculation of CCSD(T) energies. Several related many-body methods which can be considered as simplified versions of the full CC method have also been implemented: quadratic CI (QCISD), the coupled electron pair approximation (CEPA0 CEPA2),[12] perturbative methods (e.g., MP3, MP4), and variational singles and doubles CI (CISD). The SD part (e.g., CCSD) can use either canonical or noncanonical (e.g., localized) orbitals. The canonical and localized formulations are virtually identical for CI, CCSD, QCISD, etc. However, perturbational methods become iterative in an orbital-invariant, noncanonical form.[13]

The purpose of this paper is to evaluate competing strategies in CC calculation and describe our recent improvements. As our program is able to perform very large calculations on modest hardware, we think it would be

* Corresponding author. E-mail: janowski@uark.edu.

10.1021/ct800142f CCC: $40.75 
Published on Web 09/20/2008

interesting to present details of our current implementation with emphasis on recent improvements and additions. In particular, our method of utilizing sparsity in the external exchange (four virtual indices) part may be of interest, as it combines two techniques that are generally incompatible: the elimination of very small integrals while using high speed dense matrix routines.

Our target hardware is moderately sized PC and workstation clusters.

At the start of this project, the only available distributed memory parallel CC programs were MOLPRO[14] and a developmental version of NWChem.[15] This was surprising because the massive CPU demand and relative simplicity of CC methods makes them ideally suited to parallel processing. More recently ACES II[16] and GAMESS-US[17] joined the list. Very recently, a novel implementation in ACES III was reported.[18] The first parallel CC codes[19−23] were generally developed for the Cray supercomputers and do not perform optimally (if at all) on modern distributed memory workstation clusters.

## 2. General Remarks about Efficiency

Because of the compute-intensive nature of CC calculations, it is important to achieve the highest possible efficiency. One method of doing this is to rewrite the algorithm in such a way that the compiler can generate optimum code; for instance, by changing the ordering of loops to minimize memory access. Vectorization, known mostly for the Cray series of supercomputers, was a notable example. However, optimization of this kind depends sensitively on the hardware architecture, in particular on memory hierarchy, and may become obsolete when new architectures are introduced. For CCSD(T), a simpler and more general method is to formulate the algorithm as much as possible in terms of matrix operations, in particular matrix multiplications, and use libraries optimized for the particular hardware architecture. Matrix multiplication can be formulated for essentially all computer architectures to run at close to the theoretical maximum speed.[24] We will refer to this strategy as vectorization, despite the fact that it does not conform fully to the original meaning of this term. Our implementation is based on the efficient matrix formulation of the singles and doubles correlation problem, the self-consistent electron pair theory.[25] We use the generator state spin adaptation.[26] This reduces the flop count by a factor of 2 for the pair coupling terms compared to orthogonal spin adaptation and significantly simplifies the formulas. For instance, the CCD residuum formula is only a few lines long. In the SCEP/generator state formulation, parts of the algorithm are already expressed as matrix products and thus run at close to maximum efficiency. Other parts involve linear combinations of matrices or matrix traces. These are "DAXPY" (double precision $aX + Y$) and dot product (DDOT) operations, respectively, which run at a much lower speed because they cannot reuse the fastest (cache) memory efficiently and are therefore limited by much slower main memory access. For instance, the 3 GHz Intel Nocona processor runs a $1000 \times 1000$ matrix multiplication at 5128 Mflops/s (using the Goto library)[24] while its performance for a 1 000 000 long dot product is only 465 Mflops/s and for a DAXPY operation with the same flop count is 307 Mflops/s, over 15 times slower than matrix multiplication. We note that it is unlikely that these ratios would change with newer memory architectures, as access to a small (cache) memory is inherently faster than to a large (main) one. It is therefore imperative to reformulate all operation to run as matrix multiplications. For instance, the contribution of four virtual orbitals (the external exchange operator or particle−particle ladder), discussed in the next section, is usually the computationally most demanding part of a CCSD calculation, and its natural formulation does not take the form of a matrix multiplication.

## 3. External Exchange Operator

This is usually the most demanding part of CCSD/QCISD programs. Possible methods of calculation include the atomic orbital (AO) or molecular orbital (MO) forms. If integral prescreening is not used, or does not yield significant savings, e.g., for small molecules, the MO formulation is faster, as the number of virtual MOs is always smaller than the total number of basis functions (AO). This is significantly magnified by the fourth power, giving a factor of 2 for systems where the number of virtual orbitals is only 16% smaller than the number of AOs. However, the AO formulation allows integral direct implementations, avoiding storage of four-external integrals and enabling utilization of sparsity. As with increasing computer power, we are able to treat bigger systems and sparsity becomes increasingly important, yielding 90% savings for one of our test molecules descibed later in this paper. On the other hand, contemporary disk capacities have grown significantly, so it is not obvious that integral direct methods are more advantageous than storing integrals on disk. Parallel filesystems, such as our Array Files tool[27] can use efficiently the aggregate local disk capacity of a cluster, allowing the storage of large integral files.

Nonetheless, the AO formulation, direct or not, allows sparsity utilization, which seems to be unavoidable if we really want big coupled cluster calculations. Note that the importance of sparsity arises from the reduction of the formal operation count of the EEO evaluation, not just from the reduction of integral calculation. While the cost of the integral evaluation is usually smaller than the evaluation of the external exchange, this is not true for small systems using large, diffuse basis sets. In view of the large increase in disk capacities, nondirect AO algorithms may be a viable option.

The EEO is naturally expressed as a dot product:

$$(\mathbf{K}[\mathbf{T}^{ij}])_{\mu\lambda} = \sum_{\nu\sigma} (\mu\nu|\lambda\sigma)\mathbf{T}^{ij}_{\nu\sigma} \tag{1}$$

Here, the Greek letters denote contracted basis functions, and $i$ and $j$ denote molecular orbitals, $\mathbf{T}^{ij}$ is a matrix of amplitudes associated with excitation from the $ij$ orbital pair. If sparsity or symmetry is not used, the operations count is $^1/_2 n^2 N^4$, where $n$ is the number of occupied orbitals and $N$ is the number of AOs. This is the most expensive part of a CCSD calculation, especially if the basis set ($N$) is large. The latter is usually the case, as CCSD(T) calculations require basis sets of at least triple-$\zeta$ or quadruple-$\zeta$ quality[28] in order to obtain quantitative results. To take maximum

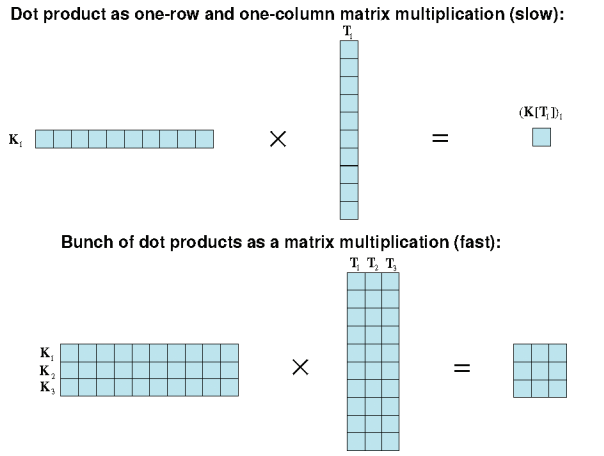



***Figure 1.*** Rearrangement of several dot products (matrix traces) into one big matrix product.

advantage of the efficiency of modern processors, the dot product shown above was reformulated as a matrix product. Figure 1 illustrates how this was done. In practice, we use symmetric and antisymmetric combinations of amplitudes and integrals as described before:[29]

$$(\mathbf{K}[\mathbf{T}^{ij}])^{\pm}_{\mu\lambda} = \frac{1}{4}\sum_{\nu\geq\sigma}[(\mu\nu|\lambda\sigma) \pm (\mu\sigma|\lambda\nu)](\mathbf{T}^{ij}_{\nu\sigma} \pm \mathbf{T}^{ij}_{\sigma\nu})(2-\delta_{\nu\sigma}) \tag{2}$$

The quantity $(\mathbf{K}[\mathbf{T}^{ij}])^{\pm}_{\mu\lambda}$, calculated for $\mu \geq \lambda$ only, gives the final external exchange by taking appropriate linear combinations:

$$\mathbf{K}[\mathbf{T}^{ij}])_{\mu\lambda} = (\mathbf{K}[\mathbf{T}^{ij}])^{+}_{\mu\lambda} + (\mathbf{K}[\mathbf{T}^{ij}])^{-}_{\mu\lambda} \tag{3}$$

and

$$\mathbf{K}[\mathbf{T}^{ij}])_{\lambda\mu} = (\mathbf{K}[\mathbf{T}^{ij}])^{+}_{\mu\lambda} - (\mathbf{K}[\mathbf{T}^{ij}])^{-}_{\mu\lambda} \tag{4}$$

This procedure reduces the formal operation count by a factor of 2, giving $^1/_4n^2N^4$, significantly better (up to a factor of 5) than in spin−orbital formulations (e.g., $^5/_4n^2N^4$ in Hirata's work).[30] An undesirable side effect of this technique is that the sparsity of integrals deteriorates somewhat.

Our previous EEO implementation was constrained by the integral program, which generated batches of full **K** matrices, as defined by $(\mathbf{K}^{\mu\lambda})_{\nu\sigma} = (\mu\nu|\lambda\sigma)$, for a given atomic orbital (AO) shell pair $\mu\lambda$. We stored as many **K** matrices as possible, using shell merging techniques, mentioned earlier by Schütz et al.,[31] constructed supermatrices out of them (see Figure 1), then read all the **T** matrices (in batches), constructed a similar supermatrix, and multiplied the two together using matrix multiplication algorithms. This technique was efficient for systems with relatively small basis sets but deteriorated significantly for systems with more than 1000 basis functions where a single matrix can occupy a large amount of memory. A good example is one of our largest calculations performed so far, benzene dimer QCISD or CCSD in the aug-cc-pVQZ basis set (1512 basis functions)[11] (see also ref 32 for more recent QCISD(T) results), where a single matrix occupies about 18 MB of memory. Because of this, we could process only batches of 100−200 $\mu\lambda$ pairs at a time. Each batch required reading the full set

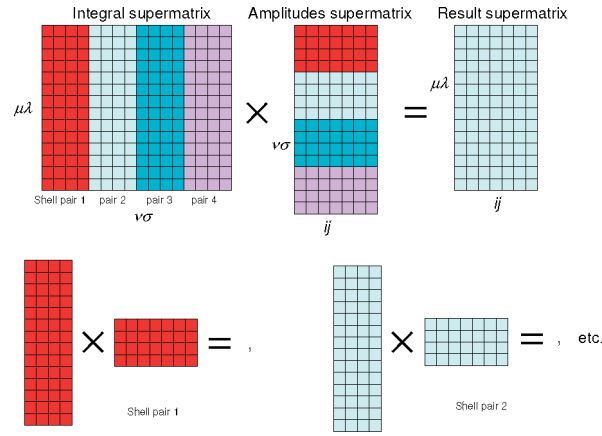



***Figure 2.*** New EEO blocking scheme.

of amplitudes from a disk, resulting in excessive input/output (I/O). The amplitudes had to be read 5000−10 000 times for the benzene dimer calculation. The set of all amplitudes for this system occupied about 8 GB, so one full EEO evaluation required the reading of 40−80 TB of data. The speed of the calculation was thus limited by I/O.

We have rearranged this algorithm using an older and somewhat slower but more flexible integral program. This program is not vectorized and therefore can generate integrals for a single quartet of AO shells. This makes it possible to subdivide the range of AO index pairs in smaller blocks, alleviating the memory problem, and in turn dramatically reducing the amount of I/O. Blocking is widely used in high-performance computers, for instance in efficient matrix multiplication routines. In the context of CC methods, the tensor contraction engine of Hirata[30] uses blocks (called "tiles"). However, this technique does not exploit sparsity. A local coupled cluster implementation was reported recently[33] that exploits integral sparsity using an approach different from ours.

We define batches of indices $\mu\lambda$, $\nu\sigma$, and $ij$. The amplitude matrices are presorted into blocks corresponding to these batches. The outermost loop is parallelized; it runs over batches of $\mu\lambda$. The next loop runs over batches of $\nu\sigma$. At this point, all integrals $(\mu\nu|\lambda\sigma)$ for the $\mu\lambda$ and $\nu\sigma$ batches are calculated. The innermost loop runs over batches of $ij$. A block of the amplitudes corresponding to $ij$ and $\nu\sigma$ is read from distributed storage, and a block of partial results are calculated by matrix multiplication of $\mathbf{T}^{ij}_{\nu\sigma}$ with $(\mu\nu|\lambda\sigma)$, using $\nu\sigma$ as the summation index. The result is accumulated in $\mathbf{K}[\mathbf{T}^{ij}]_{\mu\lambda}$. After the $ij$ and $\nu\sigma$ loops are finished, the completed $\mathbf{K}[\mathbf{T}^{ij}]_{\mu\lambda}$ matrices are written to distributed disk storage. The dominant memory requirement is for the storage of the $\mathbf{K}[\mathbf{T}^{ij}]_{\mu\lambda}$ matrices and requires only $Bn^2/2$ memory locations where $B$ is the $\mu\lambda$ batch size. Only one full read of the amplitude blocks per slave is needed, a dramatic reduction compared to our original code. e.g., using the example above, with 30 slaves, the full amplitude set is read only 30 times, not 10 000 times. The algorithm is shown in Figure 2.

One may note that, as we calculate all integrals for $\mu \geq \lambda$, only one permutational symmetry out of four is used. This requires extra CPU time compared to algorithms that use the full permutational symmetry, like that of Schütz et al.[31]

However, the advantage of avoiding multiple writings and readings of partial results greatly outweighs the increased cost of integral evaluation, even in MP2. In our opinion, integral permutational symmetry is only of minor importance in correlated calculations, as integral evaluation scales formally only as $O(N^4)$, while EEO for instance scales as $n^2N^4$. This is underscored by the fact that the main use of integral screening described below is not to economize on integral evaluation but to speed up the matrix multiplications.

The algorithm briefly described above facilitates the implementation of integral screening. For large molecules, the integral matrix contains large numbers of vanishingly small integrals. For example, for a linear decapeptide of ten glycine molecules, about 90% of the integrals are numerically zero, even with a very strict threshold of $10^{-15}$. This results in the elimination of many rows (Figure 2) from the left-hand side matrix, allowing compression and the use of efficient dense matrix multiplication routines for the resulting smaller matrices.[11] One technical problem associated with the utilization of screening is that if a larger numbers of $\nu\sigma$ pairs are treated simultaneously, the location of zero integrals becomes more-or-less random and only a few rows can be eliminated. Dividing the matrix into narrower stripes, say over single shell pairs, improves the utilization of sparsity but reduces the efficiency of matrix multiplication and results in more memory copying (e.g., the accumulation of partial results in the final matrix). A compromise must be sought which allows us to treat as many $\nu\sigma$ pairs as possible at a time while still allowing efficient utilization of integral sparsity.[11]

Row sparsity can be improved significantly if we order the $\nu\sigma$ AO index pairs appropriately. It can be theoretically justified that sorting them according to the physical space location increases regularity in the sparsity pattern of the integral matrix. This phenomenon arises from the fact that most integrals are numerically zero if, for a given $(\mu\nu|\lambda\sigma)$ integral, $\mu$ is far away from $\nu$ or $\lambda$ is far away from $\sigma$. The integral will also be small if the $\mu\nu$ pair is far away from the $\lambda\sigma$ pair, but the asymptotic behavior is different. In the former two cases, the integral decreases exponentially as a function of the distance; in the latter, it is a much weaker $1/r$ dependency. Because one row of our integral matrix (Figure 2) contains all integrals for a given $\mu\lambda$ index, it is better to group all integrals so that the $\nu$ and $\sigma$ atomic orbitals are distant from $\mu$ and $\lambda$, respectively. Unfortunately, all rows must share the same ordering pattern—otherwise we cannot use matrix multiplication. The best compromise is to group the $\nu\sigma$ pairs according to the coordinates of their centers. We divide the molecule into boxes 2−3 Å big and collect all $\nu\sigma$ AOs that belong to the same pair of boxes. This improves the sparsity for larger batches of $\nu\sigma$ pairs significantly. For instance, if a batch contains $\nu$'s from box A and $\mu$ is distant from A then the whole $\mu\lambda$ row of the integral matrix $\mathbf{K}^{\mu\lambda}_{\nu\sigma}$ vanishes. The same statement holds true if box B containing $\sigma$'s is distant from $\lambda$.

Our algorithm for sorting shell pairs is as follows. First we divide the AO set into two categories according to their exponent values. One set contains all AOs with large exponents (tight basis functions), the second one, all remaining AOs. This way we can take advantage of the better sparsity of the more compact basis functions and isolate the diffuse ones. Further subdivision, e.g., very tight functions, moderately diffuse, and very diffuse AOs, is also possible. The next two loops are over boxes. Consecutive boxes are adjacent, which improves the sparsity pattern. The inner two loops sort the $\nu\sigma$ pairs inside the boxes according to center locations (atoms). The outcome of the sorting routine is an array with the pairs sorted and additional arrays which indicate the box and atom borders and the number of pairs contained in pairs of atoms and pairs of boxes. In the next step, these arrays are used to determine the $\nu\sigma$ batches. Usually we want to cut the $\nu\sigma$ set either at the shell pair, or better, atom, or, even better, box boundaries. Our goal is to have as large batches as possible, within the memory allocation limits. Large batches increase matrix multiplication efficiency and I/O throughput. Taking all the above into account, the "cut" and batch creation takes place if the current batch is big enough or if a forthcoming batch can be large. The cuts are allowed on the atom or box borders only, with the latter preferred.

In this way, we can utilize at least 50% of the maximum sparsity by eliminating empty rows. The average size of a $\nu\sigma$ stripe is about 100−200, depending on the system and the basis set used. The row sparsity is about the same as if we divided the matrix by shell pairs, which yields much smaller $\nu\sigma$ stripes (it can potentially be just 1 for an ss pair, 3 for an sp pair, etc.).

Parallelization of the EEO part follows the scheme described previously.[11] With Array Files as the I/O software,[27] all nodes share uniform access to network files, i.e., they can access the same data using a given file and record number. Parallelization is performed over batches of $\mu\lambda$ pairs. If the available memory allows, we divide all $\mu\lambda$ pairs into as many batches as the number of slaves. If there is insufficient memory for this, the number of batches is chosen as a multiple of the number of slaves. This guarantees good load balancing. Every slave processes its own batch in essentially the same way as in the single processor program, but uses Array Files I/O routines in place of local I/O. All data are read or written via network communication. We do not use any local caching of the results or amplitudes, as local disk I/O rates are commensurate with network rates.

## 4. Perturbative Triples Implementation and Parallelization

The key part of the perturbative triples correction to the CCSD energy is the calculation of the following quantity (see e.g.refs 23, 36 or 37):

$$W^{abc}_{ijk} = P^{abc}_{ijk}\left(\sum_d \mathbf{T}^{ad}_{ij}(ck|bd) - \sum_l \mathbf{T}^{ab}_{il}(ck|lj)\right) \qquad (5)$$

Here $P$ is the permutation operator, which performs simultaneous permutation of indices *ijk* and *abc*, generating a sum of six terms, each of the above form. The indices *ijkl* denote Hartree−Fock occupied canonical molecular orbitals, *abcd*, the canonical virtuals.

$W$ is combined into the (T) energy:

$$E(T) = \sum_{a \geq b \geq c} (2 - \delta_{ab} - \delta_{bc}) \sum_{ijk} W_{ijk}^{abc} (4W_{ijk}^{abc} + W_{kij}^{abc} + W_{jki}^{abc} - 2W_{kji}^{abc} - 2W_{ikj}^{abc} - 2W_{jik}^{abc}) / D_{ijk}^{abc} \quad (6)$$

Here $D$ is a standard perturbational energy denominator. Equation 5 requires about $C(n^4V^3 + n^3V^4)$ floating point operations, but eq 6 needs only $En^3V^3$ ($C$ and $E$ are constants, $V$ is the number of virtual orbitals). For simplicity, we have omitted contributions from singles amplitudes in eq 6, which corresponds to the [T] correction of Urban et al.[8] Singles do not add significantly to the numerical effort, and therefore, our discussion will focus on the doubles part only. We have to calculate and store $W^{abc}$ only for a given *abc* triplet and all *ijk*; after calculating the energy contribution of this set it can be deleted and a new *abc* triplet processed. This requires only $n^3$ memory locations, allowing the program to calculate a number of *abc* triplets simultaneously. This is advantageous because the integrals and amplitudes in eq 5 can be reused with proper blocking, significantly reducing the I/O overhead.

Equation 6 is often written in an alternative form:[23,36]

$$E(T) = \sum_{i \geq j \geq k} (2 - \delta_{ij} - \delta_{jk}) \sum_{abc} W_{abc}^{ijk} (4W_{abc}^{ijk} + W_{bca}^{ijk} + W_{cab}^{ijk} - 2W_{cba}^{ijk} - 2W_{acb}^{ijk} - 2W_{bac}^{ijk}) / D_{ijk}^{abc} \quad (7)$$

In this form, one has to calculate $W^{ijk}$ for a given *ijk* triplet and all *abc*, requiring $V^3$ storage. This severely limits the size of systems which can be treated. For example, for 1000 virtual orbitals, we need 8 GB of fast memory to store the partial result. Of course, for a small system, eq 7 vectorizes better, as the matrices are larger. This improves the flop rate for small dimensions, although the effect saturates at a dimension of a few hundred. Notice that the form of energy expression used (eqs 6 or 7) results in different algorithms for the calculation of $W$ (eq 5) and, thus, changes the whole program design.

We have implemented triples using both algorithms. Although the matrix operations as defined by eq 7 are significantly faster for smaller systems, the gain is largely offset by increased I/O. In addition, on our PC cluster (4 GB RAM per node), it limits the calculation to about 500 virtual orbitals. Our final code uses the slower formulation in eq 6, but taking into account memory savings, I/O reduction due to blocking, and ease of implementation, it is more efficient, particularly for big systems. The pseudocode for our algorithm is shown in Figure 3. Our conclusion about the relative merit of eq 6 vs eq 7 is the opposite of what Rendell et al.[23] arrived at using the Cray supercomputers. They argued that eq 7 vectorizes better and that memory will not be a problem in the future. However, CPU speed has increased faster than RAM memory size, making the latter the bottleneck for distributed memory workstations.

In order to explain how $W^{abc}$ is calculated, we need to define a new quantity:

$$X_{ijk}^{abc} = \sum_{d} \mathbf{T}_{ij}^{ad} (ck|bd) - \sum_{l} \mathbf{T}_{il}^{ab} (ck|lj) \quad (8)$$

This allows us to cast eq 5 in a more explicit form:

$$W_{ijk}^{abc} = X_{ijk}^{abc} + X_{jki}^{bca} + X_{kij}^{cab} + X_{kji}^{cba} + X_{ikj}^{acb} + X_{jik}^{bac} \quad (9)$$

```
Loop over batches of A                    (A – batch of virtual indices)
  loop over batches of B ≤ A
    loop over batches of C ≤ B

      read & store integrals needed for all indices a ∈ A, b ∈ B, c ∈ C
      read & store amplitudes needed for all indices a ∈ A, b ∈ B, c ∈ C

      loop over all virtuals a ∈ A
        loop over all virtuals b ∈ B
          loop over all virtuals c ∈ C

            if (a = b and b = c) exit loop
            calculate W^abc (slow)
            add W^abc contribution to energy E (fast, extensively cached operation)

          endloop c
        endloop b
      endloop a

    endloop C
  endloop B
endloop A
```

***Figure 3.*** Outline of the main "driver" loop of the (T) contribution.

Zero $W^{abc}$ array. An array has indices $i, j, k$.

Call $X$ for $a, b, c$, add result to $W^{abc}$ array
Rearrange $W^{abc}$ array (swap 2nd and 3rd index)

Call $X$ for $a, c, b$, add result to $W^{abc}$ array
Rearrange $W^{abc}$ array (swap 1st and 2nd index)

Call $X$ for $c, a, b$, add result to $W^{abc}$ array
Rearrange $W^{abc}$ array (swap 2nd and 3rd index)

Call $X$ for $c, b, a$, add result to $W^{abc}$ array
Rearrange $W^{abc}$ array (swap 1st and 2nd index)

Call $X$ for $b, c, a$, add result to $W^{abc}$ array
Rearrange $W^{abc}$ array (swap 2nd and 3rd index)

Call $X$ for $b, a, c$, add result to $W^{abc}$ array

Rearrange $W^{abc}$ array (swap 1st and 2nd index, this call recovers original $ijk$ ordering).

***Figure 4.*** $W^{abc}$ construction scheme.

It is sufficient to have only one subroutine for the calculation of $X$. It is called six times for every index triple $a \geq b \geq c$ in eq 9. The calls are interleaved by another subroutine which interchanges elements of $W$ before adding the next component $X$. Swapping the destination $W$ array, instead of adding $X$ with permuted *ijk* indices simplifies the program code, as we do not need to write six different subroutines to calculate each term in eq 9. Instead, we call the same subroutine and add partial result to $W$ with two indices swapped. All operations on $W$, including energy contribution calculation are very fast, as the array $W$ is usually fairly small (2 MB for 63 occupied orbitals) and fits in the cache of most processors. With the algorithm based on the formula eq 7, the calculation of the final energy contribution becomes more expensive, because $W^{ijk}$ is large and requires access to main memory, instead of fully cached operations for $W^{abc}$. The difference in the time of evaluation of the final energy expression can be up to an order of magnitude. Figure 4 shows an outline of the $W$ calculation part of the triples code.

Vectorization inside the $X$ subroutine is explained in Figure 5. The figure shows the first term of eq 8; the same method is applied to the second term. In Figure 5, the $\mathbf{g}^{bc}$ matrix is a matrix of integrals, $(ck|bd) = (\mathbf{g}^{bc})_{dk}$. The $W^{abc}$ array is obtained by a single matrix multiplication.

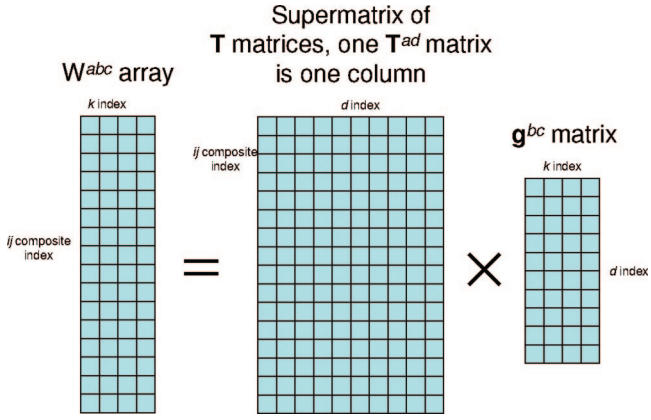



***Figure 5.*** Triples contribution (eq 8, first term), as a matrix product.

The simplicity of accessing global data through the Array Files system allows a simple master-slave self-scheduling parallel algorithm. Slaves ask the master for work and get batches of *ABC* triples to work on. When they are finished they store the partial energy in a local variable and request more work from the master. If the master responds with a "no work" message, the slaves exit their triple substitution subroutine. At the very end a global energy sum is performed over all slaves.

## 5. Discussion

We describe an improved parallel implementation of CCSD(T) and related methods. The most expensive computational steps, the calculation of the external exchange operators, and the perturbational triples energy are formulated in terms of dense matrix multiplications that take the maximum advantage of modern CPUs. Our new EEO algorithm exploits the sparsity of the atomic orbital integrals, and its communication requirements have been much reduced compared to our previous program. Benchmarks have been presented with over 1500 basis functions and 68 atoms on 2−16 computing nodes.

The main difference between our CC code and the NWChem[15] and GAMESS[17] implementations is that our program makes extensive use of disk storage and, therefore, can perform large calculations on modest size clusters. ACES II[16] uses replicated data structures. This poses limitations on the calculations by the size of the local storage.

***Table 2.*** Perturbative Triples Timings for a Test Set of Molecules

| | (nodes/proc) min | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| molecule | 1/1 | 2/2 | 4/4 | 4/8 | 8/8 | 8/16 | 16/16 | 16/32 |
| aspirin | 1196 | 601.6 | 292.5 | 157 | 151.3 | 82 | 76.7 | 43 |
| sucrose | | | 12436 | | 5981.9 | 3216 | 3034 | 1594 |
| $(glycine)_{10}$ | | | | | | | | 14319 |
| Si-ring | | | | | | | | 4168 |
| AMP | | | | | | | | 16740 |

Our code shares the general design philosophy of MOLPRO.[14] However, we use Array Files instead of global arrays. This has the merit of enabling small memory workstations to perform large calculations, at the cost of modest overhead. The I/O can be significantly reduced by employing various blocking strategies, as we have discussed already in the EEO and (T) part. The TCE (Tensor Contraction Engine)[30] also employs blocking techniques but does not address sparsity. The TCE can automatically generate code. This restricts it, in its present form, to an orthogonal spin−orbital formulation. As our flop counts show, the hand-coded version achieves substantially better performance, e.g., in the external exchange part, than the automatically generated code. However, automatic code generation is probably unavoidable for higher order CC programs.

The latest CC parallel code, ACES III,[18] uses a novel parallel programming language and automatic blocking of arrays. Preliminary results show very good parallel scaling.

## 6. Sample Timings

Table 1 presents QCISD timings obtained for a series of molecules of different size. All calculations were performed on a 16-node cluster, connected via gigabit ethernet. Every node of this cluster has a dual core processor (64-bit Intel Pentium IV Prescott, 3.0 GHz), 4 GB of RAM memory, and a 300 GB RAID0 (striped) array of disk scratch storage. Our test set includes a range of molecular sizes and symmetries: aspirin, sucrose, a linear polymer of ten glycine molecules with one symmetry plane, a fragment of a copper-based catalyst with empirical formula $CuAlSi_3O_{12}H_8$, adenosine monophosphate (AMP), and a simple calixarene with empiri-

***Table 1.*** QCISD and Sample CCSD Timings for a Test Set of Molecules

| | | | | | | time per iteration (nodes/proc) min[a] | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| molecule[b] | sym | basis | nbf[c] | corr[d] | iter[e] | 2/2 | 4/4 | 8/8 | 8/16 | 16/16 | 16/32 |
| aspirin | $C_1$ | 6-311G** | 282 | 34 | 16 | 18.9 | 11.42 | 6.4 | 4.8 | 3.8 | 3.1 |
| sucrose | $C_1$ | 6-311G** | 455 | 68 | 12 | 365.0 | 197.4 | 104 | 61.8 | 53.7 | 39.3 |
| $(glycine)_{10}$ | $C_s$ | 6-31G* | 638 | 114 | 14 | 2133 | 1099.0 | 590 | 374 | 271 | 203 |
| Si-ring | $C_1$ | VTZ3P[f] | 664 | 53 | 18 | 826 | 443.3 | 226 | 127.7 | 123.5 | 74 |
| AMP | $C_1$ | def2-tzvp[g] | 803 | 63 | 14 | 2384 | 1166 | 605.2 | | 325 | 199 |
| Calix | $C_{2h}$ | cc-pVTZ[h] | 1528 | 92 | 11[i] | | | 3506 | | 1723 | |
| | | | | | sample CCSD timings[j] | | | | | | |
| aspirin | $C_1$ | 6-311G** | 282 | 34 | 14 | 31.7 | 16.6 | 10.4 | 8.6 | 6.1 | 5.6 |
| sucrose | $C_1$ | 6-31G** | 455 | 68 | 12 | 519.2 | 236.2 | 123.0 | 91.2 | 71.7 | 65.4 |

[a] For example, 8/16 means that 16 processes were running on 8 nodes, i.e. 2 processes per node. [b] The test set is described in the Sample Timings section of this paper. [c] Number of basis set functions. [d] Number of correlated orbitals. [e] Number of iterations, until max residue $R < 10^{-6}$ and energy difference $\Delta E < 10^{-6}$. [f] Valence triple-$\zeta$ + triple polarization.[38] [g] Valence triple-$\zeta$ + polarization.[39] [h] Correlation consistent basis.[40] [i] Only 11 iterations were allowed to complete. [j] Our current integral-direct CCSD implementation requires extra integral recalculation steps, comparing to QCISD.

***Table 3.*** Timings for an External Exchange Evaluation (per Iteration): Comparison of the Old and the New Algorithms

| molecule | nodes | basis set | no. of basis functions | no. of correlated orbitals | EEO evaluation time (min) | |
|---|---|---|---|---|---|---|
| | | | | | new algorithm | old algorithm |
| benzene dimer ($C_1$ symmetry) | 32 | aug-cc-pVQZ | 1512 | 30 | 379 | 1813 |
| $(glycine)_{10}$ | 16 | 6-31G* | 679 | 114 | 72 | 148 |

***Table 4.*** Timing for Triples Evaluation in Parallel Displaced Benzene Dimer

| molecule | nodes | basis set | no. of basis functions | no. of correlated orbitals | time of triples evaluation | CPU efficiency |
|---|---|---|---|---|---|---|
| benzene dimer ($C_{2h}$ symmetry) | 32 | aug-cc-pVQZ | 1512 | 1512 | 3420 | 93% |
| as above | 32 | aug-cc-pVTZ | 828 | 30 | 350 | 91% |

cal formula $C_{32}H_{32}O_4$ and $C_{2h}$ symmetry. Each process was constrained to use no more than 1.6 GB of RAM.

The parallel scaling of QCISD jobs run with only one process per node is satisfactory, and in some cases, superlinear scaling is observed. This is most likely caused by the fact that the amount of I/O per process decreases with increasing number of processes, allowing the operating system (OS) to cache the data more efficiently in the main memory (in our case, the RAM available for disk caching was about 2.4 GB). Comparing calculations involving the same number of processes but a different number of nodes, Table 1 shows that jobs with two processes per node are significantly slower than jobs which use the same number of processor cores but with one process per node. In the former case, two processes must share access to the same disk. An additional effect arises from the deterioration of OS data caching. In our case two processes occupy 3.2 GB of main memory leaving only about 0.8 GB for the OS to use as I/O cache. Running 32 processes on 16 nodes is still faster than running only 16 processes, but the speedup is modest.

Because the amount of memory available impacts the OS data caching performance, it might seem advantageous to use shared memory parallelization (e.g., multithreading) on the same node, instead of running several separate processes per node. In our previous version of the CCSD program, we have used a multithreaded version of matrix multiplication routines.[24] If the choice is between one process per node without threading and one process per node with threading, the latter is usually (not always) faster. However, the efficiency of the threaded routines depends strongly on the size of matrices being multiplied. For very small ($10 \times 10$) matrices, it is 1388 Mflop/s on single processor, but only 62 Mflop/s with threading. This probably arises from the overhead of threads preparation. The threaded and single-threaded libraries perform comparably if the matrix dimension is about 50. For large dimensions, the threaded version outperforms the nonthreaded one almost by the theoretical factor. Our current CCSD version uses nonthreaded matrix multiplication libraries for three reasons. (1) In our blocked implementation, the matrices are relatively small, and the threaded libraries do not reach their theoretical limit. (2) Our integral program is not yet thread-safe. (3) Threading does not remove the I/O bottleneck. The latter is diminished in our current algorithm but, as the timings show, not fully eliminated. Note, however, that threading may improve I/O caching by the OS because one threaded process may occupy less memory than two independent processes running concurrently.

A different picture emerges for the calculation of the perturbative triples contribution (Table 2). Triples are very expensive in terms of CPU usage, but the time spent doing I/O is negligible. The I/O reduction results from an efficient blocking algorithm, such that the data are read in batches and reused as frequently as possible (see Figure 3). For large systems, a single process can fill its 1.6 GB of memory storage with data and use it for about 30 min without needing any I/O. This is the reason why the triples contribution scales almost linearly with the number of CPU cores and differences between jobs involving one and two processes per node are small. The calculation of the triples contribution can efficiently utilize both CPU cores because the data are reused more extensively in the triples algorithm than in the SD part.

In Table 3, the efficiency of the old and new EEO algorithms are compared. In order to demonstrate savings and compare the new results with previous ones,[11] we have switched off the use of symmetry for the benzene dimer. The speedup is dramatic, almost a factor of 5. Because the benzene dimer has very little integral sparsity with the aug-cc-pVQZ basis set, the speedup is almost exclusively due to the reduction in I/O.

Table 4 presents sample timings for one of the biggest (T) calculation we have done so far, the benzene dimer with the aug-cc-pVQZ basis set. This calculation was performed on the Red Diamond supercomputer at the University of Arkansas. It has a configuration similar to our cluster, but the processor clock is faster (3.2 GHz), the local storage is a nonstriped SCSI disk, and the network topography is such that groups of nodes share the same gigabit ethernet connection. This results in lower efficiency, due to the less efficient network and the slower speed of local storage (a single disk).

**Acknowledgment.** This work was supported by the National Science Foundation under grant number CHE-0515922 and by the Mildred B. Cooper Chair at the University of Arkansas. Acquisition of the Red Diamond

supercomputer was supported in part by the National Science Foundation under award number MRI-0421099.

## References

(1) Bartlett, R. J.; Purvis, G. D., III *Int. J. Quantum Chem.* **1978**, *14*, 561.

(2) Pople, J. A.; Krishnan, R.; Schlegel, H. B.; Binkley, J. S *Int. J. Quantum Chem.* **1978**, *14*, 545.

(3) Taylor, P. R.; Bacskay, G. B.; Hush, N. S.; Hurley, A. C. *J. Chem. Phys.* **1978**, *69*, 1971.

(4) Sinanoğlu, O. *J. Chem. Phys.* **1962**, *36*, 706.

(5) Číček, J. *J. Chem. Phys.* **1969**, *45*, 4256.

(6) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968.

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

(8) Urban, M.; Noga, J.; Cole, S. J.; Bartlett, R. J. *J. Chem. Phys.* **1985**, *83*, 4041–4048.

(9) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968.

(10) *PQS*; version 3.2, Parallel Quantum Solutions: Fayetteville, AR.

(11) Janowski, T.; Ford, A. R.; Pulay, P. *J. Chem. Theory Comput.* **2007**, *3*, 1368–1377.

(12) Meyer, W. *J. Chem. Phys.* **1973**, *58*, 1017–1035.

(13) Pulay, P.; Saebo, S. *Theor. Chim. Acta* **1985**, *69*, 357–368.

(14) 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.; McNicholas, S. J.; Manby, F. R.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Rauhut, G.; Schtz M. Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thornteinsson, T.; Werner, J.-J. *MOLPRO*, a package of ab initio programs designed by Werner H.-J. and Knowles, P. J.; version 2002.1; 2002.

(15) 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. Commun.* **2000**, *128*, 260–283.

(16) Harding, M. E.; Metzroth, T.; Gauss, J.; Auer, A. A. *J. Chem. Theory Comput.* **2007**, *4*, 64–74.

(17) Olson, R. M.; Bentz, J. L.; Kendall, R. A.; Schmidt, M. W.; Gordon, M. S. *J. Chem. Theory Comput.* **2007**, *3*, 1312–1328.

(18) Lotrich, V.; Flocke, N.; Yau, A.; Perera, A.; Deumens, E.; Bartlett, R. J. ACES III: Parallel Implementation of Electronic Structure Energy, Gradient and Hessian Calculations. In *48th Sanibel Symposium*; St. Simons Island, GA, Feb 21–26, 2008.

(19) Kobayashi, R.; Rendell, A. P. *Chem. Phys. Lett.* **1996**, *265*, 1–11.

(20) Koch, H.; de Meras, A. S.; Helgaker, T.; Christiansen, O. *J. Chem. Phys.* **1996**, *104*, 4157.

(21) Rendell, A. P.; Guest, M. F.; Kendall, R. A. *J. Comput. Chem.* **1993**, *14*, 1429.

(22) Rendell, A. P.; Lee, T. J.; Lindh, R. *Chem. Phys. Lett.* **1992**, *194*, 84.

(23) Rendell, A. P.; Lee, T. J.; Komornicki, A. *Chem. Phys. Lett.* **1991**, *178*, 462–470.

(24) Goto, K.; van de Geijn, R. *ACM Trans. Math. Software*, submitted for publication.

(25) Meyer, W. *J. Chem. Phys.* **1976**, *64*, 2901.

(26) Pulay, P.; Saebo, S.; Meyer, W. *J. Chem. Phys.* **1984**, *81*, 1901.

(27) Ford, A. R.; Janowski, T.; Pulay, P. *J. Comput. Chem.* **2007**, *28*, 1215–1220.

(28) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic-Structure Theory*, first ed.; John Wiley & Sons Ltd.: New York, 2000; Chapter 15, pp 817−883.

(29) Saebo, S.; Pulay, P. *J. Chem. Phys.* **1987**, *86*, 914–922.

(30) Hirata, S. *J. Phys. Chem.* **2003**, *107*, 9887–9897.

(31) Schütz, M.; Lindh, R.; Werner, H.-J. *Mol. Phys.* **1999**, *96*, 719–733.

(32) Janowski, T.; Pulay, P. *Chem. Phys. Lett.* **2007**, *447*, 27–32.

(33) Schütz, M.; Werner, H.-J. *J. Chem. Phys.* **2001**, *114*, 661–681.

(34) Baker, J.; Pulay, P. *J. Comput. Chem.* **2002**, *23*, 1150–1156.

(35) Pulay, P.; Saebo, S.; Wolinski, K. *Chem. Phys. Lett.* **2001**, *344*, 543–552.

(36) Lee, T. J.; Rendell, A. P.; Taylor, P. R. *J. Phys. Chem.* **1989**, *94*, 5463–5468.

(37) Schütz, M.; Werner, H.-J. *Chem. Phys. Lett.* **2000**, *318*, 370–378.

(38) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571.

(39) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

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

CT800142F

JCTC Journal of Chemical Theory and Computation

# On the Bonding of Selenocyanates and Isoselenocyanates and Their Protonated Derivatives

Cristina Trujillo,[†] Otilia Mó,[†] Manuel Yáñez,*[,†] and Bernard Silvi*[,‡]

*Departamento de Química, C-9. Universidad Autónoma de Madrid. Cantoblanco. 28049-Madrid. Spain, and Laboratoire de Chimie Théorique, UMR-CNRS-7616, Université de Paris 06, F-94000 Ivry, France*

Received May 21, 2008

**Abstract:** The structure, bonding, and protonation of NCSeX and XNCSe (X = Me, F, Cl, Br) derivatives has been investigated at the B3LYP/6−311++G(3df,2p)//B3LYP/6−31+G(d,p) level of theory. Three different approaches, namely, ELF, AIM, and NBO indicate that three main factors are responsible for the enhanced stability of the selenocyanates with respect to the isoselenocyanates when the substituents are halogens, whereas for alkyl substituents, it is the other way around: (a) the Se−X (X = F, Cl, Br) bonds are much stronger than the Se−X (X = Me); (b) the N−X (X = F, Cl, Br) bonds are much weaker than the N−X (X = Me) ones; (c) on going from the selenocyanates to the isoselenocyanates, when the substituents are halogen atoms, there is a significant weakening of the CN bond, which becomes essentially a double bond, whereas upon methyl substitution the CN bond retains its triple bond character. The same stability trends are observed for the corresponding N-protonated species. More importantly, the calculated stability differences are rather similar to those obtained for the neutral compounds, so that selenocyanates and isoselenocyanates exhibit rather similar basicities in the gas phase. Both types of isomers behave as gas-phase nitrogen bases.

## Introduction

The selenocyanate group, SeCN, as the corresponding sulfur analogue, has the possibility to bond through selenium or nitrogen. As for the thiocyanate group[1,2] both modes of coordination, XNCSe and NCSeX, are known when X is an alkyl group.[3−8] As a matter of fact alkyl isoselenocyanates are commonly used in organic synthesis,[9,10] and many of these derivatives have been spectroscopically characterized.[4,8] Recently, the generation of the isocyanoselenic acid radical cation, $HNCSe^{+\bullet}$, from a dissociative ionization of selenourea in the gas phase has been reported.[11] However, the only known compounds in the gas phase when the substituent is a halogen atom, again in parallel to the sulfur analogues,[12−15] are those in which the halogen is attached to the Se atom,[16,17] whereas those in which the halogen is attached to the N atom are not experimentally known. These findings and the fact that, although selenium derivatives have received significant attention in the past decade,[18−48] many questions on their chemistry and bonding are still open, prompted us to undergo a study on the bonding and relative stabilities of substituted selenocyanates and isoselenocyanates, when the substituent is a methyl group, as the simplest case of alkyl group, and when the substituent is F, Cl, and Br.

We have also considered it of interest to investigate the effect that the protonation of these systems have on their bonding and, as a consequence, on the relative stabilities of selenocyanates with respect to isoselenocyanates.

## Computational Details

The geometries of the selenocyanates and isoselenocyanates included in this study, and those of their protonated species have been optimized using the B3LYP density functional theory (DFT) approach associated with a 6−31+G(d,p) basis set expansion. This method, which includes Becke's three parameter nonlocal hybrid exchange potential[49] and the

* To whom correspondence should be addressed. E-mail: manuel.yanez@uam.es (M.Y.); silvi@lct.jussieu.fr (B.S.).

[†] Universidad Autónoma de Madrid.

[‡] Université de Paris 06.

10.1021/ct800178x CCC: $40.75 
Published on Web 09/18/2008

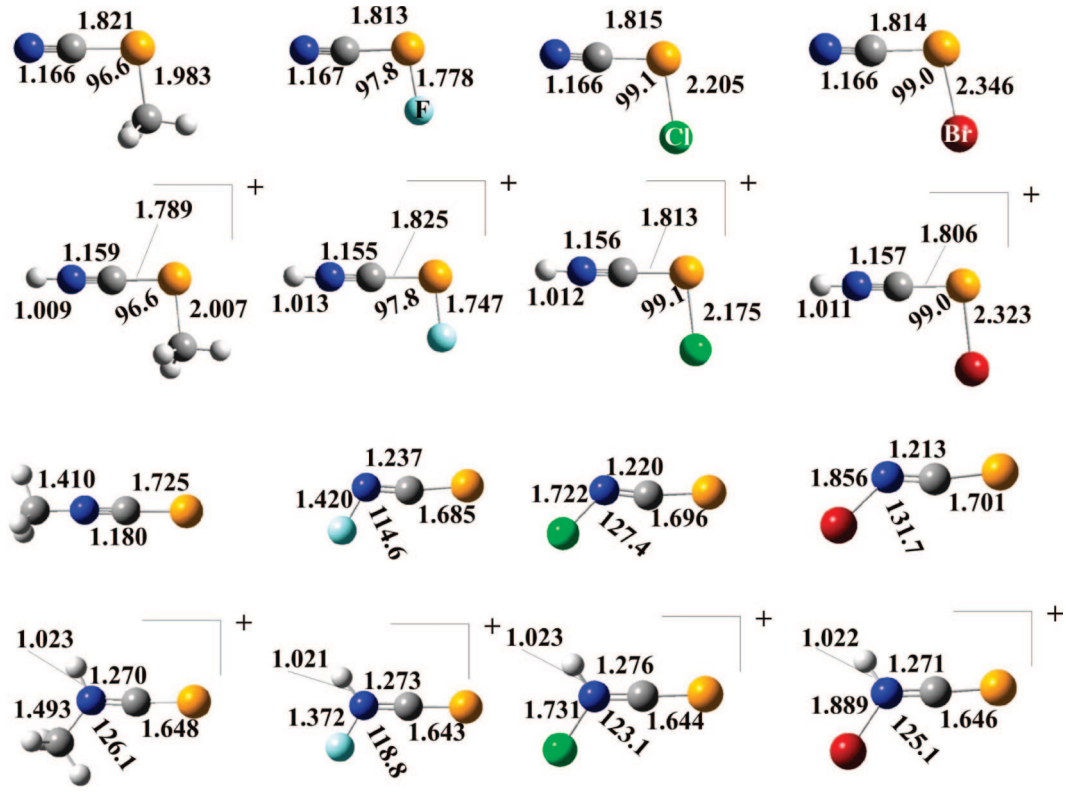



***Figure 1.*** B3LYP/6−31+G(d,p) optimized geometries for selenocyanates and isoselenocyanates and their N-protonated species. Bond lengths are in angstroms, and bond angles are in degrees.

nonlocal correlation functional of Lee, Yang, and Parr,[50] has been successfully used for the treatment of other Se-containing compounds.[51−53] To get reliable relative energies, final energies were obtained in single-point calculations using a 6−31++G(3df,2p) basis set. Harmonic vibrational frequencies were evaluated at the same level of theory used for the geometry optimizations to classify the stationary points found as local minima and to evaluate the thermal corrections necessary to obtain the free energies at 298.2 K.

The bonding was analyzed primarily by using the Becke and Edgecombe electron localization function[54] (ELF) topological approach.[55] ELF has been originally conceived as a local measure of the Fermi hole curvature around a reference point within the Hartree−Fock approximation, another interpretation in terms of "local excess of kinetic energy due to Pauli principle" was further proposed by Savin et al.[56] legitimating the calculation of the function with Kohn−Sham orbitals. More recently, it was shown that the ELF kernel can be rigorously derived by considering the number of same spin pairs contained in a sample around the reference point.[57,58] Thanks to a cosmetic Lorentz transform, ELF is confined in the [0,1] interval; 1 corresponds to regions dominated by an opposite spin pair or by a single electron, whereas low values are found at the boundaries between such regions. The partition of the molecular space is carried out by the gradient dynamical technique, which yields basins of attractors closely related to Gillespie's electronic domains which are a generalization of the ideas of Lewis. The valence shell of a molecule consists of two types of basin: polysynaptic basins (generally disynaptic), which belong to two atomic valence shells, and the monosynaptic ones, which belong to only one valence shell and which qualitatively correspond to nonbonding valence density. The valence basins are labeled by V followed by a list of the atomic symbols of the centers of the valence shells, that is, V(A) and V(A,B) for a monosynaptic and a disynaptic basin. The basin populations and the associated covariance matrix are calculated by integration of the one electron and pair densities over the volume of the basins enabling a phenomenological interpretation of the population analysis in terms of the superposition of mesomeric structures.[59]

ELF grids and basin integrations have been computed with the TopMod package.[60] The ELF isosurfaces have been visualized with the Amira 3.0 software.[61] The atoms in molecules (AIM) theory,[62] based in a topological analysis of the electron density, is a complementary tool for the investigation of bonding characteristics. In the framework of this approach, we have located the bond critical points (BCP) of each compound because the electron density at these points offers quantitatively valid information on the strength and the multiple bond character of the linkage. Moreover insights on the delocalization in terms of delocalization indexes[63] can be obtained by a covariance analysis of the atomic populations. This question can be also investigated by evaluation of the Wiberg bond orders[64] in the framework of the natural bond orbital (NBO) approach,[65] which also permits to estimate the weights of the different resonant structures contributing to the stability of the system through the natural resonance theory (NRT) as implemented in the NBO-5.0 suite of programs.[66] On the other hand, some further insight into details of the bonding can be gained by means of a second-order perturbation analysis of the Fock matrix, which usually provides information on the interactions between occupied and empty molecular orbitals. All

***Table 1.*** Proton Affinities (PA, kJ mol$^{-1}$), Gas-Phase Basicities (GB, kJ mol$^{-1}$), and Relative Free Energies ($\Delta\Delta$G, kJ mol$^{-1}$)

| compound | PA | GB | $\Delta\Delta G$ (neutral) | $\Delta\Delta G$ (protonated) |
|---|---|---|---|---|
| NCSeF | 745.8 | 714.6 | 0.0 | 0.0 |
| FNCSe | 751.2 | 719.9 | 173.8 | 168.5 |
| NCSeCl | 766.1 | 734.9 | 0.0 | 0.0 |
| ClNCSe | 757.3 | 725.6 | 100.3 | 109.6 |
| NCSeBr | 775.8 | 744.5 | 0.0 | 0.0 |
| BrNCSe | 765.3 | 733.9 | 91.8 | 102.5 |
| NCSeMe | 813.6 | 782.9 | 0.0 | 0.0 |
| MeNCSe | 791.7 | 759.0 | −23.3 | 0.6 |

***Table 2.*** ELF Valence Basin Populations for Selenocyanates and Isoselenocyanates

| CNSeX | | | | | | |
|---|---|---|---|---|---|---|
| X | V(Se) | V(Se,X) | V(Se,C) | V(C,N) | V(N) | V(X) |
| F | 2 × 2.35 | | 2.19 | 2 × 2.16 | 3.24 | 2 × 3.79 |
| Cl | 2 × 2.38 | 0.90 | 2.16 | 2 × 2.17 | 3.25 | 2 × 3.33 |
| Br | 2 × 2.42 | 0.92 | 2.17 | 2 × 2.16 | 3.26 | 2 × 3.44 |
| $CH_3$ | 2 × 2.40 | 1.43 | 2.15 | 2 × 2.30 | 3.29 | |

| XCNSe | | | | | | |
|---|---|---|---|---|---|---|
| X | V(Se) | V(N,X) | V(Se,C) | V(C,N) | V(N) | V(X) |
| F | 2 × 2.69 | 0.51 | 1.60<br>1.51 | 2 × 1.56 | 3.12 | 2 × 3.43 |
| Cl | 2 × 2.77 | 0.98 | 1.50<br>1.56 | 2 × 1.58 | 2.98 | 2 × 3.21 |
| Br | 2 × 2.79 | 0.92 | 1.35<br>1.64 | 2 × 1.60 | 3.0 | 2 × 3.33 |
| $CH_3$ | 5.87 | 1.82 | 2.89 | 3 × 1.85 | | |

these bonding analysis have been carried out at the same level used for the geometry optimization.

## Results and Discussion

The optimized geometries of the different compounds under investigation and their most stable protonated species are given in Figure 1. The calculated total energies are summarized in Table S1 of the Supporting Information.

**Relative Stability and Protonation.** Although protonation at the halogen, selenium, and nitrogen atoms have been considered, in all cases, for both selenocyanates and isoselenocyanates, nitrogen protonation is by far the most favorable process, so in what follows we will refer exclusively to the nitrogen protonated species. The calculated proton affinity and gas-phase basicity, defined as the negative of the enthalpy and free energy for reaction 1, respectively, are given in Table 1. This table also includes the relative stability of the isoselenocyanates with respect to the analogue selenocyanates.

$$B(g) + H^+(g) \rightarrow BH^+(g) \quad (1)$$

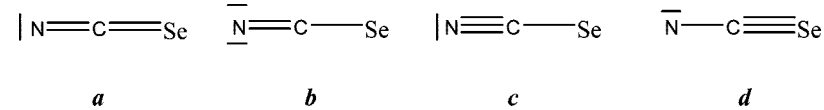
The first conspicuous fact is that halogen selenocyanates are systematically more stable than the corresponding isoselenocyanate isomer, similar to what has been reported before for the corresponding thio-derivatives.[15,17] The opposite is found for the methyl derivative, where the isoselenocyanate isomer is predicted to be 23.3 kJ mol$^{-1}$ more stable than the selenocyanate one. The stability gap decreases in the order F > Cl > Br. Importantly, the same stability trends are observed for the corresponding N-protonated species, with the exception of the methyl derivative. For the halogen derivatives, the stability gap between isoselenocyanate and selenocyanate forms is rather similar to that calculated for the neutral forms, and as a consequence, both isoselenocyanate and selenocyanate halogen derivatives exhibit rather similar proton affinities and gas-phase basicities. For the methyl derivatives, both protonated forms are almost degenerate, and since the neutral isoselenocyanate is ∼23 kJ mol$^{-1}$ more stable than the selenocyanate, its PA is about 23 kJ mol$^{-1}$ smaller.

**Bonding Analysis of the Neutral Compounds.** To gain some insight into the origin of the stability trends discussed in the previous section, it is necessary to carry out an analysis of the bonding similarities and dissimilarities of both families of compounds. Table 2 presents the valence basin populations of both series of compounds.

The bonding of the CNSe moiety in both families of compounds could be described, in principle, in terms of the superposition of the four mesomeric structures depicted in Scheme 1.

The values in Table 2, indicate that for the selenocyanates, the characteristics of the CN and CSe bonds are almost independent of the substituent and point to a significant contribution of the mesomeric form c, as reflected in the population of the V(C,N) basin. The population of the V(C,Se) basin clearly indicates that the contribution of mesomeric form **a**, although small is not negligible. In agreement with this description, the weight of these mesomeric forms estimated by means of the NRT approach are about 87% and 11% for forms **c** and **a**, respectively, with negligible contributions from form **b**. The NRT result is not apparently fully consistent with the ELF population of the V(N) basin which requires that the weights of formally ionic structures such as **b** to be greater than 50%. However, the NRT mesomeric structures being of the Coulson−Fisher type are polarized and therefore implicitly account for the ionic component. Also consistent with the ELF populations, the weights of these forms are practically independent of the substituent, as well as the Wiberg bond orders, that for the CN bond is 2.81 and for the CSe bond, 1.12. As expected, it can be observed that in the selenocyanate derivatives the ionic character of the Se−X (X = F, Cl, Br) bond decreases

***Scheme 1***



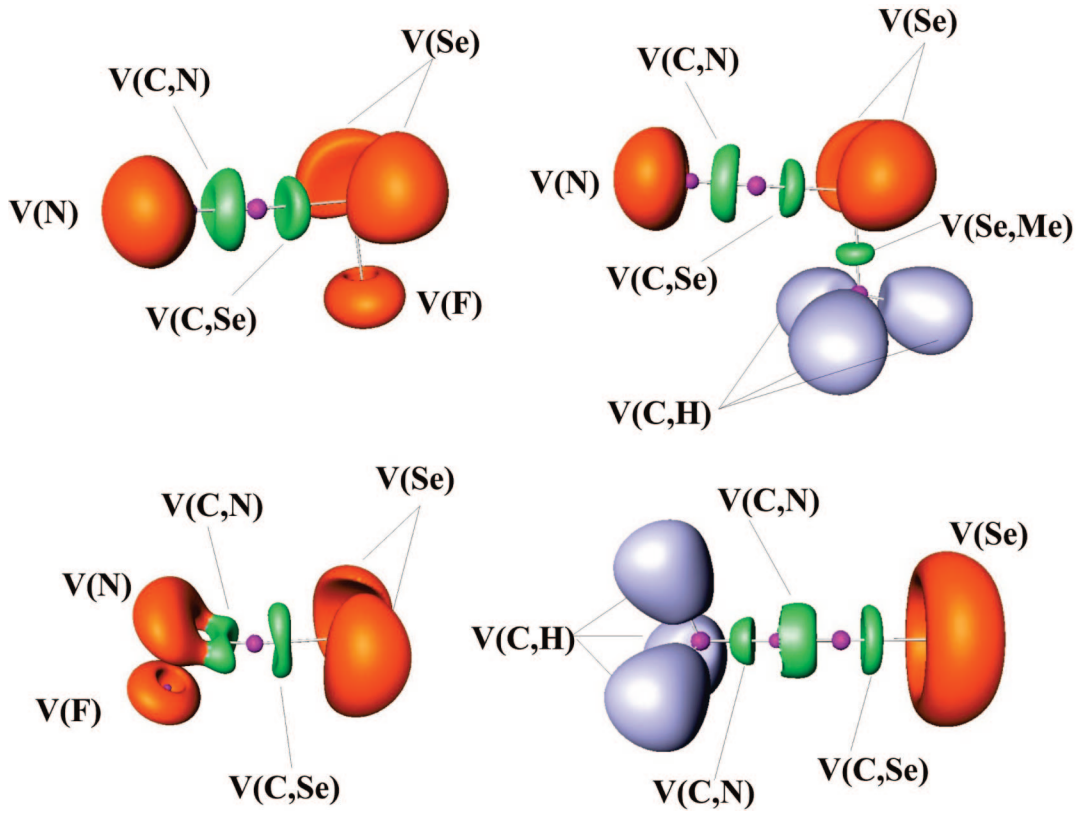



***Figure 2.*** Three-dimensional representations of ELF isosurfaces with ELF = 0.80 for the F and Me derivatives of selenocyanates and isoselenocyanates. Blue lobes correspond to V(C,H) basins, orange lobes correspond to V(N), V(Se), and V(F) basins associated with N, Se, and F lone-pairs. Green lobes correspond to V(C,Se) and V(C,N) basins.

with the halogen atom electronegativity as testified by the V(Se,X) basin population. Coherently, the Wiberg bond orders for these linkages increase slightly on going from F to Br (0.73, 0.95, 0.98), whereas the electron density at the BCP decreases (0.141, 0.110, 0.099 au, respectively) as a consequence of the size increase of the susbtituent. The same behavior is observed for the delocalization indexes (0.96, 1.16, 1.22). The constancy along the series of the characteristics of the CN and CSe bonds are also mirrored in the constant values of the electron densities at the BCPs, which for the CSe bond vary from 0.173 to 0.171, whereas for the CN bond, it is constant and equal to 0.457. For these bonds, the delocalization indexes also remain nearly constant, ~1.18 for CSe and ~2.34 for CN.

There are significant changes on going from the selenocyanates to the corresponding isoselenocyanates. The first important change affects to the CSe bond, for which two connected disynaptic basins, pointing to a certain C═Se double bond character (see Figure 2) are located with a population about 1 $e^-$ greater than for the selenocyanate isomers. The delocalization indexes, close to 2.0, are consistent with this interpretation.

The population of the V(C,N) disynaptic basins, as well as the CN delocalization index, also clearly decreases. Both changes point to a significant participation of mesomeric form **a** and, accordingly, to a parallel decrease of the participation of form **c**. A NRT analysis, actually shows that on going from the selenocyanate to the isoselenocyanate derivative, the weight of form **a** dramatically increases (from 11% to 61% for the F derivative and to 52% for the Cl and Br derivatives). Concomitantly, the weight of the mesomeric form **c** decreases from 87% to 37% in the case of F derivative and to 46% in the case of the Cl and Br derivatives.

The aforementioned changes in the bonding patterns are reflected in both the electron densities at the BCPs and at the molecular force field. As a matter of fact, the electron density at the CN BCP decreases from 0.457 au in the selenocyanates to 0.411 au in the iso-derivatives. The electron density at the CSe BCP, which for the selenocyanates was 0.172 au, on average, becomes 0.197 au in average for the isoselenocyanates. As far as the stretching frequencies are concerned, while for selenocyanates the CN and CSe stretching modes appear at 2255 and 545 $cm^{-1}$, respectively, for the iso-derivatives, they appear in the regions of 1980 and 840 $cm^{-1}$, respectively, the latter being strongly coupled with the NX stretching mode.

Bonding changes are also significant for the Me derivative, but while for the halogen derivatives form **a** becomes dominant, for the Me derivative, form **c** still weights more than form **a**, even though their participation is rather similar (56% and 43%, respectively). This is consistent with the fact that, whereas in the methyl derivative the XNCSe skeleton is linear, for the halogen derivatives it is not and the XCN angle varies from 114.6° to 131.7° on going from the F- to the Br-substituted compound. Once more, the electron densities at the BCPs are in agreement with the previous discussion. The electron density at the CN BCP decreases by about 0.05 au, whereas that the CSe BCP increases by about 0.03 au.

***Table 3.*** Valence Basin Populations for N-Protonated Selenocyanates and Isoselenocyanates Derivatives

| X | V(Se) | V(Se,X) | V(Se,C) | V(C,N) | V(N) | V(N,H) | V(X) |
|---|---|---|---|---|---|---|---|
| F | 2 × 2.26 | | 2.52 | 2.81 | | 2.37 | 2 × 3.77 |
| | | | | 2.25 | | | |
| Cl | 2 × 2.32 | 0.94 | 2.53 | 2.35 | | 2.36 | 2 × 3.28 |
| | | | | 2.74 | | | |
| Br | 2 × 2.33 | 0.99 | 2.54 | 2.29 | | 2.35 | 6.76 |
| | | | | 2.82 | | | |
| $CH_3$ | 2 × 2.30 | 1.35 | 2.59 | 2.15 | | 2.33 | |
| | | | | 2.98 | | | |

| X | V(Se) | V(N,X) | V(Se,C) | V(C,N) | V(N) | V(N,H) | V(X) |
|---|---|---|---|---|---|---|---|
| F | 2 × 2.37 | 0.66 | 1.81 | 1.33 | 0.76 | 2.09 | 2 × 2.22 |
| | | | 1.88 | 1.33 | 0.76 | | 2.25 |
| Cl | 2 × 2.42 | 1.12 | 1.78 | 1.30 | 0.77 | 2.06 | 2 × 2.63 |
| | | | 1.87 | 1.30 | 0.77 | | 2 × 0.53 |
| Br | 2 × 2.44 | 1.12 | 1.73 | 1.32 | 0.80 | 2.02 | 2 × 2.02 |
| | | | 1.87 | 1.32 | 0.80 | | 2.54 |
| $CH_3$ | 2 × 2.47 | 1.63 | 1.76 | 1.36 | 0.59 | 2.04 | |
| | | | 1.81 | 1.36 | 0.59 | | |

**Bonding Analysis of the Protonated Compounds.** The valence basin populations of the N-protonated selenocyanates and isoselenocyanates are summarized in Table 3.

It can be observed that N protonation leaves the number of basins of selenocyanate derivatives unchanged because a density transfer toward V(C,N) is possible since the SeC bond has mostly a single bond character. Hence, following the prescription of the "least topological change",[67] the protonation occurs in the most populated basin provided the number of basins remains constant, in agreement with the fact the N-protonation is strongly favored with respect to X- or Se-protonation. Interestingly, there is also a rather good linear correlation between the calculated proton affinities and the population of the V(N) basin, which is the one directly involved in the protonation process (see Figure 3). The same correlation cannot be established for the isoselenocyanate derivatives because the number of basins is not preserved. In this series, the number of basins is increased by 2. As CSe has a strong double bond character, a charge of about 1 $e^{-1}$ from V(N) would make the carbon hypervalent, therefore the opposite transfer (i.e., toward V(N)) is observed and the former V(N) basin gives rise to one V(N,H) basin and two symmetrically disposed V(N) basins. The trigonal bipyramid arrangement of the basins around the N center is consistent with Gillespie's VSEPR rules.

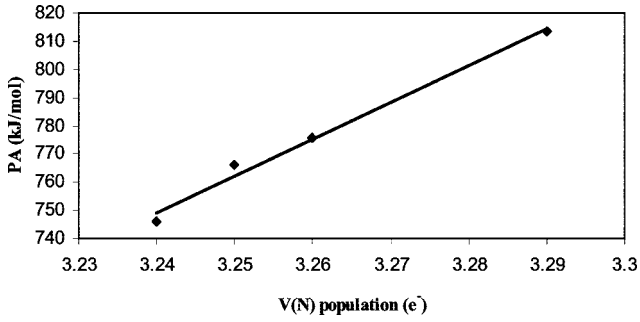



***Figure 3.*** Linear correlation between the calculated proton affinity (PA) and the population of the V(N) basin for selenocyanate derivatives.

It is also apparent that N-protonation triggers a certain electron density reorganization, through a slightly larger participation of mesomeric form **a**, leading to an increase of the population of the V(C,Se) basin. In fact, the weight of this form, according to the NRT analysis, increases up to 16% and the BO of the CSe bond becomes 1.21. Quite interestingly, also for the isoselenocyanates there is a reinforcement of the CSe bond upon N-protonation, but even stronger than that observed for the selenocyanate analogues. This as nicely illustrated by the increase in both the population of the V(C,Se) basin, from about 3.1 $e^-$ to 3.7 $e^-$, on average, and the electron density at the BCP (from 0.19 to 0.21 au), which therefore results in a significant increase of the Wiberg BO, which for the neutral species was around 2.0 and for the N-protonated ones becomes typically around 2.4. Simultaneously, the CN bond loses part of its double bond character (its BO being only 1.48). This dramatic increase in the CSe BO and in the population of the V(C,Se) basin points to a significant participation of mesomeric form **d** (see Scheme 1), which is indeed confirmed by a NRT analysis, which shows this form to become dominant (52%): the second dominant one (43%) is form **a**.

**Relationship between Stability and Bonding.** One question remains still to be answered, why are the selenocyanates halogen derivatives more stable, in general, than the corresponding iso-analogues, whereas for the methyl substituent, it is the other way around? As we have mentioned above, the characteristics of the CN and CSe bonds within each family of isomers are rather constant when the substituents are halogens but differ significantly when the substituent is a methyl group. One of the most striking differences is that on going from the selenocyanates to the iso-derivatives, the CN linkage becomes significantly weaker for the halogen derivatives, but not much for the methyl derivative. As a matter of fact, the ELF of the methyl-isoselenocyanate (see Figure 2) clearly shows a cylindrical symmetry of the CN basin, compatible with a dominant triple bond character, whereas for the F derivative, two connected basins are found in this region, which are consistent with a double bond character. The electron density at the CN BCP also reflects this difference, being larger (0.420 au) for the Me derivative than for the F derivative (0.410 au). This change would imply a larger destabilization of the system on going from the selenocyanate isomer to the iso-analogue when the substituent is F than when the substituent is Me. To this, some other important differences in the characteristics of the Se−X and N−X bonds are to be added. As shown in Figure 2, the Se−Me bond has a clear covalent character, whereas the Se−X bond, when X is a halogen atom, is strongly ionic in character, as reflected by the absence of valence basins in the Se−X region. The rather large polarizability of Se renders this interaction very strong. The important consequence is that the Se−X (X = F, Cl, Br) bonds are stronger than the Se−Me bond. As a matter of fact, the calculated Se−F bond dissociation energy, for instance, is 59 kJ $mol^{-1}$ greater than the Se−Me bond dissociation energy. On the other hand the N−Me bond in the iso-derivatives is significantly stronger than the N−X (X = F, C, Br) as clearly illustrated by the populations of the corresponding V(N,X) basins. This is not

surprising if one takes into account the fact that the N−X bond involves two very electronegative atoms when X is a halogen, and as was the case in the $F_2$ molecule, the bond is weak because electron density accumulates preferentially in the vicinity of both atoms rather than in the internuclear region.[68] This is clearly confirmed by the N−X dissociation energies in the isoselenocyanates derivatives. For example, the N−Me bond dissociation energy is 149 kJ $mol^{-1}$ greater than the N−F bond dissociation energy.

## Conclusions

According to our analysis of the bonding in selenocyanates and isoselenocyanates derivatives in terms of ELF, AIM, and NBO approaches, there are three main factors which explain why the former are more stable than the latter when the substituents are halogens, whereas for alkyl substituents, it is the other way around: (a) the Se−X (X = F, Cl, Br) bonds are much stronger than the Se−X (X = Me) bonds; (b) the N−X (X = F, Cl, Br) bonds are much weaker than the N−X bonds (X = Me) ones; (c) on going from the selenocyanates to the iso-selenocyanates, there is a significant weakening of the CN bond, which becomes essentially a double bond, when the substituents are halogen atoms, whereas upon methyl substitution the CN bond retains its triple bond character.

The same stability trends are observed for the corresponding N-protonated species. More importantly, the calculated stability differences are rather similar to those obtained for the neutral compounds, with the only exception of the methyl derivative. The obvious consequence is that selenocyanates and isoselenocyanates exhibit rather similar basicities in the gas-phase. Only for the methyl derivative, the latter isomer is 21 kJ $mol^{-1}$ more basic than the former. Both types of isomers behave as nitrogen bases in the gas phase.

**Acknowledgment.** This work has been partially supported by the DGI Project No. CTQ2006-08558/BQU, by the Project MADRISOLAR (reference S-0505/PPQ/0225) of the Comunidad Autónoma de Madrid and by Consolider on Molecular Nanoscience CSD2007-00010. C.T. acknowledges a FPI grant from the Ministerio de Educación y Ciencia of Spain. A generous allocation of computing time at the CCC of the UAM is also acknowledged.

**Supporting Information Available:** B3LYP/6−311+G(3df,2p) total energies and thermal corrections to the enthalpy and entropies for selenocyanates, iso-selenocyanates, and their nitrogen protonated species. This material is available free of charge via the Internet at http://pubs.acs.org.

### References

(1) Norbury, A. H. *Adv. Inorg. Chem. Radiochem.* **1965**, *17*, 231.

(2) Patai, S., Ed. *the Chemistry of Cyanates and Their Thio Derivatives*; Wiley: New York, 1977; Vol. Parts 1 and 2.

(3) Weaver, W. E.; Whaley, W. M. *J. Am. Chem. Soc.* **1946**, *68*, 2115.

(4) Franklin, W. J.; Werner, R. L.; Ashby, R. A. *Spectrochim. Acta* **1974**, *A-30*, 387.

(5) Arase, A.; Masuda, Y. *Chem. Lett.* **1976**, 1115.

(6) Tamura, Y.; Adachi, M.; Kawasaki, T.; Kita, Y. *Tetrahedron Lett.* **1979**, 2251.

(7) Franklin, W. J.; Werner, R. L. *Tetrahedron Lett.* **1965**, 3003.

(8) Franklin, W. J.; Werner, R. L.; Ashby, R. A. *Spectrochim. Acta* **1974**, *A 30*, 1293.

(9) Huang, Y.; Chen, R. Y. *Synth. Commun.* **2000**, *30*, 377.

(10) Sommen,; G, L.; Linden, A.; Heimgartner, H. *Eur. J. Org. Chem.* **2005**, 3128.

(11) Gerbaux, P.; Dechamps, N.; Flammang, R.; Reddy, P. N.; Srinivas, R. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 151.

(12) Richards, R. J.; Davis, R. W.; Gerry, M. C. L. *J. Chem. Soc., Chem. Commun.* **1980**, 915.

(13) Frost, D. C.; Macdonald, C. B.; McDowell, C. A.; Westwood, N. P. C. *J. Am. Chem. Soc.* **1981**, *103*, 4423.

(14) Jonkers, G.; Grabandt, O.; Mooyman, R.; Delange, C. A. *J. Electron Spectr.* **1982**, *26*, 147.

(15) Leung, H.; Suffolk, R. J.; Watts, J. D. *Chem. Phys.* **1986**, *109*, 289.

(16) Jonkers, G.; Mooyman, R.; Delange, C. A. *Mol. Phys.* **1981**, *43*, 655.

(17) Li, Y. M.; Qiao, Z. M.; Sun, Q.; Zhao, J. C.; Li, H. Y.; Wang, D. X. *Inorg. Chem.* **2003**, *42*, 8446.

(18) Medarde, M.; Lopez, J. L.; Morillo, M. A.; Tome, F.; Adeva, M.; Sanfeliciano, A. *Tetrahedron Lett.* **1995**, *36*, 8097.

(19) González, A. I.; Mó, O.; Yáñez, M. *J. Phys. Chem. A* **1999**, *103*, 1662.

(20) Panda, A.; Mugesh, G.; Singh, H. B.; Butcher, R. J. *Organometallics* **1999**, *18*, 1986.

(21) Krief, A.; Delmotte, C.; Colaux-Castillo, C. *Pure Appl. Chem.* **2000**, *72*, 1709.

(22) Nguyen, T. M.; Lee, D. *Org. Lett.* **2001**, *3*, 3161.

(23) Braverman, S.; Zafrani, Y.; Gottlieb, H. E. *Tetrahedron* **2001**, *57*, 9177.

(24) Kennedy, R. A.; Mayhew, C. A. *Phys. Chem. Chem. Phys.* **2001**, *3*, 5511.

(25) Sanz, P.; Mó, O.; Yáñez, M. *Chem.—Eur. J.* **2002**, *8*, 3999.

(26) Nguyen, T. M.; Guzei, I. A.; Lee, D. *J. Org. Chem.* **2002**, *67*, 6553.

(27) Uehlin, L.; Fragale, G.; Wirth, T. *Chem.—Eur. J.* **2002**, *8*, 1125.

(28) Thompson, K. C.; Canosa-Mas, C. E.; Wayne, R. P. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4133.

(29) Sanz, P.; Mó, O.; Yáñez, M. *New* **2002**, *26*, 1747.

(30) Graiff, C.; Predieri, G.; Tiripicchio, A. *Eur. J. Inorg. Chem.* **2003**, 1659.

(31) Musaev, D. G.; Geletii, Y. V.; Hill, C. L.; Hirao, K. *J. Am. Chem. Soc.* **2003**, *125*, 3877.

(32) Bachrach, S. M.; Demoin, D. W.; Luk, M.; Miller, J. V. *J. Phys. Chem. A* **2004**, *108*, 4040.

(33) Poleschner, H.; Seppelt, K. *Chem.—Eur. J.* **2004**, *10*, 6565.

(34) Bajor, G.; Veszpremi, T.; Riague, E. H.; Guillemin, J. C. *Chem.—Eur. J.* **2004**, *10*, 3649.

(35) Cardey, B.; Enescu, M. *ChemPhysChem* **2005**, *6*, 1175.

(36) Tuononen, H. M.; Suontamo, R.; Valkonen, J.; Laitinen, R. S.; Chivers, T. *J. Phys. Chem. A* **2005**, *109*, 6309.

(37) Guillemin, J. C.; Riague, E. H.; Gal, J. F.; Maria, P. C.; Mó, O.; Yáñez, M. *Chem.—Eur. J.* **2005**, *11*, 2145.

(38) Bleiholder, C.; Werz, D. B.; Koppel, H.; Gleiter, R. *J. Am. Chem. Soc.* **2006**, *128*, 2666.

(39) Iwamoto, T.; Sato, K.; Ishida, S.; Kabuto, C.; Kira, M. *J. Am. Chem. Soc.* **2006**, *128*, 16914.

(40) Dutton, J. L.; Tuononen, H. M.; Jennings, M. C.; Ragogna, P. J. *J. Am. Chem. Soc.* **2006**, *128*, 12624.

(41) Coles, M. P. *Curr. Org. Chem.* **2006**, *10*, 1993.

(42) Anderson, J. S. M.; Ayers, P. W. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2371.

(43) Berski, S.; Gajewski, G.; Latajka, Z. *J. Mol. Struct.* **2007**, *844*, 278.

(44) Pearson, J. K.; Boyd, R. J. *J. Phys. Chem. A* **2007**, *111*, 3152.

(45) Salon, J.; Sheng, J.; Jiang, J. S.; Chen, G. X.; Caton-Williams, J.; Huang, Z. *J. Am. Chem. Soc.* **2007**, *129*, 4862.

(46) Shishkina, S. V.; Shishkin, O. V.; Desenko, S. M.; Leszczynski, J. *J. Phys. Chem. A* **2007**, *111*, 2368.

(47) Kaur, D.; Sharma, P.; Bharatam, P. V.; Kaur, M. *Int. J. Quantum Chem.* **2008**, *108*, 983.

(48) Pearson, J. K.; Boyd, R. J. *J. Phys. Chem. A* **2008**, *112*, 1013.

(49) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.

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

(51) Pearson, J. K.; Ban, F. Q.; Boyd, R. J. *J. Phys. Chem. A* **2005**, *109*, 10373.

(52) Trujillo, C.; Mó, O.; Yáñez, M.; Tortajada, J.; Salpin, J.-Y. *J. Phys. Chem. B* **2008**, *112*, 5479.

(53) Trujillo, C.; Mó, O.; Yáñez, M. *ChemPhysChem*, **2008**, [Online] DOI: 10.1002/cphc.200800215.

(54) Becke, A. D.; Edgecombe, K. E. *J. Chem. Phys.* **1990**, *92*, 5397.

(55) Silvi, B.; Savin, A. *Nature* **1994**, *371*, 683.

(56) Savin, A.; Becke, A. D.; Flad, J.; Nesper, R.; Preuss, H.; Vonschnering, H. G. *Angew. Chem., Int. Ed. Engl.* **1991**, *30*, 409.

(57) Silvi, B. *J. Phys. Chem. A* **2003**, *107*, 3081.

(58) Kohout, M.; Pernal, K.; Wagner, F. R.; Grin, Y. *Theor. Chem. Acc.* **2004**, *112*, 453.

(59) Silvi, B. *Phys. Chem. Chem. Phys.* **2004**, *6*, 256.

(60) Noury, S.; Krokidis, X.; Fuster, F.; Silvi, B. *Comput. Chem.* **1999**, *23*, 597.

(61) *Amira 3.0* Template Graphics Software Inc.: San Diego, CA, 2002.

(62) Bader, R. F. W. *Atoms in Molecules. A Quantum Theory*; Clarendon Press: Oxford, U.K., 1990.

(63) Fradera, X.; Austen, M. A.; Bader, R. F. W. *J. Phys. Chem. A* **1999**, *103*, 304.

(64) Wiberg, K. B.; Bader, R. F. W.; Lau, C. D. H. *J. Am. Chem. Soc.* **1987**, *109*, 985.

(65) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899.

(66) Glendening, E. D.; Badenhoop, J. K.; Reed, A. E.; Carpenter, J. E.; Bohmann, J. A.; Morales, C. M.; Weinhold, F. *NBO-5.0*; Theoretical Chemistry Institute, University of Wisconsin: Madison, WI, 2004; http://www.chem.wisc.edu/~nbo5 (accessed March 23, 2008).

(67) Fuster, F.; Silvi, B. *Chem. Phys.* **2000**, *252*, 279.

(68) Cremer, D.; Kraka, E. *Angew. Chem., Int. Ed. Engl.* **1984**, *96*, 612.

CT800178X

JCTC Journal of Chemical Theory and Computation

# Efficiency and Accuracy of the Generalized Solvent Boundary Potential for Hybrid QM/MM Simulations: Implementation for Semiempirical Hamiltonians

Tobias Benighaus and Walter Thiel*

*Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470, Mülheim an der Ruhr, Germany*

Received May 27, 2008

**Abstract:** We report the implementation of the generalized solvent boundary potential (GSBP) [Im, W.; Bernèche, S.; Roux, B. *J. Chem. Phys.* **2001**, *114*, 2924] in the framework of semiempirical hybrid quantum mechanical/molecular mechanical (QM/MM) methods. Application of the GSBP is connected with a significant overhead that is dominated by numerical solutions of the Poisson−Boltzmann equation for continuous charge distributions. Three approaches are presented that accelerate computation of the values at the boundary of the simulation box and in the interior of the macromolecule and solvent. It is shown that these methods reduce the computational overhead of the GSBP significantly with only minimal loss of accuracy. The accuracy of the GSBP to represent long-range electrostatic interactions is assessed for an extensive set of its inherent parameters, and a set of optimal parameters is defined. On this basis, the overhead and the savings of the GSBP are quantified for model systems of different sizes in the range of 7000 to 40 000 atoms. We find that the savings compensate for the overhead in systems larger than 12 500 atoms. Beyond this system size, the GSBP reduces the computational cost significantly, by 70% and more for large systems (>25 000 atoms).

## 1. Introduction

In the past decade, hybrid quantum mechanical/molecular mechanical (QM/MM) methods have gained popularity for the simulation of biomolecules and are now frequently used for the calculation of free energy differences.[1−11] In the context of this development, the treatment of long-range electrostatic interactions in QM/MM simulations attracted significant attention. The accurate description of long-range electrostatic interactions was found to be imperative for meaningful simulations of biomolecular systems, since electrostatic interactions strongly influence their structure and function.[12−15] While the development of efficient and accurate methods to treat these interactions has been an active area of research in the field of classical simulations for a long time, these techniques are only recently adapted to QM/MM methods due to the technical difficulties introduced by the QM atoms.

For the case of periodic boundary conditions (PBC), Ewald summation is an established method to compute the electrostatic energy and forces of an infinite periodic array of systems without significant truncations.[16−18] Therefore, the applicability of the Ewald summation method has been extended to hybrid QM/MM simulations with semiempirical QM Hamiltonians.[19−21] Unfortunately, application of these methods to large nonperiodic biomolecules is affected by serious problems. The imposed periodicity may lead to significant artifacts[22−24] and even qualitatively wrong results unless the molecule is solvated in a solvent box of adequate size.[25] Thus, the number of solvent molecules is necessarily large and increases the computational costs massively, such that the Ewald summation method can only be used for small- to medium-sized biomolecules.

Very often, however, one is interested in simulating the behavior of a large biomolecule in infinite dilution, and alternative approaches were devised to facilitate these computations. For biochemical reactions that proceed in a localized region of the macromolecule, boundary potentials

* Corresponding author. E-mail: thiel@mpi-muelheim.mpg.de.

10.1021/ct800193a CCC: $40.75 
Published on Web 09/19/2008

are an especially attractive approach.[26−37] Within this approach, the system is subdivided into an inner region, containing the active site and the adjacent part of the enzyme, and an outer region, containing the rest of the enzyme and the outer solvent molecules. While the inner region is simulated atomistically, the effect of the outer region onto the inner region is mimicked by the boundary potential. Ideally, the boundary potential is designed such that the statistical properties of the inner region interacting with the boundary potential are the same as those of the full solvated macromolecule. Although this may be formulated rigorously as an integration over the outer region degrees of freedom,[37] an efficient implementation necessitates the introduction of further approximations.

In the generalized solvent boundary potential (GSBP), developed by Im et al. in 2001, the outer region solvent molecules are described by a continuous polarizable dielectric and the outer region charge distribution by fixed point charges.[38] Electrostatic interactions with the outer region macromolecule and solvent molecules are separated into a solvent-shielded static field created by the outer region point charges interacting with the dielectric, and a dynamic reaction field induced by interaction of the inner region charge distribution with the dielectric. A great advantage of the GSBP is that the dynamic reaction field term can handle irregularly shaped macromolecule/solvent boundaries. Accuracy and efficiency of the GSBP in classical simulations were validated by studies on aspartyl-tRNA synthetase[38,39] and the KcsA potassium channel.[38] In 2005, Cui and co-workers adapted the GSBP method to the QM/MM framework as a means to treat long-range electrostatic interactions in QM/MM simulations accurately and to describe QM/MM and MM/MM interactions in a balanced way.[40] Here, the self-consistent-charge density-functional tight-binding (SCC-DFTB)[41] method was chosen as the QM Hamiltonian. The accuracy of the SCC-DFTB/MM/GSBP approach was evaluated by comparison to results from Ewald/PBC calculations on small model systems. The SCC-DFTB/MM/GSBP method was found to provide quantitatively very similar results at significantly lower computational costs compared to Ewald/PBC methods.[42−44] The fixation of the outer region atoms is a fundamental assumption in the GSBP that allows for a closed-form expression for the electrostatics.[38] While this assumption is valid in many cases, the use of the GSBP was found to be problematic if the macromolecule underwent major conformational changes during the course of a reaction.[44] For the investigation of localized processes in large macromolecules, the SCC-DFTB/MM/GSBP approach proved to be an efficient and accurate method and was applied subsequently to study several biological systems.[45−48]

The use of the GSBP is connected with a significant overhead. Initially, the solvent-shielded static field and the matrix representation of the reaction field Green's function have to be calculated. Computation of the reaction field matrix implies solving several hundred linearized Poisson−Boltzmann (PB) equations and is therefore rather demanding. Furthermore, the accuracy of the GSBP and the costs of its overhead strongly depend on the choice of parameters that are inherent to the GSBP and the finite-difference solution of the PB equation. To the best of our knowledge, a systematic determination of the best parameters for the GSBP has not been pursued up to date. In this study, we determine a set of parameters that provide the accuracy that is necessary to mimic the effect of the outer region at optimal computational costs. On the basis of these parameters, we quantify the overhead and the savings related to the GSBP, and estimate the minimum system size for which the GSBP is more efficient than standard approaches using nontruncated Coulombic electrostatics. Moreover, we present improved algorithms that decrease the costs for computation of the reaction field matrix significantly.

Previously, the GSBP method was adapted to the hybrid QM/MM framework exclusively in combination with the SCC-DFTB Hamiltonian for the QM region.[40] In light of the success of reaction-specific parametrizations of semiempirical methods based on the neglect of diatomic differential overlap (NDDO) approximation in QM/MM simulations[49−53] and the widespread use of NDDO-based QM/MM methods in general,[11] we found it desirable to adapt and implement the GSBP as an efficient means to treat long-range electrostatics in NDDO-based QM/MM simulations. This interest is further substantiated by recent findings that NDDO-based methods are more reliable for certain properties and systems,[54] and that the use of SCC-DFTB may be problematic for specific systems.[55] Accordingly, we adapted the GSBP for NDDO-based QM/MM approaches and present the implementation in this work.

## 2. Theory

In this section, we briefly review the theoretical background of the GSBP for classical MM simulations[38] and its QM/MM implementation.[40] Thereafter, the adaptation to NDDO-based QM/MM methods and strategies to accelerate computation of the reaction field matrix are presented.

**2.1. GSBP for MM Methods.** Consider a macromolecule $R$ surrounded by $N$ solvent molecules. In a boundary potential approach, the system is subdivided into an inner region that contains the inner part of the macromolecule and the $n$ inner solvent molecules, and an outer region that contains the outer part of the macromolecule and the $N - n$ outer solvent molecules. Statistical expectation values depending only on the degrees of freedom of the inner region ($\mathbf{R}_i$, 1,..., $n$) can be calculated by integrating out the outer region contributions. The influence of the outer region on the inner region can be described rigorously by means of the potential of mean force (PMF) $W(\mathbf{R}_i, 1,..., n)$.

$$e^{-\beta W(\mathbf{R}_i,1,...\mathbf{n})} = \frac{1}{C}\int' d\mathbf{R}_o\, d(n+1) \ldots dN e^{-\beta U(\mathbf{R},1,...,N)} \quad (1)$$

Here, $C$ denotes an arbitrary integration constant, and the primed integral indicates integration over the degrees of freedom of the outer region ($\mathbf{R}_o$, $n + 1$,..., $N$) including only those configurations with all outer region atoms outside the inner region. Beglov and Roux demonstrated that the PMF is related to the reversible thermodynamic work necessary to assemble the inner region.[37]

$$W(\mathbf{R}_i, 1, ..., n) = U(\mathbf{R}_i, 1, ..., n) + \Delta W_{cr} + \Delta W_{np}(\mathbf{R}_i, 1, ..., n) + \Delta W_{elec}(\mathbf{R}_i, 1, ..., n) \quad (2)$$

The contribution to the PMF that arise from the configurational restrictions and the nonpolar and the electrostatic

interactions are denoted $\Delta W_{cr}$, $\Delta W_{np}$, and $\Delta W_{elec}$, respectively. $U$ is the potential energy of the isolated inner region that includes bonded and nonbonded (van der Waals and electrostatic) terms.

The goal of the GSBP is to provide an efficient and accurate approximation to the electrostatic contribution to the PMF. Therefore, the outer region solvent molecules are described by a polarizable dielectric continuum (PDC) and the outer region macromolecule by fixed point charges. The electrostatic contributions to the PMF now consist of the direct Coulombic interactions of inner and outer region ($U^{io}_{elec}$), and the solvation free energy resulting from interaction with the PDC ($\Delta W^{solv}_{elec}$). Representing the charge distribution of the outer macromolecule and the inner region by point charges $q_A$, the electrostatic solvation free energy can be calculated as

$$\Delta W^{solv}_{elec} = \frac{1}{2}\sum_A q_A \phi_{rf}(\mathbf{r}_A) \tag{3}$$

where the reaction field potential $\phi_{rf}(\mathbf{r})$ is the difference of a reference electrostatic potential computed in vacuum, $\phi_v(\mathbf{r})$, and the electrostatic potential computed in solution, $\phi_s(r)$. The electrostatic potentials are obtained by solving the linearized Poisson−Boltzmann (PB) equation

$$\nabla[\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})] - \bar{\kappa}^2(r)\phi(\mathbf{r}) = -4\pi\rho(\mathbf{r}) \tag{4}$$

with the charge density of all explicit atoms $\rho(\mathbf{r})$, the space-dependent dielectric constant $\epsilon(\mathbf{r})$, and modified Debye−Hückel screening factor $\bar{\kappa}(\mathbf{r})$.[56] The solvation free energy term is problematic, since during sampling of the inner region configurations the PB equation would have to be solved for each configuration which is prohibitively expensive. To isolate the dynamic properties, the charge distribution is separated into an inner and outer part.

$$\rho(\mathbf{r}) = \rho_i(\mathbf{r}) + \rho_o(\mathbf{r}) \tag{5}$$

In consequence, the electrostatic solvation free energy splits up into three terms: outer−outer, inner−outer, and inner−inner contributions.

$$\Delta W^{solv}_{elec} = \Delta W^{oo}_{elec} + \Delta W^{io}_{elec} + \Delta W^{ii}_{elec} \tag{6}$$

The first term, $\Delta W^{oo}_{elec}$, stems from the interaction of the outer region charge distribution with the self-induced reaction field and is constant throughout sampling. The inner−outer contribution arises from the interaction of the inner region charge distribution with the reaction field that is induced by the outer region charge distribution. Calculation of the Coulombic interaction of the inner and outer region can be combined very efficiently with the calculation of the inner−outer contribution to the solvation free energy.

$$\Delta W^{io}_{elec} + U^{io}_{elec} = \sum_{A\in inner} q_A \phi^o_{rf}(\mathbf{r}_A) + U^{io}_{elec}$$

$$= \sum_{A\in inner} q_A \phi^o_s(\mathbf{r}_A) \tag{7}$$

The outer region being fixed, the electrostatic potential of the outer region charges in solution, $\phi^o_s(\mathbf{r})$, has to be computed only once and is valid for all inner region configurations. As the interaction with all outer region point charges and solvent molecules is substituted by an interaction with a static potential, computational costs are reduced massively. However, computation of the inner−inner contribution

$$\Delta W^{ii}_{elec} = \frac{1}{2}\sum_{A\in inner} q_A \phi^i_{rf}(\mathbf{r}_A) \tag{8}$$

remains problematic because $\phi^i_{rf}(\mathbf{r})$ depends on the inner region configuration. To circumvent repeated solution of the PB equation, the inner region charge distribution is projected onto a set of basis functions $\{b_n\}$:

$$\rho_i(\mathbf{r}) = \sum_n c_n b_n(\mathbf{r}) \tag{9}$$

For a set of orthonormal basis functions, the expansion coefficients $c_n$ are the generalized multipole moments $Q_n$ of the charge distribution:

$$Q_n = \sum_{A\in inner} q_A b_n(\mathbf{r}_A) \tag{10}$$

Finally, the reaction field Green's function which determines the inner reaction field potential

$$\phi^i_{rf}(\mathbf{r}) = \int d\mathbf{r}'\ \rho_i(\mathbf{r}')G_{rf}(\mathbf{r},\mathbf{r}') \tag{11}$$

is projected onto the same basis set. Using $M_{mn}$, the matrix representation of $G_{rf}$, the inner−inner electrostatic contribution to the PMF can be expressed as

$$\Delta W^{ii}_{elec} = \frac{1}{2}\sum_{mn} Q_m M_{mn} Q_n \tag{12}$$

This leads to the final expression for the electrostatic contribution to the PMF

$$\Delta W_{elec} = \sum_{A\in inner} q_A \phi^o_s(\mathbf{r}_A) + \frac{1}{2}\sum_{mn} Q_m M_{mn} Q_n \tag{13}$$

In MD simulations employing the GSBP, the inner region atoms move on the PMF surface that is defined as

$$W(\mathbf{R}_i, 1, ..., n) = U(\mathbf{R}_i, 1, ..., n) + \Delta W_{cr} + \Delta W_{np} + \sum_{A\in inner} q_A \phi^o_s(\mathbf{r}_A) + \frac{1}{2}\sum_{mn} Q_m M_{mn} Q_n \tag{14}$$

**2.2. GSBP Implementation for NDDO-Based QM/MM Methods.** Extension of the GSBP to general QM/MM methods necessitates further subdivision of the inner region into QM and MM regions, since it is natural to assume that the QM region lies within the inner region. Consequently, the inner region charge distribution splits up into QM and MM charge distributions that interact separately with the static outer region field, $\phi^o_s$, and the reaction field Green's function, $G_{rf}$. Equation 13 has to be modified as follows to account for these changes:

$$\Delta W_{elec} = \sum_{A\in MM} q_A \phi^o_s(\mathbf{r}_A) + \int d\mathbf{r}\rho^{QM}(\mathbf{r})\phi^o_s(\mathbf{r}) + \frac{1}{2}\sum_{mn} Q^{QM}_m M_{mn} Q^{QM}_n + \sum_{mn} Q^{QM}_m M_{mn} Q^{MM,cs}_n + \frac{1}{2}\sum_{mn} Q^{MM}_m M_{mn} Q^{MM}_n \tag{15}$$

The main issue that arises when introducing QM atoms into the GSBP framework is the representation of the QM charge distribution in the terms that describe the interaction with the outer region field and the reaction field. As NDDO-based semiempirical QM methods use only a minimum set of relatively tight basis functions, we decided to represent the QM charge distribution by a set of Mulliken charges.[57] Now, the QM-dependent terms in eq 15 can be calculated in close analogy to the MM terms

$$\int \mathrm{d}\mathbf{r}\, \rho^{\mathrm{QM}}(\mathbf{r})\phi_{\mathrm{s}}^{\mathrm{o}}(\mathbf{r}) = \sum_{A\in \mathrm{QM}} q_{\mathrm{A}}^{\mathrm{Mull}}\phi_{\mathrm{s}}^{\mathrm{o}}(\mathbf{r}_A) \tag{16}$$

and

$$Q_n^{\mathrm{QM}} = \int \mathrm{d}\mathbf{r}\, \rho^{\mathrm{QM}}(\mathbf{r})b_n(\mathbf{r}) = \sum_{A\in \mathrm{QM}} q_A^{\mathrm{Mull}} b_n(\mathbf{r}_A) \tag{17}$$

Here, $q_A^{\mathrm{Mull}}$ are the Mulliken charges representing the QM charge distribution, and $Q_n^{\mathrm{QM}}$ are the multipole moments of the QM charge distribution. Still, we are facing two technical difficulties. First, electrostatic interactions at the QM−MM boundary need to be treated with special care to avoid overpolarization of the QM electron density. Thus, the QM electron density does not interact with the full MM charge distribution but with a modified one. In this work, we implemented the GSBP for use in combination with the charge-shift scheme,[58] and therefore, the QM charge distribution interacts with the reaction field potential that is induced by the charge-shifted MM charges ($\mathrm{MM^{cs}}$) (fourth term in eq 15), with

$$Q_n^{\mathrm{MM,cs}} = \sum_{\mathrm{A}\in \mathrm{MM^{cs}}} q_{\mathrm{A}} b_n(\mathbf{r}_{\mathrm{A}}) \tag{18}$$

Second, using electronic embedding,[59] the QM wave function interacts with all MM point charges and the PDC. Hence, the GSBP contributions have to be accommodated at the level of the self-consistent field (SCF) iterations during optimization of the wave function by addition of the following terms to the Fock matrix.

$$F_{\mu\nu}^{\mathrm{GSBP}} = -\frac{1}{2}\delta_{\mu\nu}[\Omega_{\mathrm{C}} + \Omega_{\mathrm{D}}] - \frac{1}{2}\delta_{\mu\nu}\sum_{A\in \mathrm{QM}} q_A^{\mathrm{Mull}}[\Gamma_{\mathrm{CA}} + \Gamma_{\mathrm{DA}}]; \quad \mu\in \mathrm{C},\ \nu\in \mathrm{D} \tag{19}$$

Here, $\mu$ and $\nu$ denote basis functions attached to the QM atoms C and D, respectively. The atom-dependent matrices $\Omega_{\mathrm{C}}$ and $\Gamma_{\mathrm{CA}}$ are defined as

$$\Omega_{\mathrm{C}} = \phi_{\mathrm{s}}^{\mathrm{o}}(\mathbf{r}_{\mathrm{C}}) + \sum_{mn} b_m(\mathbf{r}_{\mathrm{C}})M_{mn}Q_n^{\mathrm{MM,cs}} \tag{20}$$

and

$$\Gamma_{\mathrm{CA}} = \sum_{mn} b_m(\mathbf{r}_{\mathrm{C}})M_{mn}b_n(\mathbf{r}_{\mathrm{A}}) \tag{21}$$

Moreover, the GSBP also affects the atomic forces, and its contribution to the analytic gradient can be evaluated by taking the first derivative of the GSBP contribution to the PMF with respect to the atomic coordinates. In the case of a QM atom, the analytic derivative takes the following form

$$\frac{\partial}{\partial \mathbf{r}_{\mathrm{A}}}\Delta W_{\mathrm{elec}} = q_{\mathrm{A}}^{\mathrm{Mull}}\frac{\partial}{\partial \mathbf{r}_{\mathrm{A}}}\phi_{\mathrm{s}}^{\mathrm{o}}(\mathbf{r}_{\mathrm{A}}) + \sum_{\mathrm{B}\in \mathrm{QM}} \frac{\partial q_{\mathrm{B}}^{\mathrm{Mull}}}{\partial \mathbf{r}_{\mathrm{A}}}\phi_{\mathrm{s}}^{\mathrm{o}}(\mathbf{r}_{\mathrm{B}}) + \sum_{mn}\left[\frac{\partial}{\partial \mathbf{r}_{\mathrm{A}}}Q_m^{\mathrm{QM}}\right]M_{mn}[Q_n^{\mathrm{QM}} + Q_n^{\mathrm{MM,cs}}] \tag{22}$$

where the derivatives of the QM multipole moments are calculated as

$$\frac{\partial}{\partial \mathbf{r}_{\mathrm{A}}}Q_m^{\mathrm{QM}} = q_{\mathrm{A}}^{\mathrm{Mull}}\frac{\partial}{\partial \mathbf{r}_{\mathrm{A}}}b_m(\mathbf{r}_{\mathrm{A}}) + \sum_{\mathrm{B}\in \mathrm{QM}} \frac{\partial q_{\mathrm{B}}^{\mathrm{Mull}}}{\partial \mathbf{r}_{\mathrm{A}}}b_m(\mathbf{r}_{\mathrm{B}}) \tag{23}$$

In contrast to a previous implementation of the GSBP for hybrid QM/MM approaches,[40] we found it necessary for NDDO-based QM methods to include the contribution from coupled Mulliken charge derivatives, $\partial q_{\mathrm{A}}^{\mathrm{Mull}}/\partial \mathbf{r}_{\mathrm{B}}$, to compute accurate gradients of the QM atoms. Using only one-center Mulliken charge derivatives, $\partial q_{\mathrm{A}}^{\mathrm{Mull}}/\partial \mathbf{r}_{\mathrm{A}}$, a mean absolute deviation (MAD) of the components of the QM gradient in the range of $10^{-3}$ au was observed (compared with finite-difference reference values). Incorporating the contribution from the coupled Mulliken charge derivatives reduces the MAD to the order of $10^{-5}$ au which is sufficiently accurate. Although Mulliken charge derivatives take a very simple form in the NDDO approximation

$$\frac{\partial}{\partial \mathbf{r}_{\mathrm{B}}}q_A^{\mathrm{Mull}} = -\sum_{\alpha\in \mathrm{A}}\frac{\partial}{\partial \mathbf{r}_{\mathrm{B}}}P_{\alpha\alpha} \tag{24}$$

their computation is complicated, since the coupled-perturbed SCF (CPSCF) equations have to be solved to calculate the derivatives of the SCF density matrix.[60,61]

In the case of an MM atom, the evaluation of the derivative of the GSBP contribution to the PMF is less demanding:

$$\frac{\partial}{\partial \mathbf{r}_{\mathrm{A}}}\Delta W_{\mathrm{elec}} = q_{\mathrm{A}}\frac{\partial}{\partial \mathbf{r}_A}\phi_{\mathrm{rf}}^{\mathrm{o}}(\mathbf{r}_{\mathrm{A}}) + q_{\mathrm{A}}\sum_{mn}\left[\frac{\partial}{\partial r_{\mathrm{A}}}b_n(r_{\mathrm{A}})\right]M_{mn}[Q_n^{\mathrm{QM}} + Q_n^{\mathrm{MM}}] \tag{25}$$

**2.3. Computation of the Reaction Field Matrix.** Although the reaction field matrix is formally the matrix representation of the reaction field Green's function, the computation of this matrix follows a different approach that exploits the fact that its *mn*th element corresponds to the interaction of $b_m$ with the reaction field induced by $b_n$.[38]

$$M_{mn} = \int \mathrm{d}\mathbf{r}\, b_n(\mathbf{r})\phi_{\mathrm{rf}}(\mathbf{r};b_m(\mathbf{r})) \tag{26}$$

To calculate $\phi_{\mathrm{rf}}(\mathbf{r};b_m(\mathbf{r}))$, it is necessary to solve the PB equation with the dielectric boundary defined by the macromolecule, and the charge distribution defined by $b_m$ in the inner region and set to zero in the outer region, for vacuum and solvent conditions. Since a standard GSBP calculation employs about 400 basis functions,[38] computation of the reaction field matrix implies solving the PB equation about 800 times. This procedure is computationally expensive and dominates the GSBP-related overhead. Therefore, three approaches to accelerate computation of the reaction field matrix are presented in this section.

*2.3.1. Coarsening of the Inner Region.* In finite-difference solutions of the PB equation, the boundary values are commonly set using the Debye−Hückel expression[56]

$$\phi_i = \sum_j \frac{q_j e^{-\kappa r_{ij}}}{\epsilon r_{ij}} \tag{27}$$

that implies summation over all point charges $q_j$ for each boundary point $\phi_i$. With a continuous charge distribution in the inner region, determination of the boundary values becomes computationally expensive. In the original GSBP work, a focusing procedure[62] is used to reduce these computational costs. In this procedure, the PB equation is first solved for a rough outer grid (grid I) with large spatial extent. Subsequently, a fine inner grid (grid II) focusing on the inner region with boundary values defined by grid I is used to calculate an accurate electrostatic potential. However, even when such a focusing procedure is used, determination of the boundary values of grid I still has a significant share of the computational costs. Since the boundary points of grid I are far from the inner region and the "charge" in the outer region is zero, a less accurate representation of the basis function in the inner region is expected to be sufficient. Therefore, we introduce the "coarsening of the inner region" (CIR) approximation that utilizes a very rough grid (grid III) to represent the "charge" distribution that is only used to determine the boundary values of grid I. The mesh size of grid III is the product of the new CIR factor and the mesh size of grid I; i.e., a CIR factor of 1.0 corresponds to a standard focusing procedure.

*2.3.2. Linear Interpolation.* In view of the large distance between the boundary points of grid I and the inner region, it is evident that the boundary values are slowly varying. Therefore, we introduce a simple interpolation scheme that reduces the number of explicitly determined boundary values significantly. On the edges every second and on the faces every fourth boundary value is calculated using the Debye−Hückel expression. The remaining boundary values are determined by linear interpolation from the adjacent four or two boundary points. This scheme is illustrated in Figure 1. For an example grid with $100^3$ points, the linear interpolation scheme reduces the number of explicitly determined boundary values from 58 416 to 14 802.

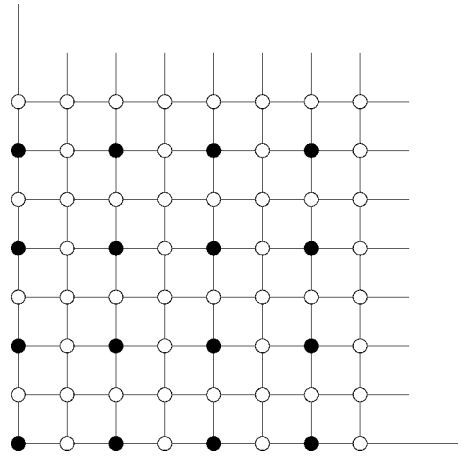


***Figure 1.*** Interpolation scheme used to define boundary values in finite-difference solutions of the PB equation. Black circles represent boundary points that are set using the Debye−Hückel expression. White circles represent boundary points that are set by interpolation.

*2.3.3. Modified Stripping.* In a finite-difference solution to the PB equation with zero salt conditions, the potential at a particular grid point, $\phi_0$, is updated using the potential at the six nearest neighbors, $\phi_i$, the dielectric constants $\epsilon_i$ at the midpoints between $\phi_0$ and $\phi_i$, and the charge $q_0$ assigned to that grid point.

$$\phi_0 = \frac{\sum_{i=1}^{6} \epsilon_i \phi_i + 4\pi q_0/h}{\sum_{i=1}^{6} \epsilon_i} \tag{28}$$

Here, $h$ is the distance between two grid points. This procedure implies 13 additions, 7 multiplications, and 1 division per grid point. Honig et al. demonstrated[63] that the number of mathematical operations can be reduced significantly for most grid points. For a point with zero charge that is surrounded by a uniform dielectric constant, eq 28 simplifies to

$$\phi_0 = \frac{1}{6}\sum_{i=1}^{6} \phi_i \tag{29}$$

Updating these points requires only 6 additions and 1 multiplication. This procedure is termed "stripping" because the points are updated separately.[63] As a continuous charge distribution is used in the computation of the reaction field matrix, there are no points without charge in the inner region. Therefore, we apply a "modified stripping" approach and drop the zero charge condition: for all points surrounded by a uniform dielectric constant with arbitrary charge, we simplify eq 28 as follows:

$$\phi_0 = \frac{1}{6}\sum_{i=1}^{6} \phi_i + \frac{2\pi q_0}{3h\epsilon} \tag{30}$$

Although three additional operations per grid point (one addition, one multiplication, and one division) are necessary compared to the standard stripping approach, modified stripping offers computational savings since it is applicable to a significantly larger number of grid points.

## 3. Computational Details

The GSBP was implemented in a developmental version of the modular program package ChemShell.[58] The energy and gradient evaluations for the QM part were performed with the MNDO2004 program that was modified locally to account for the GSBP contribution. The AM1 method was chosen as the QM Hamiltonian.[64] The SCF convergence criterion was $10^{-8}$ eV. For the MM part, the DL_POLY[65] code was employed to run the CHARMM22 force field.[66] The PB equation was solved with our new ChemShell PB module that uses the optimal successive over-relaxation method in combination with Gauss−Seidel relaxation to compute the electrostatic potential.[63,67] A convergence criterion of $2 \times 10^{-5}$ au was employed for the maximum absolute change in every grid point. If not stated otherwise, the dielectric constants of the macromolecule, $\epsilon_m$, and the solvent, $\epsilon_s$, were set to 1 and 80, respectively. van der Waals radii from the CHARMM22 force field were used to define

***Table 1.*** Average Absolute Percentage Deviation [%] of the Electrostatic Interaction between Inner and Outer Region Computed from the PB Electrostatic Potential with Different Mesh Sizes of the Inner and Outer Grid

| | inner grid size (Å) | | | |
|---|---|---|---|---|
| outer grid size (Å) | 0.25 | 0.40 | 0.60 | 0.80 |
| 1.00 | 0.13 | 0.14 | 0.22 | 0.55 |
| 1.25 | 0.16 | 0.28 | 0.26 | 0.61 |
| 1.50 | 0.25 | 0.28 | 0.28 | 0.62 |
| 1.75 | 0.21 | 0.28 | 0.28 | 0.64 |
| 2.00 | 0.23 | 0.16 | 0.24 | 0.61 |
| 2.50 | 0.25 | 0.27 | 0.27 | 0.60 |

the dielectric boundary. All calculations for which timings are reported were performed serially on 2.6 GHz AMD Opteron machines with 16 GB of memory.

## 4. Optimization of Parameters

The accuracy and the efficiency of the GSBP strongly depend on the values that are chosen for its inherent parameters. In this section, we determine a set of optimal values for the mesh sizes of the inner and outer grid, the CIR factor, and assess the accuracy of the approximations that were introduced in section 2.3.

**4.1. Static Outer Region Field.** The reliability of the static solvent-shielded outer region field, $\phi_s^o$, to mimic the electrostatic potential in the inner region is judged by comparison to the exact Coulombic potential. In vacuum environment, i.e., with $\epsilon_m = \epsilon_s = 1$, the electrostatic interactions between the inner and outer region have to be identical when using the electrostatic potentials from solution of the PB equation or the Coulomb expression. A model system consisting of a threonine molecule solvated in a TIP3P water ball with 30 Å radius and 4175 water molecules was set up for this study. By means of classical molecular dynamics (MD) simulation, 10 different configurations of this model system were generated. For each configuration, the center of the inner region was taken to be the $C_\alpha$ carbon of threonine. All molecules with any atom within 18 Å from the center were assigned to the inner region. Depending on the configuration, the inner region contained between 2858 and 2978 atoms. As the electrostatic interaction energy varies with the size of the inner region, we averaged over the absolute percentage deviation in the electrostatic interaction energy.

On average, the vacuum electrostatic interaction between the inner and outer region ($U_{elec}^{io}$) was $-3096.2 \pm 243.5$ kcal/mol. The average absolute percentage deviation was calculated for all combinations of outer grid mesh sizes of 1.0, 1.25, 1.5, 1.75, 2.0, and 2.5 Å, and inner grid mesh sizes of 0.25, 0.4, 0.6, and 0.8 Å. The results given in Table 1 indicate that the interaction energy calculated from the PB electrostatic potential is very accurate. All mesh size combinations provide average deviations $<0.3\%$ if the inner grid spacing is $\leq 0.6$ Å, indicating that 0.6 Å is reasonable choice for the inner grid spacing. For the outer grid, no reliable correlation was found between mesh size and accuracy.

To ensure that the accuracy is not euphemized by cancelation of errors, we also assess the reliability of the electrostatic forces in the inner region. For this purpose, the MAD of the electrostatic force components of all atoms inside spherical active regions with radii of 15 and 17 Å were computed for each configuration. In Table 2, the average of the MADs is given for all mesh size combinations. For both active regions, the accuracy of the electrostatic forces seems to be rather independent of the mesh sizes. Within 15 Å of the center of the inner region, computation of the electrostatic forces based on the potential from the PB equation is quite accurate with average MADs around $4 \times 10^{-5}$ to $8 \times 10^{-5}$ au. The average deviation increases by a factor of 4−5 if the radius of the active region is extended to 17 Å. For each active region, there is only a very weak correlation between accuracy and mesh size. However, the accuracy strongly depends on the size of the active region. In Figure 2, the average MAD for one mesh size combination (0.6 and 1.75 Å) is plotted as a function of the radius of the active region. This figure shows that the accuracy is very high for radii of up to 16 Å, then the deviation increases

***Table 2.*** Average Mean Absolute Deviation ($10^{-4}$ au) of the Electrostatic Forces Computed from the PB Electrostatic Potential for Different Mesh Sizes of the Inner and Outer Grid Used for Solving the PB Equation[a]

| | inner grid size (Å) | | | |
|---|---|---|---|---|
| outer grid size (Å) | 0.25 | 0.40 | 0.60 | 0.80 |
| | 15 Å Active Region | | | |
| 1.00 | 0.63 | 0.34 | 0.52 | 0.72 |
| 1.25 | 0.52 | 0.64 | 0.58 | 0.58 |
| 1.50 | 0.77 | 0.68 | 0.60 | 0.59 |
| 1.75 | 0.76 | 0.79 | 0.76 | 0.75 |
| 2.00 | 0.74 | 0.39 | 0.55 | 0.76 |
| 2.50 | 0.86 | 0.66 | 0.59 | 0.61 |
| | 17 Å Active Region | | | |
| 1.00 | 2.13 | 1.98 | 2.36 | 2.85 |
| 1.25 | 2.02 | 2.27 | 2.42 | 2.72 |
| 1.50 | 2.28 | 2.31 | 2.43 | 2.73 |
| 1.75 | 2.27 | 2.42 | 2.58 | 2.87 |
| 2.00 | 2.26 | 2.03 | 2.39 | 2.88 |
| 2.50 | 2.37 | 2.28 | 2.43 | 2.75 |

[a] Active regions with radii of 15 and 17 Å were chosen.

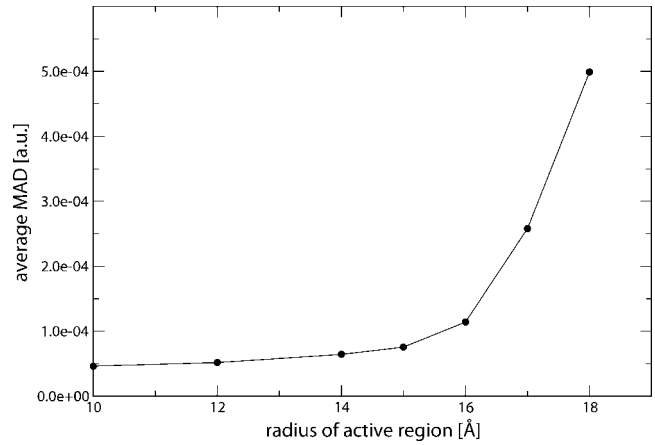



***Figure 2.*** Average mean absolute deviation (MAD) of the electrostatic forces of all atoms inside the active region as a function of the radius of the active region. Mesh sizes of 0.6 and 1.75 Å were used for the inner and outer grid, respectively. The radius of the inner region is 18 Å (see text).

strongly. This behavior is identical for all mesh size combinations.

Keeping in mind the size of the inner region (18 Å, see above), we conclude that a grid-based PB potential is not adequate to represent the details of the electrostatic potential of the outer region in close proximity to the outer region. Although these inaccuracies are likely to have only an insignificant effect on the region of interest if the size of the inner region is adequate, we recommend to keep all atoms in the outer 2−3 Å layer of the inner region fixed, since such a frozen layer will increase the reliability of the GSBP. A mesh size of 0.6 Å for the inner grid seems to provide an ideal tradeoff between accuracy and computational costs. The results are not clearcut concerning the outer grid mesh size where the accuracy seems to be rather independent of the mesh size. This indicates that the electrostatic potential is only slowly varying at the boundary of the inner grid. To be on the safe side, we opted for an outer grid spacing of 1.75 Å.

**4.2. Reaction Field Matrix.** In section 2.3, three approaches to accelerate computation of the reaction field matrix were presented. While the modified stripping technique provides speed-up without loss of accuracy, the CIR and the linear interpolation approaches are approximations to define the boundary values more efficiently. Therefore, the computational savings and the associated loss of accuracy of these methods have to be analyzed.

For this assessment, one configuration of our model system with an inner region of 2978 atoms was selected. Spherical harmonics with multipole moments up to 20th order ($L = 0-19$), i.e., 400 basis functions, were used to represent the charge distribution. The previously determined best mesh size combination of 0.6 and 1.75 Å for the inner and outer grid was employed. The accuracy of the reaction field matrix was assessed by comparing the GSBP results for the solvation free energy of the inner region ($\Delta W^{ii}_{elec}$) to the results of a finite-difference solution of the PB equation without a basis set representation, i.e., in the complete basis set limit.

The accuracy and the costs for computation of the reaction field matrix were tested for CIR factor values of 1.0, 1.5, 2.0, 2.5, and 3.0 in combination with the standard Debye−Hückel (DH) method and the DH expression with linear interpolation (DHLI). The results are given in Table 3. The combination of a CIR factor of 1.0 with DH boundary values corresponds to the standard GSBP method that reproduced the free solvation energy very well. With the selected basis set, about 97% of the free solvation energy is recovered. These results are certainly satisfying and support the finding of Im et al. that the solvation free energy is sufficiently converged with a basis set of this size.[38] If the CIR factor is increased to 1.5, 2.0, or 2.5, the deviation increases by only 0.2 kcal/mol from 3.73 to 3.93 kcal/mol. At the same time, computational costs are reduced by 54% from 8.22 h to only 3.79 h. The DHLI method proves to be similarly efficient. With a CIR factor of 1.0 and DHLI boundary values, the deviation increases by only 0.01 kcal/mol relative to DH boundary values and the computational costs are reduced by 45% to 4.53 h. Unfortunately, these two methods cannot be combined without loss of efficiency. The DHLI method in combination with a CIR factor of 2.5 yields a deviation of 3.94 kcal/mol (i.e., 0.21 kcal/mol higher relative to the standard GSBP method), but the computational costs are merely reduced from 3.79 to 3.38 h relative to the DH method with a CIR factor of 2.5. Considering the computation time for zero boundary values, it is understandable that the CIR and the DHLI method cannot be combined without loss of efficiency. Using zero boundary values, the relative computation time drops to 40.65%. Hence, in a standard reaction field matrix computation, about 60% of the computation time is used to define the boundary values. As either method, DHLI or CIR, reduces the computational costs for this step to only a fraction, combining the two methods gives only marginal extra savings. Overall, the combination of DHLI with a CIR factor of 2.5 reduces the computational costs by about 60% with minimal loss of accuracy. We also note that the computation time for this calculation increases by 10% without modified stripping (Table 3).

***Table 3.*** Accuracy and Computational Costs of the Reaction Field Matrix Calculation Using Different Approximations To Define the Boundary Values (See Text)

| boundary[a] | CIR factor | $\Delta W^{ii}_{elec}$ (kcal/mol) | deviation[b] (kcal/mol) | time (h) | rel time (%) |
|---|---|---|---|---|---|
| DH | 1.0 | −120.27 | 3.73 | 8.22 | 100.00 |
| DH | 1.5 | −120.07 | 3.93 | 4.83 | 58.81 |
| DH | 2.0 | −120.10 | 3.91 | 4.01 | 48.78 |
| DH | 2.5 | −120.08 | 3.93 | 3.79 | 46.08 |
| DH | 3.0 | −119.21 | 4.80 | 3.56 | 43.28 |
| DHLI | 1.0 | −120.26 | 3.74 | 4.53 | 55.11 |
| DHLI | 1.5 | −120.06 | 3.95 | 3.68 | 44.80 |
| DHLI | 2.0 | −120.08 | 3.92 | 3.54 | 43.05 |
| DHLI | 2.5 | −120.06 | 3.94 | 3.38 | 41.19 |
| DHLI no MS[c] | 2.5 | −120.06 | 3.94 | 3.71 | 45.17 |
| DHLI | 3.0 | −119.20 | 4.81 | 3.33 | 40.54 |
| ZERO | − | −116.38 | 7.62 | 3.34 | 40.65 |

[a] DH, Debye−Hückel; DHLI, Debye−Hückel with linear interpolation; ZERO, all boundary values are set to zero. [b] Deviation = calcd − ref. [c] No modified stripping.

In summary, we have found that the GSBP yields reliable results for the electrostatic potential and the free solvation energy at moderate computational costs using a recommended parameter set with an inner grid spacing of 0.6 Å, an outer grid spacing of 1.75 Å, a CIR factor of 2.5, and DHLI boundary values.

Concluding this section, it seems worthwhile to reiterate the reference that was used to assess the performance of the GSBP. We have confirmed that the GSBP provides an accurate representation of the electrostatic potential that arises from the fixed outer region point charges and the PDC. This does not necessarily imply that a biomolecular simulation with the GSBP will be realistic in a chemical sense. Whether application of the GSBP is reasonable, and which choice is appropriate for physical parameters like the size of the inner region or the dielectric constants, is highly system-specific and beyond the scope of this study.

## 5. GSBP Efficiency

As the application of the GSBP is linked with a significant overhead, it is of interest to quantify the computational costs

***Table 4.*** Computation Times Related to Nontruncated Coulombic Electrostatics and the GSBP Approach for Different System Sizes[a]

| | | computation time [s] | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| radius (Å) | atoms | overhead | normal step | GSBP step | QM saving | MM saving | $Q_n/Q_n^x/\Omega/\Gamma$[b] | saving | steps[c] |
| 25.0 | 7205 | 12166.1 | 2.0 | 5.4 | −0.2 | 1.3 | 4.5 | −3.4 | − |
| 27.5 | 9632 | 12104.2 | 3.8 | 5.4 | −0.1 | 3.0 | 4.5 | −1.6 | − |
| 30.0 | 12449 | 12243.9 | 6.2 | 5.4 | 0.0 | 5.3 | 4.5 | 0.8 | 15461 |
| 32.5 | 15806 | 12538.2 | 8.8 | 5.2 | 0.1 | 7.8 | 4.3 | 3.6 | 3503 |
| 35.0 | 19670 | 12399.3 | 13.3 | 5.3 | 0.3 | 12.1 | 4.4 | 8.0 | 1555 |
| 37.5 | 24110 | 12590.8 | 18.7 | 5.3 | 0.5 | 17.3 | 4.3 | 13.4 | 937 |
| 40.0 | 29234 | 12511.0 | 25.2 | 5.4 | 0.7 | 23.6 | 4.5 | 19.8 | 633 |
| 42.5 | 35042 | 12761.3 | 33.4 | 5.8 | 0.9 | 31.6 | 4.9 | 27.6 | 462 |
| 45.0 | 41468 | 12697.4 | 43.7 | 5.3 | 1.1 | 41.6 | 4.4 | 38.4 | 331 |

[a] Single-step computation times are average values from a sample of 100 steps. [b] Computation of additional terms related to the GSBP. [c] Number of steps necessary to compensate for the GSBP overhead.

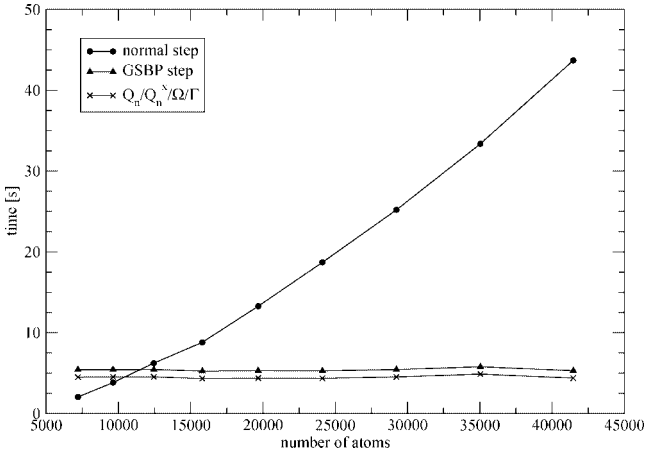



***Figure 3.*** Computation times (s) for a single MD step using nontruncated Coulombic (normal step) or GSBP electrostatics (GSBP step) as a function of the system size. Furthermore, the computation times for the GSBP-related terms ($Q_n/Q_n^x/\Omega/\Gamma$) are plotted.

and savings related to the GSBP. In this section, the efficiency of the GSBP is documented for model systems of different sizes that were generated by solving one threonine molecule in TIP3P water balls with radii increasing from 25 to 45 Å. As in the previous calculations, the inner region was centered on the $C_\alpha$ carbon of threonine and contains all molecules with any atom within 18 Å of the center. While the inner region consists of 2738 atoms for all models, the overall system size increases from 7205 atoms to 41 468 atoms with increasing radius.

A detailed analysis of the computation times related to a MD simulation using either a standard approach with nontruncated Coulombic electrostatics or the GSBP is given in Table 4 and illustrated in Figure 3. This data provides interesting insights into the applicability and efficiency of the GSBP. First of all, the computation time for the GSBP overhead, i.e., calculation of the reaction field matrix and the static field, is almost constant and increases only slightly from 3.4 to 3.5 h when increasing the system size by a factor of 6. Also the computation time of a single GSBP MD step is almost constant at 5.4 s. For a standard MD step with full electrostatics in contrast, the computation time increases from 2.0 to 43.7 s. Accordingly, impressive savings per step can be achieved if the GSBP is used for extended systems.

However, the GSBP is not always more efficient than full electrostatics. For the two smallest systems, even a single MD step is computationally more expensive with the GSBP (in addition to the initial overhead). This can be attributed to two factors. First, with the GSBP several additional terms have to be computed for each step, such as the $\Omega$ and $\Gamma$ matrices that allow interaction with the QM code, the multipole moments, $Q_n$, and their derivatives, $Q_n^x$ (see eqs 19 and 23). Especially, the computation of all multipole moment derivatives for each degree of freedom is laborious and increases the GSBP step time by roughly 4.5 s for all system sizes. Second, evaluation of the QM energy and gradient is computationally more expensive with the GSBP, since the QM part takes 0.3 s with the standard approach and 0.5 s with the GSBP. This can be traced back to the calculation of the SCF density derivatives that is not necessary in a pure QM/MM calculation. However, these factors are dominant only for small systems. With increasing system size, evaluation of the QM energy and gradient becomes more efficient in the GSBP, due to the fact that the calculation of the numerous one-electron integrals in the standard electronic embedding procedure becomes more expensive than the solution of the CPSCF equation for large systems with 12 000 atoms and more. Moreover, we note that introduction of coupled Mulliken charge derivatives (to ensure accurate gradients) increases the computational costs of the GSBP method only marginally. The computation time for the MM part remains constant at about 1.0 s when using the GSBP, providing the main contribution to the GSBP savings.

Overall, in the chosen example, we start to see minor savings for a system with 12 500 atoms. Assuming an MD step size of 1 fs, the first 15 ps of simulation time are needed to compensate for the GSBP overhead, and afterward the computation time per step is reduced by 13%. Hence, in typical semiempirical QM/MM MD simulations, the breakeven point between the GSBP and Coulombic electrostatics without truncations appears at a system size of around 12 500 atoms. Significant savings are achieved for larger systems. In simulations of the 37.5 Å system with 24 110 atoms, only 937 steps are necessary to compensate for the GSBP overhead, and subsequently, the computation time per step decreases by more than 70% from 18.7 to 5.3 s. For larger systems even more impressive savings are observed (Table

4). Since in theoretical biochemistry one is frequently interested in QM/MM simulations of biomolecular systems with 25 000 atoms and more, the GSBP method offers an efficient approach to perform such simulations at a fraction of the computational costs compared to Coulombic electrostatics without truncation.

## 6. Conclusions

In this work, we have presented the implementation of the GSBP for QM/MM approaches using NDDO-based semiempirical QM methods. Moreover, three methods to accelerate computation of the reaction field matrix were introduced: coarsening of the inner region, linear interpolation of Debye−Hückel boundary values, and modified stripping. We found that a combination of these methods reduces the computational costs for assembling the reaction field matrix by 60% with only minimal loss of accuracy. Furthermore, we studied the accuracy of the GSBP as a function of its inherent parameters, and defined a set of parameter values that offer an ideal tradeoff between accuracy and computational costs. On the basis of theses values, the computational overhead and the savings of the GSBP were quantified in QM/MM MD simulations for model systems containing from around 7000 to more than 40 000 atoms. The breakeven point where the savings in comparison to nontruncated Coulombic electrostatics roughly compensate for the overhead was determined at around 12 500 atoms. For larger systems, the GSBP showed an impressive performance. Compensation for the overhead was achieved in less than 1000 MD steps, and subsequently, the computation time per step decreased by 70% and more compared to nontruncated Coulombic electrostatics.

The GSBP is thus an efficient and accurate method to perform semiempirical QM/MM MD simulations on large biomolecular systems without neglecting or truncating long-range electrostatics if the outer layer of the inner region is fixed. It is clearly desirable to achieve similar computational savings by applying the GSBP in combination with higher-level QM/MM methods. Work in this direction is currently under way in our laboratory.

**Acknowledgment.** This work was supported by the Max Planck Initiative on Multiscale Materials Modelling. T.B. gratefully acknowledges a Kekulé scholarship from the Fonds der Chemischen Industrie.

### References

(1) Hu, H.; Yang, W. *Annu. Rev. Phys. Chem.* **2008**, *59*, 573–601.

(2) Zhang, Y.; Liu, H.; Yang, W. *J. Chem. Phys.* **2000**, *112*, 3483–3492.

(3) Cisneros, G. A.; Liu, H.; Zhang, Y.; Yang, W. *J. Am. Chem. Soc.* **2003**, *125*, 10384–10393.

(4) Ridder, L.; Rietjens, I. M. C. M.; Vervoort, J.; Mulholland, A. J. *J. Am. Chem. Soc.* **2002**, *124*, 9926–9936.

(5) Kaminski, G. A.; Jorgensen, W. L. *J. Phys. Chem. B* **1998**, *102*, 1787–1796.

(6) Acevedo, O.; Jorgensen, W. L.; Evanseck, J. D. *J. Chem. Theory Comput.* **2007**, *3*, 132–138.

(7) Rod, T. H.; Ryde, U. *J. Chem. Theory Comput.* **2005**, *1*, 1240–1251.

(8) Strajbl, M.; Hong, G.; Warshel, A. *J. Phys. Chem. B* **2002**, *106*, 13333–13343.

(9) Senn, H. M.; Thiel, S.; Thiel, W. *J. Chem. Theory Comput.* **2005**, *1*, 494–505.

(10) Kästner, J.; Senn, H. M.; Thiel, S.; Otte, N.; Thiel, W. *J. Chem. Theory Comput.* **2006**, *2*, 452–461.

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

(12) Warshel, A.; Papazyan, A. *Curr. Opin. Struct. Biol.* **1998**, *2*, 211–217.

(13) Sagui, C.; Darden, T. A. *Annu. Rev. Biophys. Struct.* **1999**, *28*, 155–179.

(14) Davis, M. E.; McCammon, J. A. *Chem. Rev.* **1990**, *90*, 509–521.

(15) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. *Science* **2004**, *303*, 186–195.

(16) Ewald, P. *Ann. Phys.* **1921**, *369*, 253–287.

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

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

(19) Nam, K.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2005**, *1*, 2–13.

(20) Gao, J.; Alhambra, C. *J. Chem. Phys.* **1997**, *107*, 1212–1217.

(21) Walker, R. C.; Crowley, M. F.; Case, D. A. *J. Comput. Chem.* **2008**, *29*, 1019–1031.

(22) Hünenberger, P. H.; McCammon, J. A. *Biophys. Chem.* **1999**, *78*, 69–88.

(23) Hünenberger, P. H.; McCammon, J. A. *J. Chem. Phys.* **1999**, *110*, 1856–1872.

(24) Kuwajima, S.; Warshel, A. *J. Chem. Phys.* **1988**, *89*, 3751–3759.

(25) Weber, W.; Hunenberger, P. H.; McCammon, J. A. *J. Phys. Chem. B* **2000**, *104*, 3668–3675.

(26) Friedman, H. L. *Mol. Phys.* **1975**, *29*, 1533–1543.

(27) Wang, L.; Hermans, J. *J. Phys. Chem.* **1995**, *99*, 12001–12007.

(28) Berkowitz, M.; McCammon, J. A. *Chem. Phys. Lett.* **1982**, *90*, 215–217.

(29) Brooks, C. L., III; Karplus, M. *J. Chem. Phys.* **1983**, *79*, 6312–6325.

(30) Brunger, A.; Brooks, C. L., III; Karplus, M. *Chem. Phys. Lett.* **1984**, *105*, 495–500.

(31) Brunger, A.; Brooks, C. L., III; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 8458–8462.

(32) Lee, F. S.; Warshel, A. *J. Chem. Phys.* **1992**, *97*, 3100–3107.

(33) Alper, H.; Levy, R. M. *J. Chem. Phys.* **1993**, *99*, 9847–9852.

(34) Essex, J. W.; Jorgensen, W. L. *J. Comput. Chem.* **1995**, *16*, 951–972.

(35) Warshel, A.; King, G. *Chem. Phys. Lett.* **1985**, *121*, 124–129.

(36) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451–5459.

(37) Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100*, 9050–9063.

(38) Im, W.; Bernèche, S.; Roux, B. *J. Chem. Phys.* **2001**, *114*, 2924–2937.

(39) Banavali, N. K.; Im, W.; Roux, B. *J. Chem. Phys.* **2002**, *117*, 7381–7388.

(40) Schaefer, P.; Riccardi, D.; Cui, Q. *J. Chem. Phys.* **2005**, *123*, 014905/1–14.

(41) Elstner, M.; Porezag, D.; Jungnickel, G.; Elstner, J.; Haugk, M.; Frauenheim, T.; Suhai, T.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.

(42) 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*, 6458–6469.

(43) Riccardi, D.; Cui, Q. *J. Phys. Chem. A* **2007**, *111*, 5703–5711.

(44) Riccardi, D.; Schaefer, P.; Cui, Q. *J. Phys. Chem. B* **2005**, *109*, 17715–17733.

(45) König, P. H.; Ghosh, N.; Hoffmann, M.; Elstner, M.; Tajkhorshid, E.; Frauenheim, T.; Cui, Q. *J. Phys. Chem. A* **2006**, *110*, 548–563.

(46) Ma, L.; Cui, Q. *J. Am. Chem. Soc.* **2007**, *129*, 10261–10268.

(47) Zhu, X.; Jethiray, A.; Cui, Q. *J. Chem. Theory Comput.* **2007**, *3*, 1538–1549.

(48) Riccardi, D.; Konig, P.; Prat-Resina, X.; Yu, H.; Elstner, M.; Frauenheim, T.; Cui, Q. *J. Am. Chem. Soc.* **2006**, *128*, 16302–16311.

(49) Gonzalez-Lafont, A.; Truong, T. N.; Truhlar, D. G. *J. Phys. Chem.* **1991**, *95*, 4618–4627.

(50) Rossi, I.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *233*, 231–236.

(51) Bash, P. A.; Ho, L. L.; MacKerell, A. D., Jr.; Levine, D.; Hallstrom, P. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 3698–3703.

(52) Lau, E. Y.; Kahn, K.; Bash, P. A.; Bruice, T. C. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 9937–9942.

(53) Ridder, L.; Rietjens, I. M. C. M.; Vervoort, J.; Mulholland, A. J. *J. Am. Chem. Soc.* **2002**, *124*, 9926–9936.

(54) Otte, N.; Scholten, M.; Thiel, W. *J. Phys. Chem. A* **2007**, *111*, 5751–5755.

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

(56) Klapper, I.; Hagstrom, R.; Fine, R.; Sharp, K.; Honig, B. *Proteins* **1986**, *1*, 47–59.

(57) Mulliken, R. S. *J. Chem. Phys.* **1962**, *36*, 3428–3439.

(58) Sherwood, P.; et al. *J. Mol. Struct. (THEOCHEM)* **2003**, *632*, 1–28.

(59) Bakowies, D.; Thiel, W. *J. Phys. Chem.* **1996**, *100*, 10580–10594.

(60) Patchkovskii, S.; Thiel, W. *J. Comput. Chem.* **1996**, *17*, 1318–1327.

(61) Yamaguchi, Y.; Osamura, Y.; Goddard, J. D.; Schaefer, H. F., III *A New Dimension to Quantum Chemistry: Analytic Derivative Methods in Ab Initio Molecular Electronic Structure Theory*; Oxford University Press: Oxford, UK, 1994; pp 128−132.

(62) Gilson, M. K.; Sharp, K. A.; Honig, B. H. *J. Comput. Chem.* **1987**, *9*, 327–335.

(63) Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435–445.

(64) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(65) Smith, W.; Forester, T. *J. Mol. Graph.* **1996**, *14*, 136–141.

(66) MacKerell, A. D., Jr.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(67) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterlig, W. T. *Numerical Recipes in C*; Cambridge University Press: London, 1988; pp 673−680.

CT800193A

JCTC Journal of Chemical Theory and Computation

# Assessment of Density Functionals for Intramolecular Dispersion-Rich Interactions

Tanja van Mourik*

*School of Chemistry, University of St. Andrews, North Haugh Fife KY16 9ST, Scotland, U.K.*

Received June 19, 2008

**Abstract:** A range of density functional theory methods, including conventional hybrid and meta-hybrid functionals, a double-hybrid functional, and DFT-D (DFT augmented with an empirical dispersion term) were assessed for their ability to describe the three minima along the $\phi_{Gly}$ rotational profile of one particular Tyr-Gly conformer. Previous work had shown that these minima are sensitive to intramolecular dispersion and basis set superposition error, the latter rendering MP2 calculations with small to medium-sized basis sets unsuitable for describing this molecule. Energy profiles for variation of the $\phi_{Gly}$ torsion angle were compared to an estimated CCSD(T)/CBS reference profile. The hybrid functionals and the meta-hybrid PWB6K failed to predict all three minima; the meta-hybrid functionals M05−2X and M06−2X and the nonhybrid meta functional M06-L as well as the double-hybrid mPW2-PLYP and the B3LYP-D method did find all three minima but underestimated the relative stability of the two with rotated C-terminus. The best performance was delivered by the most elaborate density functional theory model employed: mPW2-PLYP-D. Only M06−2X and mPW2-PLYP-D predicted the correct order of stability of the three minima.

## 1. Introduction

Intramolecular interactions with aromatic residues are known to play a crucial role in defining the secondary structure of peptides and proteins.[1] Unfortunately, such interactions are inherently difficult to describe computationally. Interactions involving $\pi$-electron clouds are affected by electrostatic as well as dispersion forces, and thus, the computational method must be able to describe these forces accurately. Most well-established density functionals like B3LYP do not describe dispersion forces correctly and are therefore not suitable for studying molecules containing aromatic rings. Unfortunately, second-order Møller−Plesset perturbation theory (MP2), which is the simplest correlated ab initio method, also has problems describing $\pi$-interactions correctly: whereas MP2 does describe both electrostatic and dispersion interactions, it produces large intramolecular BSSE (basis set superposition error) values unless very large basis sets are employed.[2−4] We encountered these problems recently while studying the tyrosyl-glycine (Tyr-Gly) dipeptide. MP2 single-point calculations (at B3LYP-optimized geometries) predicted that the six most stable conformers contain a folded "book" conformation, whereas B3LYP favored extended conformers.[5] MP2 geometry optimization of the book conformers significantly changed their geometries, increasing their degree of foldedness and stability relative to extended conformations. Further studies on the two book conformers that changed most dramatically from B3LYP to MP2 geometry optimization, labeled "book4" and "book6" in refs 2 and 3, showed that neither B3LYP nor MP2, when coupled with the medium-sized basis set 6−31+G(d), is able to predict the correct geometry of these conformers.[2,3] For book6, the much more folded structure predicted by MP2 appeared to be entirely an artifact caused by intramolecular BSSE, and for this conformer, B3LYP essentially predicted the correct structure.[3] The situation is slightly more complicated for book4. Here, the B3LYP and MP2 optimized structures mainly differ in the orientation of the C-terminus, as characterized by the $\phi_{Gly}$ Ramachandran angle (see Figure 1). Potential energy profiles were created by optimizing the

* Corresponding author phone: +44 (0)1334 463822; fax: +44 (0)1334 463808; e-mail: tanja.vanmourik@st-andrews.ac.uk.

10.1021/ct800231f CCC: $40.75 
Published on Web 09/13/2008

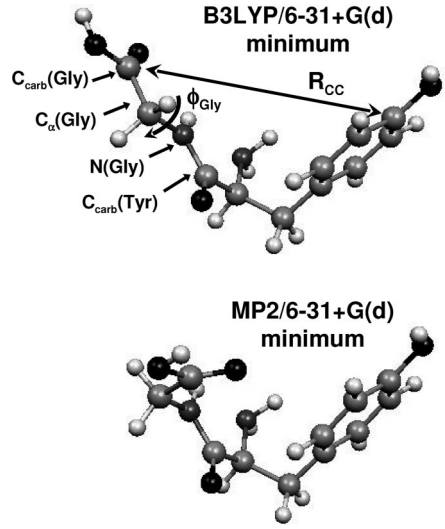



***Figure 1.*** The B3LYP/6−31+G(d) and MP2/6−31+G(d) optimized geometries of the Tyr-Gly conformer book4.

book4 structure for fixed $\phi_{Gly}$ values.[2] Single-point calculations were performed with df-MP2[6] (density-fitted MP2) and df-LMP2[6−9] (density-fitted local MP2) using large basis sets (aug-cc-pVDZ - aug-cc-pVQZ[10,11]) to reduce the BSSE. The "df" approximation significantly reduces the cost of the two-electron-four-index integrals,[6] thereby allowing the use of much larger basis sets than would be feasible with canonical MP2. The "local" approximation is also designed to reduce computational cost. In addition, this approximation reduces the size of the BSSE.[12−15] The df-(L)MP2 calculations showed that there are three minima along the energy profile. However, neither B3LYP/6−31+G(d) nor MP2/6−31+G(d) found all three minima, which was attributed to intramolecular BSSE in the MP2 calculations and missing dispersion in the B3LYP calculations. Thus, neither is a suitable level of theory to study interactions with $\pi$-electron clouds.

Similar problems with intramolecular BSSE have been encountered recently in other systems containing aromatic rings. For example, large intramolecular BSSE effects in MP2 calculations were responsible for predicting the wrong order of stability of the Phe-Gly-Phe tripeptide,[16] whereas a more extreme example of the effect of intramolecular BSSE is provided by the series of [*n*]helicenes.[16] Here, the large number of $\pi$-$\pi$ interactions per benzene ring caused such a large intramolecular BSSE effect in the MP2 calculations that clearly absurd results were obtained.

Evidently, reducing the intramolecular BSSE in MP2 calculations by using very large basis sets readily becomes intractable for large molecular systems. An alternative, computationally more efficient approach may be to make use of recent efforts to include dispersion in density functional theory (DFT). As DFT is much less basis-set dependent than MP2, intramolecular BSSE effects are much smaller. Over the last years, many new functionals have been developed to rectify the deficiencies (in particular, the inability to describe dispersion correctly) of earlier functionals. In the current work we assess a number of modern functionals for their ability to reproduce all three minima in the book4 rotational energy profile. The functionals considered fall into the following categories: hybrid GGA (generalized gradient approximation) functionals, which contain a percentage of exact Hartree−Fock (HF) exchange, and hybrid-meta GGAs, which in addition explicitly depend on the kinetic energy density. We also used one meta-GGA functional (no Hartree−Fock exchange). The hybrid GGAs considered include B3LYP,[17,18] B97−1,[19,20] X3LYP,[21] and BHandH (or BH&H).[22] B3LYP is by far the most popular functional, representing about 80% of the total occurrences in the literature over 1990−2006.[23] B97−1 was found to be among the functionals that gave the best results for a combination of thermochemical kinetics and nonbonded interactions[24,25] and was the best-performing functional without kinetic energy in a study on H-bonded and stacked structures of formic acid tetramers and formamide tetramers.[26] The X3LYP functional was designed for noncovalent interactions.[21] It describes hydrogen bonding accurately[27,28] but was found to fail for stacking interactions.[29] BHandH was found to give good results for dispersion systems but overestimates hydrogen-bonding interactions.[30] The meta-hybrid GGAs considered in this work include PWB6K,[31] M05−2X,[24] and M06−2X,[32] all originating from the Truhlar group. The nonhybrid (local) meta-GGA M06-L[33] also originates from this group. PWB6K was developed for thermochemistry and nonbonded interactions[31] and has consistently shown very good performance for noncovalent interactions.[26,34,35] The M05−2X, M06−2X, and M06-L functionals were developed by Truhlar et al. as part of two suites of functionals, the M05[31,36] and M06[33,37,38] series, intended to yield broad applicability in chemistry. The M06 suite was built on the experience gained with the M05 functionals and essentially supersedes these.[37] M06−2X was shown to perform very well for aromatic−aromatic stacking interactions,[37] though in a recent study on the glycyl-phenylalanyl-alanine peptide this functional failed to reproduce the relative order of stability of sixteen low-lying peptide conformers.[39] M06-L was found to be the only local functional that outperformed B3LYP using a test set including data for main-group thermochemistry, barrier heights, noncovalent interactions, and transition metal chemistry.[37]

We also used a double-hybrid density functional, mPW2-PLYP,[40] which was found to perform very well for weak interactions.[41] Double-hybrid functionals can be seen as a mixture of hybrid DFT and MP2: in addition to mixing in a portion of exact Hartree−Fock exchange ($E_X^{HF}$), as done in hybrid GGAs, double-hybrid functionals also mix in a fraction of MP2 correlation energy ($E_C^{MP2}$), calculated with the hybrid DFT orbitals.[42] In the case of mPW2-PLYP, the exchange and correlation functionals $E_X^{GGA}$ and $E_C^{DFT}$ are provided by mPW[43] and LYP,[18] respectively. The total exchange-correlation energy $E_{XC}$ is then given by

$$E_{XC} = (1-a)E_X^{GGA} + aE_X^{HF} + (1-b)E_C^{DFT} + bE_C^{MP2} \quad (1)$$

For mPW2-PLYP, the HF-exchange mixing parameter $a$ and the MP2 correlation mixing parameter $b$ are 0.55 and 0.25, respectively.[40]

An alternative method to overcome the deficiencies of density functionals is to augment the functional with an empirical dispersion term. In this work we have used the

DFT-D method of Grimme,[44,45] where the dispersion energy is described by a damped potential of the form $C_6R^{-6}$:

$$E_{disp} = -s_6 \sum_{i-1}^{N_{at}-1} \sum_{j=i+1}^{N_{at}} \frac{C_6^{ij}}{R_{ij}^6} f_{dmp}(R_{ij}) \quad (2)$$

Here, $N_{at}$ is the number of atoms, $C_6^{ij}$ is the dispersion coefficient for atom pair $ij$, $s_6$ is a global scaling factor only dependent on the density functional used, and $R_{ij}$ is the interatomic distance. The damping function $f_{dmp}$ is given by[44,45]

$$f_{dmp}(R_{ij}) = \frac{1}{1 + e^{-d(R_{ij}/R_r - 1)}} \quad (3)$$

Here, $R_r$ is the sum of the van der Waals radii. For mPW2-PLYP the scaling parameter $s_6$ and the damping factor $d$ are 0.40 and 20, respectively.[40]

It was shown that Grimme's parametrization yielded interaction energies that deviated on average by less than 10% from reference CCSD(T) values for a benchmark set consisting mainly of DNA base pairs and amino acid pairs.[46] For $\pi$-stacked structures DFT-D gave results in good agreement with the reference SCS-MP2 (spin-component-scaled MP2) results.[47,48] However, larger deviations were found for anisole-water and anisole-ammonia, where B3LYP-D yielded overestimated interaction energies with the ammonia or water located too close to the anisole molecule.[49] This was attributed to either an overestimated dispersion correction or to double-counting of electron correlation effects by the DFT and van der Waals parts of the method, which is expected to be worse at short-range distances.

In the current paper we show that the meta functionals (except the older PWB6K) and the double-hybrid functional as well as the DFT-D methods considered clearly outperform the conventional hybrid functionals in describing the Tyr-Gly conformer studied. The overall best agreement with the estimated CCSD(T)/CBS results is provided by the mPW2-PLYP-D method (double-hybrid functional, augmented with an empirical dispersion term).

## 2. Methodology

**2.1. DFT Energy Profiles.** The MP2/6−31+G(d) and B3LYP/6−31+G(d) geometries of the Tyr-Gly conformer book4 mainly differ in the orientation of the C-terminus, as characterized by the Ramachandran angle $\phi_{Gly}$ (equaling the $C_{carb}$(Tyr)-N(Gly)-$C_\alpha$(Gly)-$C_{carb}$(Gly) dihedral angle – see Figure 1). Energy profiles for rotation around the glycine N(Gly)-$C_\alpha$(Gly) bond were determined by single-point energy calculations at structures with $\phi_{Gly}$ values ranging from 40−310°. These structures were obtained by geometry optimization at fixed $\phi_{Gly}$ values at the M05−2X/6−31+G(d) level of theory. The relative energies used to create the energy profiles were computed relative to the energy of the conformer with $\phi_{Gly} = 180°$. The profiles were computed with the B3LYP,[17,18] B97−1,[19,20] X3LYP,[21] Gaussian's version of Becke's half-and-half functional BHandH,[22] mPW2-PLYP,[40] PWB6K,[31] M05−2X,[24] M06−2X,[32] M06-L,[33] B3LYP-D,[44,45] and mPW2-PLYP-D[50] density functional methods. The M05−2X profiles were computed with the 6−31+G(d),[51−53] aug-cc-pVDZ, and aug-cc-pVTZ[10,11] basis sets. All other DFT profiles were calculated with aug-cc-pVDZ only. The X3LYP, B97−1, and BHandH calculations were done with Gaussian 03,[54] the PWB6K, M05−2X, M06−2X, and M06-L calculations were done with NWChem,[55] whereas the mPW2-PLYP, mPW2-PLYP-D, B3LYP, and B3LYP-D calculations were performed with ORCA.[56] The mPW2-PLYP calculations invoked the RI (resolution of the identity) approximation (similar to the density-fitting approach) for the MP2 part, using automatic generation of a general-purpose fitting basis set. The B3LYP calculations employed the VWN1[57] correlation functional (Gaussian's definition of the B3LYP functional). The Gaussian calculations used the "UltraFine" integration grid (containing 99 radial shells and 590 angular points per shell), the NWChem calculations employed the "xfine" grid (125 radial and 1454 angular shells), and the ORCA calculations employed "grid 6" (default GaussChebyshev radial grid coupled with 590 angular Lebedev points).

**2.2. The Reference Profile.** Single-point df-LMP2[6−9] calculations were performed at the M05−2X/6−31+G(d) geometries using the aug-cc-pVDZ and aug-cc-pVTZ basis sets. The corresponding aug-cc-pVDZ-MP2fit and aug-cc-pVTZ-MP2fit fitting basis sets[58] were used for both the df-HF and df-LMP2 parts of the calculation. The profiles were also computed with df-HF/aug-cc-pVQZ, whereas df-MP2 and df-LMP2 calculations with the aug-cc-pVQZ/aug-cc-pVQZ-MP2fit basis set combination were carried out for selected geometries only ($\phi_{Gly}$ = 80, 130, 180, 240, and 285°). In the local calculations all pairs were treated as strong pairs, as recommended to avoid discontinuities on the potential energy surface due to orbital domain changes.[59,60] In addition, single-point calculations with df-LCCSD-(T0)[12,61−63] (density-fitted local coupled cluster with single, double and perturbative noniterative local triple excitations) were performed with the aug-cc-pVDZ/aug-cc-VDZ-MP2fit basis set combination. Df-LCCSD(T0) calculations with aug-cc-pVTZ/aug-cc-pVTZ-MP2fit were done for $\phi_{Gly} = 180$ and 285°, to provide a one-point test of the $E_{CCSD(T)corr}$/aug-cc-pVDZ higher-order correlation correction term (see below). The default selection of the pair classes was used. In the df-LMP2 and df-LCCSD(T0) calculations the two most diffuse functions of each angular momentum function were ignored in the localization to yield better-localized orbitals. A completion criterion of 0.99 was employed for the orbital domain selection. Despite treating all pairs as strong pairs in the LMP2 calculations, at some points along the df-LMP2 and df-LCCSD(T0) profiles the orbital domains changed slightly, leading to steps in the potential energy curve. We redid those calculations using the same orbital domains as used for $\phi_{Gly} = 180°$, except in the case of df-LMP2/aug-cc-pVQZ, where the orbital domains of $\phi_{Gly} = 80°$ were used, as these were the same as those of $\phi_{Gly}$ = 130, 240, and 285° (i.e., only the 180°-domains were different). The df-LMP2 and df-LCCSD(T0) calculations were done with Molpro 2006.[64]

Complete basis set (CBS) CCSD(T) limits were estimated as follows: the aug-cc-pVDZ and aug-cc-pVTZ df-LMP2 correlation energies, and for selected geometries also the aug-

cc-pVTZ/aug-cc-pVQZ correlation energies, were extrapolated to the CBS limit using the two-point extrapolation scheme of Halkier et al.[65]

$$E_{\mathrm{MP2correl,CBS}} = \frac{X^3}{X^3-(X-1)^3}E_{\mathrm{correl},X} - \frac{(X-1)^3}{X^3-(X-1)^3}E_{\mathrm{correl},X-1} \quad (4)$$

Here, $X$ is the cardinal number of the largest basis set used in the extrapolation ($X = 3$ for aug-cc-pVDZ/aug-cc-pVTZ extrapolation; $X = 4$ for aug-cc-pVTZ/aug-cc-pVQZ extrapolation). The extrapolated correlation energy contributions were then added to the df-HF/aug-cc-pVQZ total energies. A higher-order correlation correction term, $E_{\mathrm{CCSD(T)corr}}$, was added by calculating the difference between the df-LMP2 and df-LCCSD(T0) total energies computed with the aug-cc-pVDZ basis set. It was shown previously that the basis set dependence of the CCSD(T) correction term for interaction energies is small,[66−68] and therefore, the aug-cc-pVDZ basis set should give an accurate estimate of the higher-order correlation correction.

Thus, the total energies required for the estimated CCSD(T)/CBS reference profile were computed as

$$E_{\mathrm{CCSD(T)/CBS}} = E_{\mathrm{HF}}(\mathrm{avqz}) + E_{\mathrm{MP2correl,CBS}}(\mathrm{avdz/avtz}) + E_{\mathrm{CCSD(T)corr}}(\mathrm{avdz}) \quad (5)$$

where avnz (n = d, t, q) is an abbreviated notation for aug-cc-pVnZ (n = D, T, Q). As above, the relative energies for the CCSD(T)/CBS profile were computed relative to the value at $\phi_{\mathrm{Gly}} = 180°$.

## 3. Results

**3.1. The B3LYP/6−31+G(d) and MP2/6−31+G(d) Profiles.** In previous work we had found that B3LYP/6−31+G(d) geometry optimization predicted an extended structure of the Tyr-Gly conformer book4, with a $\phi_{\mathrm{Gly}}$ angle of 180°, whereas MP2/6−31+G(d) geometry optimization yielded a more folded structure, with a $\phi_{\mathrm{Gly}}$ angle of 74°.[5] Figure 1 contrasts the B3LYP and MP2 optimized structures. The MP2 structure is more compact, with closer contact between the glycine/C-terminus and the tyrosine aromatic ring. In further work on this molecule, relaxed potential energy profiles at fixed values of $\phi_{\mathrm{Gly}}$, computed at these two levels of theory, were compared to the profile computed with df-LMP2/aug-cc-pVTZ.[2] These profiles, constructed from the work presented in ref 2, are shown in Figure 2. The df-LMP2/aug-cc-pVTZ method is expected to give the correct profile, as it is nearly BSSE-free, describes dispersion, and was found to produce relative energies in close agreement with df-LCCSD(T0) results for the Tyr-Gly conformer book6[3] (indicating that a possible overestimation of dispersion by the MP2 method, as is often seen in stacking and other weakly bound interactions,[69] is very small for this molecule).

The df-LMP2/aug-cc-pVTZ profile, displayed in Figure 2, clearly shows three minima at roughly 80, 180, and 280°. The minimum at 80° is the global minimum, whereas the 280°-minimum is shallowest. The MP2/6−31+G(d) profile reproduces the minima at 80 and 280° but completely misses the 180°-minimum. In previous work we showed that this is due to large intramolecular BSSE effects in the MP2/6−31+G(d) calculations.[2] In contrast, the B3LYP profile only shows the minimum at 180°. The other two minima are absent, presumably due to missing dispersion in the B3LYP calculations.

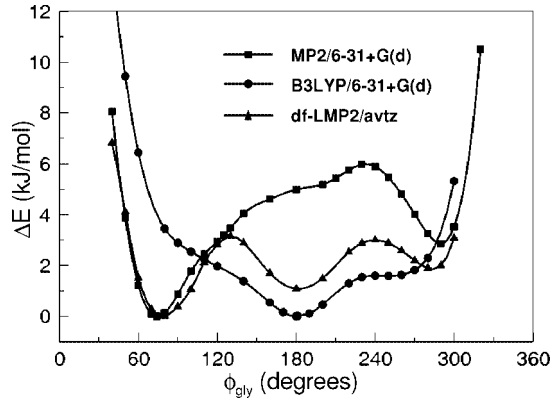



***Figure 2.*** B3LYP/6−31+G(d), MP2/6−31+G(d), and df-LMP2/avtz potential energy profiles for rotation around the $\phi_{\mathrm{Gly}}$ N(Gly)-$C_\alpha$(Gly) bond (avtz = aug-cc-pVTZ). The B3LYP profile used structures optimized with B3LYP/6−31+G(d) at fixed $\phi_{\mathrm{Gly}}$ angles. The MP2 and df-LMP2 profiles were computed using structures optimized with MP2/6−31+G(d) at fixed $\phi_{\mathrm{Gly}}$ angles. The minimum energy points in the profiles (B3LYP: $\phi_{\mathrm{Gly}} = 180°$; MP2: $\phi_{\mathrm{Gly}} = 74°$; df-LMP2: $\phi_{\mathrm{Gly}} = 80°$) were taken as the reference point for the relative energies.

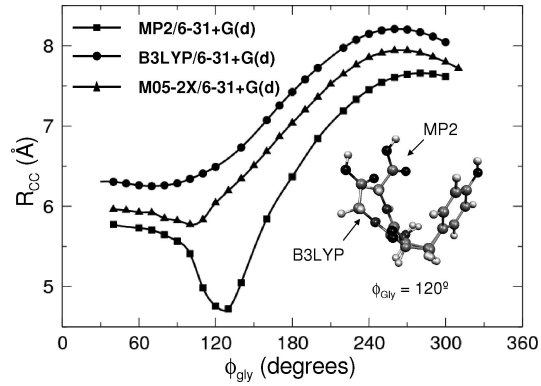



***Figure 3.*** Variation of the $R_{CC}$ distance as a function of the $\phi_{\mathrm{Gly}}$ torsion angle in the partially optimized structures obtained with B3LYP/6−31+G(d), MP2/6−31+G(d), and M05−2X/6−31+G(d). The inset shows a comparison of the B3LYP and MP2 geometries at $\phi_{\mathrm{Gly}} = 120°$.

**3.2. Tyr-Gly Geometries along the Energy Profile.** In our previous work,[2] the B3LYP profile was computed using partially optimized structures (at fixed values of $\phi_{\mathrm{Gly}}$) obtained with B3LYP/6−31+G(d), whereas the MP2 and df-LMP2 profiles used the partially optimized MP2/6−31+G(d) geometries. However, the MP2 and B3LYP structures differ to some extent. In the MP2 structures the glycine/C-terminus chain is closer to the tyrosine ring, as exemplified by the distance $R_{CC}$ between the $C_{carb}$(Gly) and tyrosine C(OH) atoms, shown in Figure 3. Over the whole $\phi_{\mathrm{Gly}}$ range the $R_{CC}$ distances are shorter in the MP2 structures

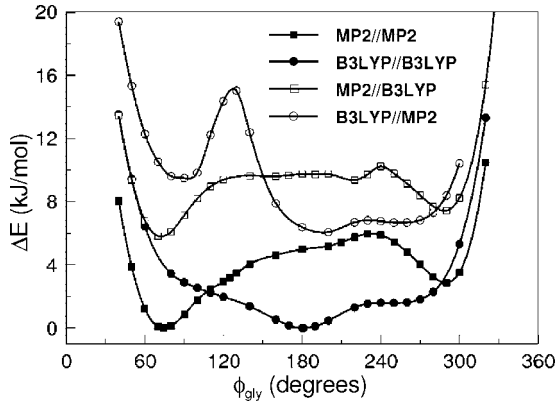



***Figure 4.*** B3LYP/6−31+G(d) and MP2/6−31+G(d) potential energy profiles for rotation around the N(Gly)-$C_\alpha$(Gly) bond, using the B3LYP/6−31+G(d) and MP2/6−31+G(d) sets of geometries. In "Method1//Method2" Method1 is the method used for the calculation of the single-point energies, whereas Method2 is the method used to obtain the geometries. The minimum energy points in the profiles (B3LYP//B3LYP: $\phi_{Gly} = 180°$; MP2//MP2: $\phi_{Gly} = 70°$) were taken as the reference point for the relative energies.

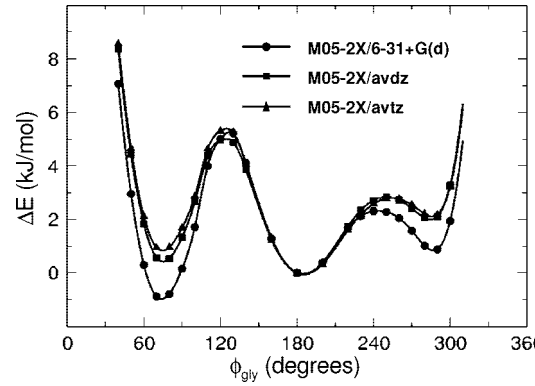



***Figure 5.*** M05−2X potential energy profiles for rotation around the N(Gly)-$C_\alpha$(Gly) bond computed with the 6−31+G(d), aug-cc-pVDZ (avdz), and aug-cc-pVTZ (avtz) basis sets, using the MP2/6−31+G(d) set of geometries. The energy at $\phi_{Gly} = 180°$ was taken as the reference point for the relative energies.

than in the B3LYP structures. The MP2 $R_{CC}$ profile shows a deep dip around $\phi_{Gly} = 120-130°$, and it is in this region that the MP2 and B3LYP structures differ most. The more compact MP2 structures in this region are likely a result of intramolecular BSSE, which was previously shown to exhibit a sharp peak around 120°.[2]

Figure 4 shows that the shape of the B3LYP and MP2 profiles depends on the set of partially optimized geometries (MP2 or B3LYP) used. Naturally, the B3LYP geometries are not ideal for the MP2 calculations and vice versa, as can be seen by the upward shift of the curves that use the other method's geometries. The shapes of the two MP2 profiles do not differ much, though the profile using the MP2 geometries is more energetically favorable than the one using the B3LYP geometries in the region around 120°, presumably due to the more compact structures predicted by MP2 in this region. The two B3LYP profiles, however, differ considerably. The 180°-minimum in the profile using the B3LYP geometries has shifted toward ∼200° in the profile using the MP2 geometries and has become shallower. The B3LYP profiles differ most dramatically in the region around 100−130°. The profile using the MP2 geometries now exhibits a second minimum at ∼90° with a large barrier separating the 90- and 200°-minima. The greater compactness of the MP2 structures in the region around 120° (cf. Figure 3) clearly leads to less favorable B3LYP energies, pushing up the curve around 120° and thereby creating the 90°-minimum. The MP2 energies appear less dependent on the compactness of the geometries, possibly because any repulsive energy contributions due to conformational strain or repulsive interactions between close atoms in the more compact structures (present in the MP2 as well as B3LYP calculations) are more than compensated by intramolecular dispersion and/or BSSE (mostly absent in the B3LYP calculations). The difference profile of the two B3LYP curves has a shape closely resembling the difference profile of the $R_{CC}$ distances in the MP2 and B3LYP structures, indicating that the shorter $R_{CC}$ distances in the MP2 structures are directly responsible for the large changes in the B3LYP profile (Figure S1, Supporting Information).

The Tyr-Gly geometries should preferentially be optimized using a (virtually) BSSE-free method that also describes dispersion. The B3LYP structures are probably not folded enough due to missing dispersive attraction, whereas the MP2 geometries are likely too compact due to BSSE. As we encountered problems performing partial geometry optimizations with df-LMP2 using Molpro, we instead reoptimized the Tyr-Gly geometries at fixed $\phi_{Gly}$ torsion angles with M05−2X/6−31+G(d) using NWChem. The M05−2X functional was shown to give a benzene-methane binding energy curve in excellent agreement with CCSD(T),[32] whereas MP2 overestimated the strength of the complex and B3LYP gave a repulsive potential. We therefore expect M05−2X to give reliable geometries for the Tyr-Gly book conformers. Figure 3 shows that the M05−2X geometries are more compact than the B3LYP geometries (as expected from the presence of dispersion) but do not show the sharp increase in compactness around 120−130° as exhibited by the MP2 structures (which is likely due to BSSE). The geometries therefore appear very plausible. Unless stated otherwise, all subsequent results were obtained using the M05−2X/6−31+G(d) geometries.

**3.3. Basis Set Convergence of the M05−2X Profile.** Figure 5 shows M05−2X profiles computed with three different basis sets, using the MP2/6−31+G(d) geometries. All profiles nicely show three minima, demonstrating the functional's superiority compared to B3LYP. The 6−31+G(d) basis set appears to overestimate the stability of the minima at ∼80° and ∼280°, as compared to the results obtained with the larger basis sets. The aug-cc-pVDZ and aug-cc-pVTZ curves are very similar, indicating that the results are essentially converged at the aug-cc-pVDZ basis set level. The remainder of this study therefore employed the aug-cc-pVDZ basis set.

**3.4. The CCSD(T) Reference Profile.** The CCSD(T) reference profile was obtained from the total energies computed according to eq 4, using the M05−2X/6−31+G(d) geometries. The reference profile has four main sources of

***Table 1.*** Comparison of the Relative Energies (in kJ/mol) of Key Structures Computed at Different Levels of Theory[a]

| | $\phi_{Gly}$ (in deg) | | | |
|---|---|---|---|---|
| method | 80 | 130 | 240 | 285 |
| df-HF/avdz | 5.44 | 3.29 | 3.03 | 2.04 |
| df-HF/avtz | 6.13 | 3.29 | 2.90 | 2.53 |
| df-HF/avqz | 6.33 | 3.26 | 2.86 | 2.53 |
| df-LMP2/avdz | −1.04 | 1.75 | 2.21 | 0.31 |
| df-LMP2/avtz | −1.10 | 1.55 | 2.00 | 0.88 |
| df-LMP2/avqz | −1.00 | 1.50 | 2.00 | 1.04 |
| LMP2/CBS(avdz/avtz)[b] | −1.22 | 1.43 | 1.93 | 0.92 |
| LMP2/CBS(avtz/avqz)[c] | −1.06 | 1.49 | 2.03 | 1.15 |
| MP2/CBS(avtz/avqz)[d] | −1.24 | 1.60 | 2.17 | 1.37 |
| df-LCCSD(T0)/avdz | −1.36 | 1.65 | 2.12 | −0.29 |
| df-LCCSD(T0)/avtz | — | — | — | 0.46 |
| LCCSD(T0)/CBS(avdz/avtz)[e] | −1.54 | 1.34 | 1.85 | 0.32 |
| LCCSD(T0)/CBS(avtz/avqz)[f] | −1.38 | 1.39 | 1.94 | 0.55 |

[a] The energy at $\phi_{Gly} = 180°$ is taken as the reference point for the relative energies; avdz = aug-cc-pVDZ, avtz = aug-cc-pVTZ, and avqz = aug-cc-pVQZ. The "df" designation is omitted from the CBS entries, as the density fitting approximation should not affect the estimated CBS limits noticeably. [b] LMP2/CBS limit estimated by adding to the df-HF/aug-cc-pVQZ energies the df-LMP2 correlation energy extrapolated using the aug-cc-pVDZ and aug-cc-pVTZ values. [c] LMP2/CBS limit estimated by adding to the df-HF/aug-cc-pVQZ energies the df-LMP2 correlation energy extrapolated using the aug-cc-pVTZ and aug-cc-pVQZ values. [d] MP2/CBS limit estimated by adding to the df-HF/aug-cc-pVQZ energies the df-MP2 correlation energy extrapolated using the aug-cc-pVTZ and aug-cc-pVQZ values. [e] LCCSD(T0)/CBS limit estimated by adding the CCSD(T) correction term, computed with aug-cc-pVDZ, to the LMP2/CBS (avdz/avtz) energies. [f] LCCSD(T0)/CBS limit estimated by adding the CCSD(T) correction term, computed with aug-cc-pVDZ, to the LMP2/CBS (avtz/avqz) energies.

uncertainty: (i) the degree of basis-set convergence of the HF/aug-cc-pVQZ energies, (ii) the accuracy of the aug-cc-pVDZ/aug-cc-pVTZ extrapolation of the MP2 correlation energies, (iii) the accuracy of the CCSD(T) higher-order correlation correction term, and (iv) the accuracy of the local approximation. Errors due to the density fitting approximation are essentially negligible.[6]

Comparison of the df-HF energy profiles computed with aug-cc-pV(D/T/Q)Z shows that the aug-cc-pVDZ basis set slightly overestimates the relative stability of the 80- and 285°-minima (Figure S2a, Supporting Information). However, the aug-cc-pVTZ and aug-cc-pVQZ profiles are very similar (relative energy differences between 0.1−0.4 kJ/mol for $\phi_{Gly}$ in the 40−100° and 285−310° intervals and nearly 0 kJ/mol in the 100−285° range), indicating that the aug-cc-pVQZ energies are sufficiently converged with respect to the basis set quality. Note that the smaller relative energy differences in the midregion are a direct result of our choice to position the profiles at 0 kJ/mol for $\phi_{Gly} = 180°$. The close similarity of the aug-cc-pVTZ and aug-cc-pVQZ results can also be deduced from the relative energies of the minimum- and maximum-energy points (80, 130, 180, 240, and 285°) along the profile listed in Table 1. These are indicative of the relative stability of the three minima and of the barriers between them. Differences between aug-cc-pV5Z and aug-cc-pVQZ energies would probably be roughly half of those between aug-cc-pVQZ and aug-cc-pVTZ, so that the HF/aug-cc-pVQZ relative energies are estimated to be accurate within ∼0.2 kJ/mol (for $\phi_{Gly} = 40-60°$) and considerably more accurate for larger $\phi_{Gly}$ angles. The HF/aug-cc-pVQZ relative energies of the 80- and 285°-minima are estimated to be too small by approximately 0.14 and 0.04 kJ/mol, respectively.

Due to the large amount of computational resources required for df-LMP2/aug-cc-pVQZ calculations, these were only performed for the minimum- and maximum-energy points (Table 1). Df-LMP2/aug-cc-pVDZ significantly overestimates the relative stability of the 285°-minimum (see also Figure S2b, Supporting Information). Extrapolation to the CBS limit (using the aug-cc-pVDZ and aug-cc-pVTZ energies) remedies this: the avdz/avtz-extrapolated relative energy of the 285°-structure is almost identical to that computed with df-LMP2/aug-cc-pVQZ. Reassuringly, the CBS(avdz/avtz) and CBS(avtz/avqz) results are very close to each other (differences of at most 0.2 kJ/mol for the minimum and maximum points). Also note that the same overestimation of the relative stability of the 285°-minimum by aug-cc-pVDZ occurs for the df-LCCSD(T0) method (Table 1). The df-LCCSD(T0) relative energy changes from a negative value (−0.29 kJ/mol) when computed with aug-cc-pVDZ to a positive value (0.46 kJ/mol) computed with aug-cc-pVTZ. We estimate that the CBS limits are accurate within ∼0.2 kJ/mol. Based on the differences between the avdz/avtz and avtz/avqz extrapolated results, the avdz/avtz extrapolation likely overestimates the relative stability of the 80- and 285°-minima by approximately 0.2 and 0.12 kJ/mol, respectively.

Sinnokrot and Sherrill found that the CCSD(T) correlation correction term for interaction energies is rather insensitive to the basis set size, as long as the basis set contains diffuse functions.[67] Thus, for the three different benzene dimer configurations (sandwich, T-shaped, and parallel-displaced), the correction term differed by less than 0.2 kJ/mol for the aug-cc-pVDZ or aug-cc-pVTZ(-f/-d) basis sets. Jurečka and Hobza found larger differences (between 0.2 and 0.5 kJ/mol) for CCSD(T) correction terms computed with aug-cc-pVDZ and cc-pVTZ, for dimers consisting of formamide and formamidine units.[66] However, the cc-pVTZ results are expected to be less accurate than the aug-cc-pVDZ ones (because of lacking diffuse functions in cc-pVTZ), so that the larger differences between the correction terms computed with these two basis sets are probably mainly due to errors in the cc-pVTZ rather than in the aug-cc-pVDZ values. To assess the accuracy of the $E_{CCSD(T)corr}$ term for the current molecule, we recomputed with aug-cc-pVTZ its contribution to the relative energy of the conformer with $\phi_{Gly} = 285°$ (which is the conformer minimum with the largest $E_{CCSD(T)corr}$ correction term). The difference between the $E_{CCSD(T)corr}$/aug-cc-pVDZ and $E_{CCSD(T)corr}$/aug-cc-pVTZ contribution to the relative energy of this structure is only 0.18 kJ/mol. In agreement with the results of Sinnokrot and Sherrill, we therefore estimate the aug-cc-pVDZ correction terms to be accurate within ∼0.2 kJ/mol. The $E_{CCSD(T)corr}$/aug-cc-pVDZ correction likely overestimates the relative stability of the 80- and 285°-minima by approximately 0.1 and 0.2 kJ/mol, respectively.

The error introduced by the local approximation cannot be simply deduced from the differences in the local and canonical energies calculated with finite basis sets, as these are partially caused by the much reduced BSSE in the local

calculations. In previous work we showed that df-MP2/aug-cc-pVTZ calculations on the book4 conformer produce large BSSE values of 5−6 kJ/mol, and even with the aug-cc-pVQZ basis set the BSSE still amounts to 2−3 kJ/mol.[2] However, when employing the local approximation the BSSE is below 1 kJ/mol already for the aug-cc-pVTZ basis set. Thus, the local energies may in fact be more reliable than the corresponding nonlocal energies, unless very large basis sets are employed. However at the CBS limit the BSSE is nonexistent, and we therefore estimated the error of the local approximation by comparing the MP2/CBS and LMP2/CBS limits (see Table 1). Relative to the 180°-minimum, the df-LMP2 method slightly overestimates the relative stability of the 285°-structure (by 0.18 kJ/mol), whereas the relative stability of the 80°-structure is slightly underestimated (by 0.22 kJ/mol). The local error in the $E_{CCSD(T)corr}$ term is likely much smaller, due to expected cancelation of errors in the df-LMP2 and df-LCCSD(T0) results.

From this analysis, we expect the reference relative stabilities of the 80°- and 285° minima to be overestimated by approximately 0.3 and 0.6 kJ/mol, respectively.

**3.5. Performance of Density Functionals.** Figure 6 shows the profiles computed with the different density functional theory methods considered in this work. None of the hybrid functionals yields the correct profile shape (Figure 6a). BHandH finds a minimum around 70°; B3LYP, X3LYP, and B97−1 show a very shallow minimum around 90°; but all hybrid functionals miss the 280°-minimum. The meta functionals perform much better (Figure 6b). The PWB6K profile does not show a minimum in the 280° area, but the M05 and M06-type functionals tested, including the local M06-L functional, nicely predict three minima along the profile, though the relative stability of the 80°- and 280°-minima are underestimated as compared to the reference profile. Only M06−2X predicts the 80°-minimum to be the most stable of the three minima, in agreement with the reference profile. However, the barrier between the 80°- and 180°-minima is much overestimated. The double-hybrid mPW2-PLYP functional and B3LYP-D also underestimate the relative stability of the 80°- and 280°-minima, but the overestimation of the barrier heights is less severe than for the meta functionals (Figure 6c). The mPW2-PLYP-D method performs best, particularly in the 60−240° region. It has been shown that for dispersion complexes double-hybrid functionals capture only about 50% of the interaction energy.[70] It is therefore not surprising that adding a dispersion term to mPW2-PLYP improves the performance of this functional. Even so, like all density functional methods, mPW2-PLYP-D underestimates the relative stability of the 280°-minimum.

The van der Waals/dispersion contributions computed by the DFT-D methods are displayed in Figure 7. Also shown is the df-LMP2/aug-cc-pVTZ correlation energy. Not surprisingly, the van der Waals term is smaller for mPW2-PLYP than for B3LYP, as the nonlocal perturbation term in the double-hybrid functional already accounts for part of the van der Waals energy. For both functionals, the van der Waals term is largest around ∼60°, where the contact between the C-terminus and the aromatic ring is closest (cf. Figure 3),

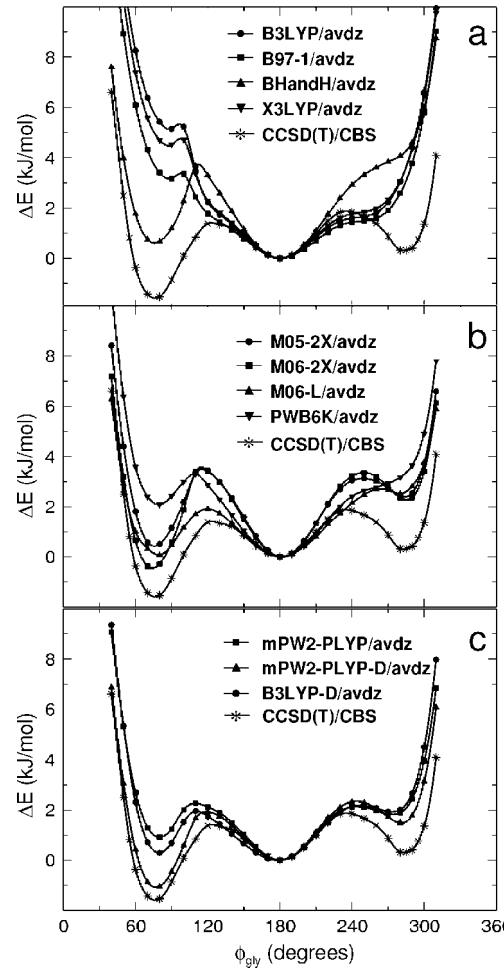



***Figure 6.*** Potential energy profiles for rotation around the N(Gly)-$C_\alpha$(Gly) bond computed with different density functionals and the aug-cc-pVDZ (avdz) basis set, using the M05−2X/6−31+G(d) set of geometries. The energy at $\phi_{Gly} = 180°$ was taken as the reference point for the relative energies. (a) Profiles computed with the hybrid functionals B3LYP, B97−1, BHandH, and X3LYP. (b) Profiles computed with the meta functionals PWB6K, M05−2X, M06−2X, and M06-L. (c) Profiles computed with mPW2-PLYP and B3LYP-D.

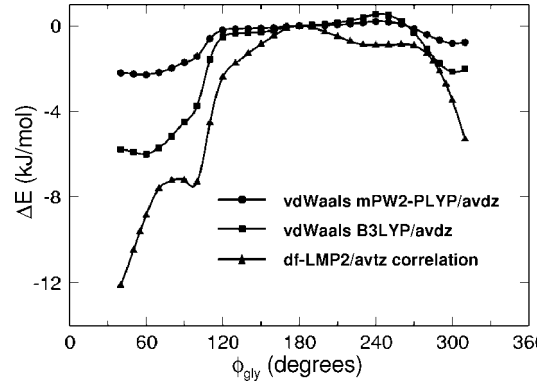



***Figure 7.*** A comparison of the van der Waals contributions for B3LYP-D and mPW2-PLYP-D and the df-LMP2/aug-cc-pVTZ correlation contribution. The energy at $\phi_{Gly} = 180°$ was taken as the reference point for the relative energies.

and where the dispersion energy can be expected to be large. It should be noted, however, that the van der Waals

contributions cannot be interpreted as pure dispersion energy, but rather, the $R^{-6}$ term in combination with the damping function corrects for DFT's general incorrect description of weakly bonded systems,[45,71] i.e. corrections other than for dispersion were absorbed in the fitted $R^{-6}$ term coefficients and damping function parameters. However, the close resemblance of the MP2 correlation energy and the DFT-D van der Waals contributions indicate that the latter can be largely interpreted as dispersion. This is corroborated by recent work by Grimme et al.,[72] who used a partitioning of the interfragment MP2 correlation energy into electron pairs of different orbital type to study the intramolecular interaction in the same Tyr-Gly conformer as studied in the current work. It was shown that the MP2 correlation energy is mainly determined by the interfragment contribution, which can be interpreted as dispersion energy. Note that the different shapes of the HF, DFT-D dispersion, and MP2 correlation energy in the paper by Grimme et al. (specifically, the sharp peak at ~120° in the HF curve and the dips in the dispersion and correlation energy curves around 120°) are mainly due to the use of MP2/6−31+G(d) geometries in the study by Grimme et al., which, as shown above, exhibit very compact structures in the region around $\phi_{Gly} = 120°$.

## 4. Conclusions

We have tested a range of density functional theory methods for their ability to describe three minima along the $\phi_{Gly}$ profile of the Tyr-Gly conformer book4. These include one minimum with an extended glycine/C-terminus chain ($\phi_{Gly} = 180°$) and two more compact structures with a rotated C-terminus ($\phi_{Gly} = \sim 80$ and 280°). Previous work had shown that this is a demanding test for electronic structure methods: MP2 calculations with a medium-sized basis set miss the 180°-minimum because the potential energy surface is distorted by large intramolecular basis set superposition errors, whereas B3LYP misses the other two minima because of lacking dispersive interactions. Potential energy curves as a function of $\phi_{Gly}$ were compared to an estimated CCSD(T)/CBS reference profile. These calculations employed geometries optimized with M05−2X/6−31+G(d) for fixed $\phi_{Gly}$ values between 40 and 310°. We show here that the conventional hybrid functionals B3LYP, B97−1, BHandH, and X3LYP as well as the meta-hybrid PWB6K fail to predict all three minima; the meta-hybrid functionals M05−2X and M06−2X and the nonhybrid meta functional M06-L, on the other hand, do find all minima but underestimate the relative stability of the two with rotated C-terminus. The mPW2-PLYP double-hybrid functional and B3LYP-D (B3LYP augmented with an empirical dispersion term) slightly outperform the meta functionals by predicting barrier heights closer to those of the reference functional. However, also these underestimate the relative stability of the 80°- and 280°-minima. The best performance is delivered by the most elaborate density functional theory method tested: the double-hybrid functional augmented by an empirical dispersion term (mPW2-PLYP-D). mPW2-PLYP-D predicts the relative stability of the 80°-minimum in very close agreement with the reference profile, though the relative stability of the 280°-minimum is still slightly underestimated, even when allowing for the projected overestimation of approximately 0.4 kJ/mol of the stability of this point by the CCSD(T) reference profile. Only M06−2X and mPW2-PLYP-D predict the correct order of stability of the three minima (80°-minimum most stable, 280°-minimum least stable). It should be noted that the dispersion and correlation corrections in the mPW2-PLYP-D method only very slightly increase the computational cost: in the single-point calculations we performed, ~90% of the CPU time was spent on the calculation of the mPW-LYP energy, ~10% on the MP2 correlation correction, and a negligible percentage on the dispersion correction.

The geometries optimized at fixed $\phi_{Gly}$ values were found to be much dependent on the level of theory used. MP2/6−31+G(d) calculations obtained more compact structures than B3LYP/6−31+G(d), particularly around $\phi_{Gly} = 130°$. As previous work had shown that the intramolecular BSSE is particularly large in this region,[2] the more compact structures predicted by MP2 are probably mainly a result of this error and to a lesser extent due to dispersion forces; both effects are largely missing in the B3LYP calculations. The M05−2X geometries used to create the potential energy profiles are more compact than the B3LYP geometries (as expected from the presence of dispersion) but do not show the sharp increase in compactness around 120−130° as exhibited by the MP2 structures (because of the much smaller BSSE in DFT calculations). The M05−2X geometries are expected to be the most accurate of the three sets of geometries.

The compact MP2 geometries around 130° are not favorable for B3LYP calculations, with the result that a peak around 130° appears in the B3LYP energy profile when the MP2 structures are used. This is probably caused by repulsive energy contributions due to conformational strain and/or repulsive interactions between close atoms in the more compact MP2 structures. The MP2 profile, on the other hand, is only very slightly affected by the choice of geometries. The insensitivity of the MP2 results to the compactness of the structures is probably because the larger repulsive energy contributions in the more compact structures are more than compensated by intramolecular dispersion and/or BSSE (mostly absent in the B3LYP calculations).

The current study highlights the difficulty of reliably describing flexible molecules where intramolecular interactions with a $\pi$-electron system can be anticipated. Such interactions are affected by intramolecular BSSE (rendering MP2 methods with small to medium-sized basis sets unsuitable) and intramolecular dispersion-type interactions (providing a challenge for DFT methods). The performance of several modern DFT methods to describe such systems is quite promising.

**Acknowledgment.** We gratefully acknowledge the Royal Society for their support under the University Research Fellowship scheme and EaStCHEM for computational support via the EaStCHEM Research Computing Facility.

**Supporting Information Available:** The $R_{CC}$ difference between B3LYP/6−31+G(d) and MP2/6−31+G(d) geometries and the energy penalty for computing the B3LYP and MP2 profiles using the geometries optimized with the

other method as a function of $\phi_{Gly}$ (Figure S1); potential energy profiles for rotation around the N(Gly)-$C_\alpha$(Gly) bond computed with df-HF and df-LMP2 (Figure S2); Cartesian coordinates of the structures optimized at fixed $\phi_{Gly}$ angles using M05−2X/6−31+G(d) (Table S1); and total energies of the Tyr-Gly conformer book4 at fixed $\phi_{Gly}$ values, computed at different levels of theory (Tables S2a-e, S3a-d, and S4a-b). This material is available free of charge via the Internet at http://pubs.acs.org.

## References


(1) Steiner, T.; Koellner, G. *J. Mol. Biol.* **2001**, *305*, 535–557.

(2) Holroyd, L. F.; van Mourik, T. *Chem. Phys. Lett.* **2007**, *442*, 42–46.

(3) Shields, A. E.; van Mourik, T. *J. Phys. Chem. A* **2007**, *111*, 13272–13277.

(4) van Mourik, T.; Karamertzanis, P. G.; Price, S. L. *J. Phys. Chem. A* **2006**, *110*, 8–12.

(5) Toroz, D.; van Mourik, T. *Mol. Phys.* **2006**, *104*, 559–570.

(6) Werner, H.-J.; Manby, F. R.; Knowles, P. J. *J. Chem. Phys.* **2003**, *118*, 8149–8160.

(7) Schütz, M.; Hetzer, G.; Werner, H.-J. *J. Chem. Phys.* **1999**, *111*, 5691–5707.

(8) Hetzer, G.; Pulay, P.; Werner, H.-J. *Chem. Phys. Lett.* **1998**, *290*, 143–149.

(9) Hetzer, G.; Schütz, M.; Stoll, H.; Werner, H.-J. *J. Chem. Phys.* **2000**, *113*, 9443–9455.

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

(11) Kendall, R. A.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.

(12) Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1996**, *104*, 6286.

(13) Pedulla, J. M.; Vila, F.; Jordan, K. D. *J. Chem. Phys.* **1996**, *105*, 11091–11099.

(14) Saebø, S.; Tong, W.; Pulay, P. *J. Chem. Phys.* **1993**, *98*, 2170–2175.

(15) Schütz, M.; Rauhut, G.; Werner, H.-J. *J. Phys. Chem. A* **1998**, *102*, 5997–6003.

(16) Valdés, H.; Klusák, V.; Pitoňák, M.; Exner, O.; Starý, I.; Hobza, P.; Rulíšek, L. *J. Comput. Chem.* **2008**, *29*, 861–870.

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

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

(19) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264–6271.

(20) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554–8560.

(21) Xu, X.; Goddard III, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673–2677.

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

(23) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Phys. Chem. A* **2007**, *111*, 10439–10452.

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

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

(26) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 6624–6627.

(27) Xu, X.; Goddard III, W. A. *J. Phys. Chem. A* **2004**, *108*, 2305–2313.

(28) Santra, B.; Michaelides, A.; Scheffler, M. *J. Chem. Phys.* **2007**, *127*, 184104.

(29) Èerný, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2005**, *7*, 1624–1626.

(30) Csontos, J.; Palermo, N. Y.; Murphy, R. F.; Lovas, S. *J. Comput. Chem.* **2008**, *29*, 1344–1352.

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

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

(33) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.

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

(35) Zhao, Y.; Tischenko, O.; Truhlar, D. G. *J. Phys. Chem. B* **2005**, *109*, 19046–19051.

(36) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103.

(37) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.

(38) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126–13130.

(39) Valdés, H.; Spiwok, V.; Rezac, J.; Reha, D.; Albo-Riziq, A. G.; de Vries, M. S.; Hobza, P. *Chem. Eur. J.* **2008**, *14*, 4886–4898.

(40) Grimme, S.; Schwabe, T. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4398–4401.

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

(42) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.

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

(44) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.

(45) Grimme, S. *J. comput. Chem.* **2004**, *25*, 1463–1473.

(46) Antony, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5287–5293.

(47) Piacenza, M.; Grimme, S. *J. Am. Chem. Soc.* **2005**, *127*, 14841–14848.

(48) Piacenza, M.; Grimme, S. *ChemPhysChem* **2005**, *6*, 1554–1558.

(49) Barone, V.; Biczysko, M.; Pavone, M. *Chem. Phys.* **2008**, *346*, 247–256.

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

(51) Hariharan, P. C.; Pople, J. A. *Theor. Chem. Acc.* **1973**, *28*, 213–222.

(52) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; von Ragué Schleyer, P. *J. Comput. Chem.* **1983**, *4*, 294–301.

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

(54) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.;

Ishida, M.; Nakajima, T.; Honda, Y.; KitaoO.; 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 E.01*; Gaussian Inc.: Wallingford, CT, 2004.

(55) *NWChem, Version 5.1*; High Performance Computational Chemistry Group, Pacific Northwest National Laboratory: Richland, WA 99352, U.S.A., 2006.

(56) Neese, F. *ORCA – an ab initio, density functional and semiempirical program package, 2.6, Revision 35*; University of Bonn: 2007.

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

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

(59) Groll, E.; Leininger, T.; Manby, F. R.; Mitrushchenkov, A.; Werner, H.-J.; Stoll, H. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3353–3357.

(60) Hill, J. G.; Platts, J. A.; Werner, H.-J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4072–4078.

(61) Schütz, M.; Werner, H.-J. *Chem. Phys. Lett.* **2000**, *318*, 370–378.

(62) Schütz, M. *J. Chem. Phys.* **2000**, *113*, 9986–10001.

(63) Schütz, M.; Werner, H.-J. *J. Chem. Phys.* **2001**, *114*, 661–681.

(64) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. *MOLPRO, a package of ab initio programs, version 2006.1*. http://www.molpro.net (accessed Sep 2008).

(65) Halkier, A.; Klopper, W.; Helgaker, T.; Jørgensen, P.; Taylor, P. R. *J. Chem. Phys.* **1999**, *111*, 9157–9167.

(66) Jurečka, P.; Hobza, P. *Chem. Phys. Lett.* **2002**, *365*, 89–94.

(67) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200–10207.

(68) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104–112.

(69) Riley, K. E.; Hobza, P. *J. Phys. Chem. A* **2007**, *111*, 8257–8263.

(70) Benighaus, T.; DiStasio, R. A.; Lochan, R. C.; Chai, J.-D.; Head-Gordon, M. *J. Phys. Chem. A* **2008**, *112*, 2702–2712.

(71) Bludský, O.; Rubeš, M.; Solán, P.; Nachtigall, P. *J. Chem. Phys.* **2008**, *128*, 114102.

(72) Grimme, S.; Mück-Lichtenfeld, C.; Antony, J. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3327–3334.

CT800231F

JCTC Journal of Chemical Theory and Computation

# Massively Multicore Parallelization of Kohn−Sham Theory

Philip Brown,[†] Christopher Woods,[†] Simon McIntosh-Smith,[‡] and Frederick R. Manby*,[†]

*Centre for Computational Chemistry, School of Chemistry, University of Bristol, Bristol, BS8 1TS, United Kingdom, and ClearSpeed Technology plc, 3110 Great Western Court, Hunts Ground Road, Bristol, BS34 8HP, United Kingdom*

Received July 1, 2008

**Abstract:** A multicore parallelization of Kohn−Sham density functional theory is described, using an accelerator technology made by ClearSpeed Technology. Efficiently scaling parallelization over 2304 cores is achieved. To deliver this degree of parallelism, the Coulomb problem is reformulated to use Poisson density fitting with numerical quadrature of the required three-index integrals; extensive testing reveals negligible errors from the additional approximations.

## 1. Introduction

Recent advances in computing technology and algorithm design have allowed ab initio electronic structure theory methods to be applied to large biological molecules.[1,2] However to model these systems realistically, relevant free energy differences must be computed. To do this, one must perform dynamics calculations requiring many thousands of calculations. Electronic structure calculations are computationally demanding, so treatment of a full biological system is generally impractical. The quantum mechanical/molecular mechanical (QM/MM) method divides a system into a small QM region and a larger residue which is treated with classical techniques, significantly reducing the size of the QM calculations required.[3−5] Despite this, QM/MM dynamics[6,7] have been largely restricted to computationally inexpensive semiempirical QM techniques such as AM1,[8] PM3,[9] or tight binding.[10]

Density functional theory (DFT)[11−13] provides an excellent balance of accuracy and computational cost; however, current implementations are an order of magnitude too slow for QM/MM dynamics in enzymological problems: a 20 000 time-step dynamics calculation on a 50-atom QM region might take around 1 year using a conventional serial implementation. New accelerator technologies can provide significant performance gains compared to commodity central processing units (CPUs).

Several groups have investigated the use of graphics processing units (GPUs) to calculate two-electron repulsion integrals (ERIs),[14,15] one of the bottlenecks in many electronic structure methods. They report speedups of 8−15×[14] and 80−130×[15] for their ERI kernels. Yasuda has also implemented exchange-correlation quadrature using GPUs[16] and reports a speedup of 5−10× against a commodity CPU. However, both groups had to make significant efforts to minimize the errors caused by the lack of double-precision support on current GPUs. We feel that this would significantly complicate the programming effort for a full DFT implementation.

ClearSpeed Technology plc produces a mature, low-power accelerator, with full double-precision support. Each CSX600 chip has 96 single-instruction, multiple-data (SIMD) processing elements (PEs) with 6 KB cache each, providing roughly 33 billion floating point operations per second (FLOPS) of double-precision performance in a matrix multiplication.[17] The ClearSpeed e620 mounts two CSX600 chips with 1 GB of random access memory (RAM) on a PCI-Express card.[18] An e620 consumes roughly 33 W. The ClearSpeed-accelerated tera-scale system (CATS) allows twelve e620 boards housed in one rack-mounted server unit to be attached to a host, providing aggregate performance of ∼1 TFLOPS. We present a heterogeneous approach to accelerate DFT, combining ClearSpeed's low-power 64-bit accelerator technology in parallel with the host CPU.

* To whom correspondence should be addressed. E-mail: fred.manby@bristol.ac.uk.

[†] University of Bristol.

[‡] ClearSpeed Technology plc.

10.1021/ct800261j CCC: $40.75 
Published on Web 09/13/2008

The ClearSpeed architecture requires a very fine degree of parallelization. To use a CATS node efficiently, work must be divided over $12 \times 2 \times 96 = 2304$ processing elements. A significant effort has been made over the past years to parallelize DFT for the relatively coarse architecture of multicore workstations[19] and vector supercomputers.[22−25] However, good scaling has been achieved mainly over tens or hundreds of processors. Efforts have also focused on implementing parallel linear-scaling methods,[20,21] which are less important for the relatively small QM region in a QM/MM dynamics calculation. We therefore propose a different approach, which uses the Poisson density fitting method[26−28] to shift all of the bottlenecks into finely parallelizable numerical quadrature.

## 2. Theory

DFT has two main bottlenecks when applied to ∼50 atom systems: the evaluation of the Coulomb matrix,

$$J_{\alpha\beta} = \sum_{\gamma\delta} \gamma_{\gamma\delta}(\alpha\beta|\gamma\delta) \tag{1}$$

and the numerical quadrature to evaluate the exchange-correlation contribution to the Fock matrix

$$V^{\mathrm{xc}}_{\alpha\beta} = \int \mathrm{d}\vec{r}\; v^{\mathrm{xc}}(\vec{r})\chi_\alpha(\vec{r})\chi_\beta(\vec{r}) \approx \sum_\lambda w_\lambda v^{\mathrm{xc}}_\lambda \chi_{\alpha\lambda}\chi_{\beta\lambda} \tag{2}$$

Here and throughout, we use the notation $(\cdot|\cdot)$ to denote a 2-electron repulsion integral, so for example

$$(\alpha\beta|\gamma\delta) = \int \mathrm{d}\vec{r}_1 \int \mathrm{d}\vec{r}_2 \frac{\chi_\alpha(\vec{r}_1)\chi_\beta(\vec{r}_1)\chi_\gamma(\vec{r}_2)\chi_\delta(\vec{r}_2)}{r_{12}} \tag{3}$$

The numerical quadrature runs over points $\vec{r}_\lambda$ with weights $w_\lambda$, and $v^{\mathrm{xc}}_\lambda = v^{\mathrm{xc}}(\vec{r}_\lambda)$ and $\chi_{\alpha\lambda} = \chi_\alpha(\vec{r}_\lambda)$. For much larger systems, quadrature becomes less of a bottleneck, because screening rapidly renders this an $O(N)$ step. Diagonalization becomes a serious bottleneck (scaling as $O(N^3)$) but can be avoided (see, for example, ref 29). The Coulomb problem asymptotically scales as $O(N^2)$ if screening is used, but can be made linear-scaling through the fast multipole method.[30−32]

It is straightforward to parallelize numerical quadrature by distributing batches of integration points between processing elements. The Coulomb term is more problematic. Direct calculation of the Coulomb contribution requires four index electron repulsion integrals (ERIs), $(\alpha\beta|\gamma\delta)$, which for f shells would require a matrix of 10 000 numbers occupying 78 KB of memory. This is difficult to efficiently map to an architecture with only 6KB of local store per PE. Implementations on GPUs,[14,15] which face similar limitations, have used Rys quadrature[33] for higher angular momenta. We propose to avoid the calculation of these ERIs altogether, by a combination of density fitting and use of the Poisson equation.

**2.1. Density Fitting.** To avoid the need to calculate 4-index ERIs, we use the density fitting method, first proposed by Boys and Shavitt in 1959,[36] and extended to DFT by Baerends et al.[34] and Dunlap et al.[35] The conventional Kohn−Sham density

$$\rho(\vec{r}) = \sum_{\alpha\beta} \gamma_{\alpha\beta}\chi_\alpha(\vec{r})\chi_\beta(\vec{r}) \tag{4}$$

is approximated by an auxiliary basis, $\Xi_A$:

$$\tilde{\rho}(\vec{r}) = \sum_B d_B \Xi_B(\vec{r}) \tag{5}$$

Rewriting eq 1, the Coulomb contribution becomes

$$J_{\alpha\beta} = (\alpha\beta|\rho) \approx (\alpha\beta|\tilde{\rho}) = \sum_B d_B(\alpha\beta|B) \tag{6}$$

The fitting coefficients, $d_B$ are obtained by minimizing the Coulomb self-energy of the fitting residual

$$\Delta = \frac{1}{2}(\rho - \tilde{\rho}|\rho - \tilde{\rho}) \tag{7}$$

This leads to the linear equations

$$\sum_B J_{AB} d_B = c_A \tag{8}$$

where

$$c_A = \sum_{\gamma\delta} \gamma_{\gamma\delta}(A|\gamma\delta) \tag{9}$$

and $J_{AB} = (A|B)$. Solving the fitting equations, eq 6 can then be used to give the Coulomb contribution to the Fock matrix, $J_{\alpha\beta}$. This method uses only three-index integrals of the form $(A|\gamma\delta)$. However, analytic calculation of these three-index integrals for f shells still requires a matrix of 1000 numbers, occupying 8 KB, posing a problem for a highly parallel implementation. In principle, the Coulomb potential

$$V_A(\vec{r}) = \int \mathrm{d}\vec{r}_1 \frac{\Xi_A(\vec{r})}{r_{12}} \tag{10}$$

of each fitting function could be evaluated on a quadrature grid and a numerical integration performed. However, the Coulomb potential is long-ranged, and we found that grids optimized for exchange-correlation quadrature did not give acceptable accuracy. We therefore use the Poisson method to convert most of our Coulomb integrals to overlap integrals,[26−28] which we can then evaluate using conventional DFT quadrature.

**2.2. Density-Fitted Poisson Method.** Manby and Knowles noticed simplifications in density fitting if the density is fitted in so-called Poisson functions: these are obtained by applying the Poisson operator $\hat{P} = -(4\pi)^{-1}\nabla^2$ to Gaussian-type orbitals.[26] The density is expanded in these Poisson functions:

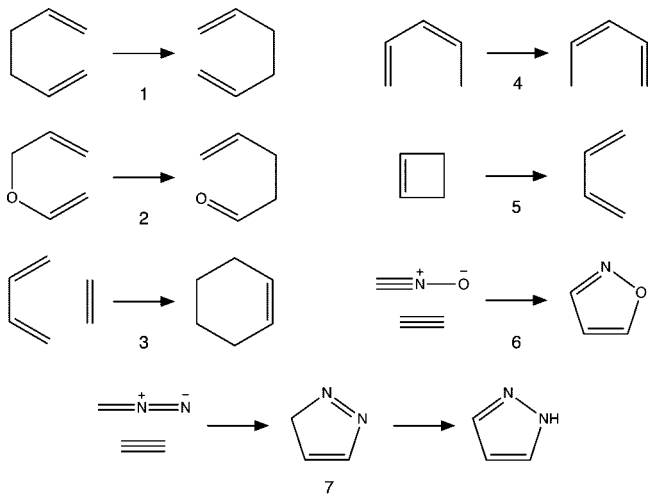
$$\tilde{\rho}(\vec{r}) = \sum_A d_A \hat{P}\Xi_A(\vec{r}) \tag{11}$$

and, using the integral identity

$$\Xi(\vec{r}_1) = \int \mathrm{d}\vec{r}_2 \frac{\hat{P}\Xi(\vec{r}_2)}{r_{12}} \tag{12}$$

the Coulomb matrix elements in the fitting basis simplify to short-ranged three-dimensional integrals, which differ from kinetic energy integrals only by a numerical factor:

$$J_{AB} = (A|B) = \int \mathrm{d}\vec{r}\; \Xi_A(\vec{r})\hat{P}\Xi_B(\vec{r}) \tag{13}$$



8 $HCN \rightarrow CNH$ 9 $HCOOC_2H_5 \rightarrow HCOOH + C_2H_4$

10 $3C_2H_2 \rightarrow C_6H_6$ 11 $HCHO \rightarrow CO + H_2$

***Figure 1.*** Selected reactions of molecules for which barriers have been computed.

Similarly, three-index Coulomb integrals, $(A|\gamma\delta)$, can be rewritten as simple overlap integrals:

$$(A|\gamma\delta) = \int \mathrm{d}\vec{r}_1 \int \mathrm{d}\vec{r}_2 \frac{\hat{P}_1 \Xi_A(\vec{r}_1)\chi_\gamma(\vec{r}_2)\chi_\delta(\vec{r}_2)}{r_{12}}$$
$$= \int \mathrm{d}\vec{r}\ \Xi_A(\vec{r})\chi_\gamma(\vec{r})\chi_\delta(\vec{r}) \qquad (14)$$

Further investigation revealed that the fitted density in eq 11 can have no total charge, dipole, or higher multipoles.[27] To alleviate this problem, a small number of ordinary basis functions are introduced, and these describe the charge and higher multipoles. The Poisson functions move the charge around and produce an accurate model density. Setting up a fitting basis with $m_c$ standard and $m_p$ Poisson functions, the fitted Coulomb matrix $J_{AB}$ can be broken down into three types of integrals: standard Coulomb integrals, standard overlaps, and scaled kinetic-energy-type integrals (eq 13). The three-index integrals, $(\alpha\beta|A)$, block into $m_c m(m+1)/2$ Coulomb integrals, and $m_p m(m+1)/2$ overlaps, where $m$ is the size of the atomic orbital basis.[27] The small number of standard Coulomb integrals and kinetic energy-like integrals are calculated explicitly, but the overlaps can calculated by quadrature.

This grid-based density-fitted Poisson method (GDFP) for the Coulomb problem has been implemented in serial within Molpro.[37] The energies were calculated on a test set of 21 reactions of small molecules containing first row elements (see Table V of ref 38) and some reactions of larger molecules, Figure 1.[39] Barrier heights for the larger reactions were also calculated. Calculations were performed with the BLYP functional, a cc-pVDZ orbital basis, a cc-pVTZ/jkfit fitting basis[40] for conventional density-fitted Kohn−Sham (DFKS), and the Poisson cc-pVTZ fitting set, described by Polly et al.[28] for density-fitted Poisson (DFP). Probability density plots of the errors of DFKS, DFP, and GDFP relative to standard KS for reaction energies and barriers (Figure 2) show that the grid-based method gives comparable accuracy to the standard DFP method and is in either case not worse than DFKS.

## 3. Implementation

The GDFP method and standard exchange-correlation quadrature have also been parallelized for the ClearSpeed accelerator technology. The ClearSpeed 3.0 software development kit (SDK) was used for the implementation. Accelerated code was written in C$^n$, a language that extends ANSI C[41] through the addition of keywords `mono` and `poly`. These specify scalar and parallel data types respectively, so `poly` data are distributed across all 96 PEs. A full set of standard C libraries as well as optimized mathematics libraries are available. The SDK also provides an implementation of BLAS double-precision matrix multiplication (DGEMM)[42] accessible from C$^n$, which gives easy access to the full 33 GFLOPS of the CSX600. However, for peak performance the matrices must have dimensions of around 1000 or more and use a blocked data format.

A fifty-atom molecule with a 6-31G* basis will have roughly 600 functions and require a quadrature grid of around 200 000 points. The atomic orbitals (AOs) evaluated on this grid therefore require roughly 900 MB of memory, assuming screening is not used. To avoid excessive communication between the accelerator and the host system, we choose to calculate and use the orbitals on the grid on the accelerator cards. We therefore pass only information about the grid, basis, Fock matrix, and density matrix between the accelerators and the host system. The grid is split equally between each CSX600 chip, and each batch, along with the entire basis, passed to the accelerators.

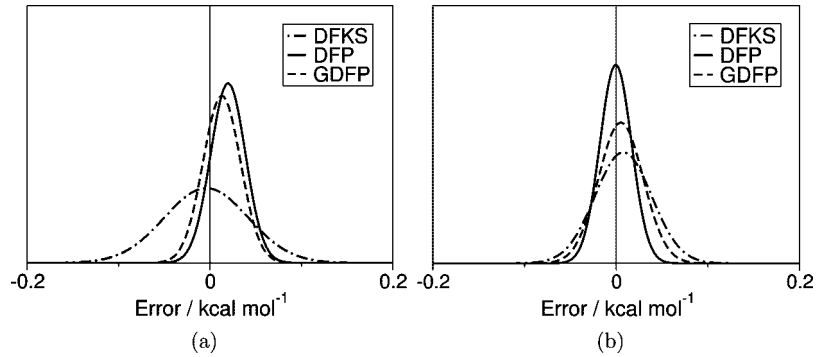



***Figure 2.*** Probability density plots of errors in (a) reaction energy and (b) barrier heights, relative to standard KS method.

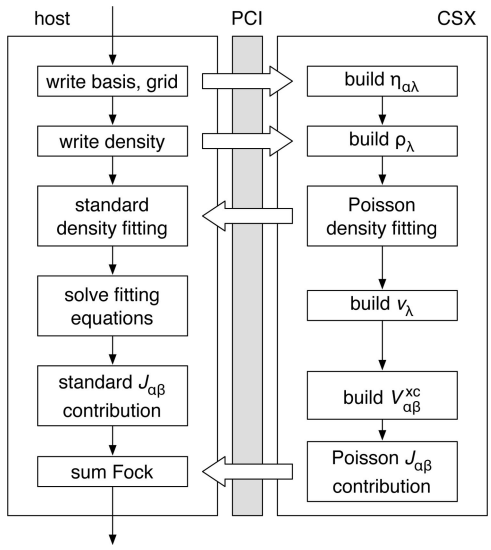



***Figure 3.*** Implementation of GDFP and exchange-correlation quadrature on a hybrid system.

The small local store available on each PE is a challenge for efficient evaluation of the orbitals on the grid,

$$\chi_{\alpha\lambda} = x_\lambda^l y_\lambda^m z_\lambda^n e^{-\alpha r_\lambda^2} \tag{15}$$

where $x_\lambda$, $y_\lambda$, and $z_\lambda$ are the distances from the grid point $\lambda$ to the center of the orbital on the respective axes, and $r_\lambda^2 = x_\lambda^2 + y_\lambda^2 + z_\lambda^2$. We choose to evaluate the entire basis, one basis group at a time, on eight grid points per PE, transferring the results to board RAM when each group has been completed. We then work through the grid in these tranches of $8 \times 96 = 768$ points. This allows us to make most efficient use of the processing power of the CSX600, while maximizing data bandwidth between PE local store and RAM. If enough RAM is available, we store the AOs on the grid to avoid recalculating them at every iteration. Otherwise we calculate the AOs in the largest block possible.

Further calculations on the card, such as building the density on the grid, are performed with all available AOs on the grid, to maximize the performance of the DGEMM calls. Figure 3 shows the implementation of one iteration the method when enough RAM is available to store the AOs on the grid. The accelerators are initialized, each given a portion of the grid and the complete basis set and the AOs on the grid calculated once, during the first iteration.

We treat the host and card environments as parallel pipelines. During every iteration of the calculation, the density matrix, $\gamma_{\alpha\beta}$, is passed to each card and a density on the grid calculated,

$$\rho_\lambda = \sum_{\alpha\beta} \gamma_{\alpha\beta}\chi_{\alpha\lambda}\chi_{\beta\lambda} \tag{16}$$

The vector

$$c_A = \sum_\lambda w_\lambda \rho_\lambda \Xi_{A\lambda}, \quad A \in \text{Poisson} \tag{17}$$

is calculated and passed back to the host. The host has calculated the conventional integrals,

$$c_A = \sum_{\alpha\beta} (\alpha\beta|A), \quad A \in \text{standard} \tag{18}$$

and can then solve

$$d_B = \sum_A [\mathbf{J}^{-1}]_{AB} c_A \tag{19}$$

for the fitting coefficients. The coefficients for the Poisson section of the fitting basis are transferred to the accelerators, and the contribution, to the Coulomb matrix,

$$J_{\alpha\beta} \leftarrow \sum_{B \in \text{Poisson}} d_B \sum_\lambda w_\lambda \chi_{\alpha\lambda}\chi_{\beta\lambda}\Xi_{B\lambda} \tag{20}$$

is built. The accelerators also calculate the exchange-correlation potential

$$v_\lambda^{xc} = f(\rho_\lambda) \tag{21}$$

analogues for density gradients, and all relevant contributions to the exchange-correlation matrix,

$$V_{\alpha\beta}^{xc} = \sum_\lambda w_\lambda v_\lambda^{xc} \chi_{\alpha\lambda}\chi_{\beta\lambda} \tag{22}$$

Meanwhile, the host calculates the conventional Gaussian contribution to the Coulomb matrix,

$$J_{\alpha\beta} \leftarrow \sum_{B \in \text{standard}} d_B(\alpha\beta|B) \tag{23}$$

All contributions to the Coulomb and exchange-correlation matrices are returned to the host and summed into the Fock matrix,

$$F_{\alpha\beta} \leftarrow \frac{1}{2} J_{\alpha\beta} + V_{\alpha\beta}^{xc} \tag{24}$$

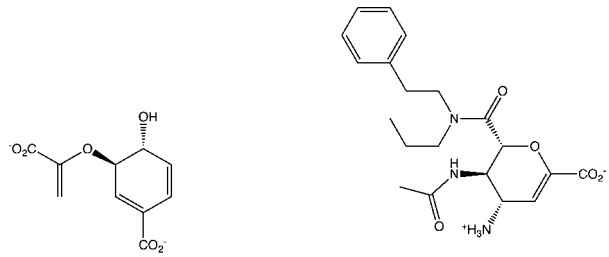
The calculation of the density on the grid and contributions to the Fock matrix are key steps as they scale $O(N^3)$ with molecular size. Fortunately they can be decomposed to a cubic-scaling matrix-multiplication, along with some matrix−vector operations with lower scaling. The calculation of the density on the grid (eq 16) breaks down the matrix multiplication,

$$\tilde{\chi}_{\alpha\lambda} = \sum_\beta \gamma_{\alpha\beta}\chi_{\beta\lambda} \tag{25}$$

and a series of dot products, one for each grid point,

$$\rho_\lambda = \sum_\alpha \chi_{\alpha\lambda}\tilde{\chi}_{\alpha\lambda} \tag{26}$$

Similar decompositions can be applied to eqs 20 and 22. Using the $C^n$ DGEMM implementation allows us to harness the full power of the ClearSpeed accelerators, and we routinely see 26 GFLOPS (80% of peak) per CSX600 in our DGEMMs and thus 624 GFLOPS aggregate on a single CATS node. All other steps, such as the calculation of the Coulomb fitting vector (eq 17) or exchange-correlation potentials (eq 21) are computationally trivial for our target molecules. Additionally, the calculation of the density on the grid can be shared between the Coulomb and exchange-correlation, enhancing our efficiency. Finally we are able to overlap the calculation on the accelerators with the computation of the conventional Coulomb integrals by the host.




(a) Chorismate (b) DPC

***Figure 4.*** Structures of test molecules.

***Table 1.*** Timings for the Accelerated GDFP Method versus Standard DFP and Density-Fitted KS Methods[a]

| | AO basis | fit basis | Wall times (s) | |
|---|---|---|---|---|
| | $N_{AO}$ | $N_{Fit}$ | Fock build | total |
| | | DPC | | |
| | 6-31G* | cc-pVDZ/jkfit | | |
| DFKS | 446 | 2581 | 2050 | 2383 |
| DFP | 446 | 3382 | 1380 | 1413 |
| grid DFP | 446 | 3382 | 80 | 112 |
| | cc-pVTZ | cc-pVTZ/jkfit | | |
| DFKS | 1218 | 3022 | 6730 | 7096 |
| DFP | 1218 | 3382 | 4043 | 4409 |
| grid DFP | 1218 | 3382 | 278 | 644 |
| | | Chorismate | | |
| | 6-31G* | cc-pVDZ/jkfit | | |
| DFKS | 240 | 1304 | 398 | 404 |
| DFP | 240 | 1680 | 283 | 289 |
| grid DFP | 240 | 1680 | 18 | 24 |
| | cc-pVTZ | cc-pVTZ/jkfit | | |
| DFKS | 592 | 1504 | 1281 | 1323 |
| DFP | 592 | 1680 | 1154 | 1194 |
| grid DFP | 592 | 1680 | 41 | 82 |
| | | $Ala_{12}$ helix | | |
| | 6-31G* | cc-pVDZ/jkfit | | |
| DFKS | 978 | 5696 | 9747 | 10489 |
| DFP | 978 | 7476 | 6577 | 7327 |
| grid DFP | 978 | 7476 | 446 | 867 |

[a] Calculations were performed on one core of a 2× Dual Core Opteron 2218 2.6 GHz server with 8 GB RAM attached to one ClearSpeed CATS (12 xe620 cards). All DFP calculations use the Poisson cc-pVTZ fitting set described by Polly et al.[28] and the BLYP exchange-correlation functional.

## 4. Results

A timing analysis of the method for chorismate ($C_{10}H_8O_6$, reactant substrate of chorismate mutase, Figure 4a[43]), a typical neuraminidase ligand, DPC ($C_{20}H_{27}O_5N_3$, Figure 4b[44]), and an alanine helix ($Ala_{12}$, $C_{37}H_{61}O_{13}N_{12}$), for various basis sets is shown in Table 1. Overall application speedup varies from 7× to 15× versus DFP and 11× to 19× compared to DFKS. If we consider only the time spent constructing the Fock matrix, we see between 15× and 30× acceleration compared to host-only DFP.

At the moment, we have not implemented any screening within the GDFP method, so the accelerators are performing considerably more work than the equivalent host implementation. We anticipate a further factor of at least 1.5× from the implementation of screening. Additionally, since the Coulomb energy has a significantly higher contribution to the total energy than the exchange-correlation, to maintain numerical stability we were forced to use the fine grid for all of the iterations of the GDFP calculation. Figure 5 shows the scaling of the method relative to number of ClearSpeed processing elements for DPC with a 6-31G* basis and the BLYP exchange-correlation functional. The method scales well over the 2304 processing elements of a CATS node.

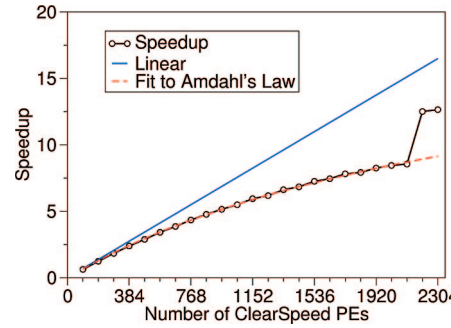



***Figure 5.*** Scaling of GDFP with a number of ClearSpeed processing elements for DPC with a 6-31G* AO basis, Poisson cc-pVTZ fitting set, and the BLYP exchange-correlation functional. The dashed line is a fit of the first 22 data points to Amdahl's law, corresponding to a 95% parallelization of the code. Calculations were performed on one core of a 2× Dual Core Opteron 2218 2.6 GHz with 8 GB RAM attached to one ClearSpeed CATS (12 × e620 cards).

***Table 2.*** Breakdown of the Calculation on DPC with a 6-31g* AO Basis, Poisson cc-pVTZ Fitting Set, and the BLYP Exchange-Correlation Functional[a]

| component | time % | time in DGEMM % |
|---|---|---|
| exchange-correlation | 58 | 17 |
| Coulomb | 27 | 9 |
| common | 15 | 11 |

[a] Calculation performed on one core of a 2× Dual Core Opteron 2218 2.6 GHz server with 8 GB RAM attached to one ClearSpeed CATS (12 × e620 cards). The "common" times are for building the orbitals and density on the grid, which are needed for evaluation of both exchange-correlation and Coulomb contributions.

Fitting to Amdahl's law reveals that we have parallelized ~95% of the calculation. We can clearly see that the remaining work on the host has become a bottleneck; the diagonalization of the Fock equations now takes ~30% of the total runtime. We can also observe a significant performance gain for the last two points, where enough memory becomes available to store the AOs on the grid, removing a significant portion of the work performed by the accelerators. It is also important to note that if we consider only time spent on the accelerator cards, we see perfect linear speedup, with a significant jump above linearity when storing the AOs becomes possible.

A breakdown of the time spent on the accelerator cards is given in Table 2 for DPC. Building the orbitals and density on the grid is common to both parts. It is worth noting that while DGEMMs account for 98% of the floating point operations, they only take ~37% of the time. This suggests significant opportunity for improving the efficiency of the implementation of other sections of code.

Due to the nature of the architecture, we pad the matrix dimensions to a multiple of 96. This leads to significant inefficiency for small molecules, especially with a small basis. Timings for ethane with one accelerator card are given

***Table 3.*** Timings for Ethane for the Accelerated GDFP Method versus Standard DFP and Density-Fitted KS Methods[a]

| | AO basis | fit basis | |
|---|---|---|---|
| | $N_{AO}$ | $N_{Fit}$ | wall time(s) |
| | 6-31G* | cc-pVDZ/jkfit | |
| DFKS | 40 | 278 | 8 |
| DFP | 40 | 362 | 8 |
| grid DFP | 40 | 362 | 5 |
| | cc-pVTZ | cc-pVTZ/jkfit | |
| DFKS | 144 | 338 | 35 |
| DFP | 144 | 362 | 30 |
| grid DFP | 144 | 362 | 8 |

[a] Calculations were performed on one core of a 2× Dual Core Opteron 265 1.8 GHz workstation with 4 GB RAM and one ClearSpeed xe620 card. All DFP calculations use the Poisson cc-pVTZ fitting set described by Polly et al.[28] and the BLYP exchange-correlation functional.

in Table 3 for two basis sets. With the large basis set, we observe a reasonable speedup of 4×.

## 5. Conclusions

We have implemented the GDFP method on ClearSpeed accelerators and demonstrated that an order of magnitude speedup is possible, with good scaling over thousands of PEs. The accelerator code shows perfect scaling over 2304 processing elements, while we see the expected behavior for the full application. There are however still many areas to improve. The introduction of screening should improve the efficiency of the method and ensure that it scales effectively to larger problem sizes. We anticipate improving the host/card load balancing at the same time, allowing the host to process batches of grid points. We also aim to implement gradients with respect to the nuclear positions, to allow the method to be used for dynamics calculations. The algorithm we have presented would map well onto GPUs, addressing some of the concerns expressed by Yasuda about the difficulty of fine-grained parallelization of the Coulomb problem.[16] Additionally, the current generation of GPUs have double-precision support,[45] greatly simplifying the implementation. Our algorithm is also suitable for implementation on standard shared memory parallel architectures, such as multicore x86, on which we expect that the algorithm would scale well.

**Acknowledgment.** The authors are indebted to several members of ClearSpeed Technology plc for input on algorithm design. They particularly thank Daniel Kidger and Thomas Bradley. The authors are grateful for funding from the School of Chemistry, University of Bristol, ClearSpeed Technology plc, and the Royal Society. Computer time from the Advanced Computing Research Centre at the University of Bristol is gratefully acknowledged.

### References

(1) Scuseria, G. *J. Phys. Chem. A.* **1999**, *103*, 4782.

(2) Gogenea, V.; Suárez, D.; van der Vaart, A.; Merz, K. W., Jr. *Curr. Opin. Struct. Biol.* **2001**, *11*, 217.

(3) Friesner, R. A.; Gullar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389.

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

(5) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700.

(6) Ridder, L.; Rietjens, I. M. C. M.; Vervoort, J.; Mulholland, A. J. *J. Am. Chem. Soc.* **2002**, *124*, 9926.

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

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

(9) Stewart, J. J. P. *J. Comput.-Aided. Mol. Des.* **1990**, *4*, 1.

(10) Frauenheim, T.; Seifert, G.; Elstner, M.; Hajnal, Z.; Jungnickel, G.; Porezag, D.; Suhai, S.; Scholz, R. *Phys. Stat. Sol. (B)* **2000**, *217*, 41.

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

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

(13) Kohn, W.; Becke, A. D.; Parr, R. G. *J. Phys. Chem.* **1996**, *100*, 12974.

(14) Yasuda, K. *J. Chem. Theory Comput.* **2008**, *4*, 1230.

(15) Ufimtsev, I. S.; Martínez, T. J. *J. Chem. Theory Comput.* **2008**, *4*, 222.

(16) Yasuda, K. *J. Chem. Theory Comput.*, in press.

(17) Clearspeed CSXL 3.0 User Guide, Section 5. http://support.clearspeed.com/documentation/software/release3 (accessed March 11, 2008).

(18) Advance e620 Product Brief. http://www.clearspeed.com (accessed March 11, 2008).

(19) Baker, J.; Füsti-Molnár, L.; Pulay, P. *J. Phys. Chem. A.* **2004**, *180*, 3040.

(20) Gan, C. K.; Challacombe, M. *J. Chem. Phys.* **2004**, *121*, 6608.

(21) Gan, C. K.; Challacombe, M. *J. Chem. Phys.* **2003**, *118*, 9128.

(22) Von Arnim, M.; Ahlrichs, R. *J. Comput. Chem.* **1998**, *19*, 1746.

(23) Sosa, C. P.; Ochterski, J.; Carpenter, J.; Frisch, M. J. *J. Comput. Chem.* **1998**, *19*, 1053.

(24) Furlani, T. F.; Kong, J.; Gill, P. M. W. *Comput. Phys. Commun.* **2000**, *128*, 170.

(25) Sosa, C. P.; Scalmani, G.; Gomperts, R.; Frisch, M. J. *Parallel Comput.* **2000**, *26*, 843.

(26) Manby, F. R.; Knowles, P. J. *Phys. Rev. Lett.* **2001**, *87*, 163001.

(27) Manby, F. R.; Knowles, P. J.; Lloyd, A. W. *J. Chem. Phys.* **2001**, *115*, 9144.

(28) Polly, R.; Werner, H.-J.; Manby, F. R.; Knowles, P. J. *Mol. Phys.* **2004**, *102*, 2311.

(29) Helgaker, T.; Larsen, H.; Olsen, J.; Jørgensen, P. *Chem. Phys. Lett.* **2000**, *327*, 397.

(30) Rokhlin, V. *J. Comput. Phys.* **1985**, *60*, 187.

(31) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chem. Phys. Lett.* **1994**, *230*, 8.

(32) Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Science* **1996**, *271*, 51.

(33) Dupuis, M.; Rys, J.; King, H. F. *J. Chem. Phys.* **1976**, *65*, 111.

(34) Baerends, E. J.; Ellis, D. E.; Ros, P. *Chem. Phys.* **1973**, *2*, 41.

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

(36) Boys, S. F.; Shavitt, I. *A Fundamental Calculation of the Energy Surface for the System of Three Hydrogen Atoms; Rep WIS-AF-13*; University of Wisconsin Naval Research Laboratory: Madison, WI, 1959.

(37) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nickla, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. *Molpro*, Version 2006.4; University College Cardiff Consultants Limited: Cardiff, U.K., 2006; http://www.molpro.net.

(38) Werner, H.-J.; Manby, F. R. *J. Chem. Phys.* **2006**, *124*, 054114.

(39) Nunn, J.; Harvey, J. N.; Manby, F. R. manuscript in preparation.

(40) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chim. Acta.* **1997**, *97*, 119.

(41) ISO/IEC 9899 - Programming languages - C. http://www.open-std.org/jtc1/sc22/wg14/www/standards.html#9899 (accessed May 22, 2008).

(42) Basic Linear Algebra Subroutines. http://www.netlib.org/blas/ (accessed May 22, 2008).

(43) Claeyssens, F.; Harvey, J. N.; Manby, F. R.; Mata, R. A.; Mulholland, A. J.; Ranaghan, K. E.; Schutz, M.; Thiel, S.; Thiel, W.; Werner, H. J. *Angew. Chem., Int. Ed.* **2006**, *45*, 6856.

(44) Birch, L.; Murray, C. W.; Hartshorn, M. J.; Tickle, I. J.; Verdonk, M. L. *J. Comput. Aid. Mol. Des.* **2002**, *16*, 855.

(45) nVidia Telsa C1060 Datasheet. http://www.nvidia.com/object/teslaroductiterature.html (accessed July 31, 2008).

CT800261J

JCTC Journal of Chemical Theory and Computation

# On the Balance of Simplification and Reality in Molecular Modeling of the Electron Density

Peter L. Warburton,*,† Jenna L. Wang,‡ and Paul G. Mezey†,§

*Scientific Modeling and Simulation Laboratory (SMSL), Department of Chemistry and Department of Physics and Physical Oceanography, Memorial University of Newfoundland, St. John's, Newfoundland A1B 3X7, Canada, Molecular Graphics and Modeling Laboratory, University of Kansas, Lawrence, Kansas 66045, and Institute for Advanced Study, Collegium Budapest, Szentháromság utca 2, 1014 Budapest, Hungary*

Received July 9, 2008

**Abstract:** Fused-sphere (van der Waals) surfaces and their variants such as solvent accessible surfaces and molecular surfaces are simple molecular models that are commonly used for many diverse purposes across a broad range of scientific disciplines due to their low computational resource demands. Fused-sphere models require atomic radii to be defined. Many different atomic radii have been proposed, with each set of radii being applicable to a relatively limited scope of molecular types or situations. The large number of differing radii sets actually serves to emphasize the simplicity of the model and its inability to accurately represent the reality of the molecule: its electron density. By measuring the similarity of fused-sphere, fuzzy fused-sphere, and calculated electron density representations of a set of small molecules via symmetric volume differences and the shape group method, it can be seen that fused-sphere models are very poor at representing the real electronic charge distribution of small molecules, especially where $\pi$ bond systems, lone pair electrons, and aromatic rings are involved. Larger molecules, conceivably, will be even more poorly represented. With advances in computational power and modeling techniques to arrive at high-quality calculated electron density representations for large molecules already in existence, abandoning the use of fused-sphere models should be considered for many applications.

## Introduction

Fused-sphere surfaces based upon the addition of rigid spherical representations of atoms defined by their van der Waals radii[1] are commonly used as simple models of molecular systems. Derivative models based on van der Waals surfaces,[2] where a spherical solvent molecule of a certain radius is rolled over the van der Waals representation, include the solvent-accessible surface and the molecular surface (solvent-excluded surface). In the solvent-accessible surface, the derivative surface is the envelope of points the center of the solvent molecule occupies as it is rolled over the van der Waals surface, while the molecular surface is the envelope of points of closest approach between the van der Waals surface and the solvent sphere surface as it is rolled along the original surface.

Applications using these fused-sphere models can fall into several categories. In one category property values such as lipophilicity[3] or electrostatic potential[4,5] are determined at points on the surface to help in defining force field[5−7] and solvation[4,7−11] models which are used to describe phenomenon ranging from molecular docking,[3] to molecular mechanics,[5−7] to calculated energetics,[8] to solute−solvent nuclear Overhauser effects.[9] In another category, fused-sphere models are used to determine molecular surface areas

* Corresponding author phone: (709)737-6939; fax: (709)737-3702; e-mail: peterw@mun.ca. Corresponding author address: Department of Chemistry, Memorial University of Newfoundland, St. John's, Newfoundland A1B 3X7, Canada.

† Memorial University of Newfoundland.

‡ University of Kansas.

§ Collegium Budapest.

10.1021/ct800268c CCC: $40.75 
Published on Web 09/10/2008

and volumes[2,12−25] which are then used to investigate many properties of proteins from folding[12] and packing density[13,14] to docking[15] and hydrophobicity[19] as well as molecular solid-state reactivity[16] and molecular connectivity indices for QSAR purposes.[25] In a third category, investigations of weak interactions in molecular systems often rely on the concept that a weak bond between atoms exists if the interatomic distance is shorter than the sum of the van der Waals radii.[26−28]

Determining the van der Waals radius for an atom is not a straightforward process, as the many proposed values of atomic radii[1,12−18,29−37] indicate. Many van der Waals radii values have been obtained from measurements of the nonbonding contact distances between atoms in crystal structures.[1,13−17,29−33] One drawback to this source of radii is that often hydrogen atoms are not clearly seen within the structure, and so they are either ignored, or the radii of the functional group including the hydrogen atoms are defined.[15,33] A second drawback is that a specific atom must either be represented by an average radius value, or by several different radii, depending on the environment it is found in.

Attempts at defining van der Waals radii based on electrostatic principles or SCF calculations[34−39] have also been made. However, these values often differ from each other because they are dependent on an appropriate cutoff condition. Various conditions proposed are cases where the radii in the model system result in appropriate binding[34] or repulsion energies,[35] or on numerical factors based on row constants, number of valence electrons and the Born exponent,[36] or on a fixed electron density value calculated from the ratio of the Dirac exchange constant and the Thomas-Fermi kinetic energy constant.[37] The use of a pseudopotential based calculations to derive adjustable atomic radii, mainly for transition metals, has also been considered,[38] as has the concept of bond valence.[39]

The defining of several different or adjustable radii[12−15,18,33,35,38] for the same atom (or functional group) arises from the concept that the atomic environments can differ due to bonding (e.g., C−H bond versus C−O bond), atom hybridization, or relative atomic position (e.g., the interior versus exterior of a protein). This use of several different radii then serves to reduce the issue of a single representative atomic radius being used to describe all like atoms but only does so to the extent that two or more representative radii are used instead. Also ignored in fused-sphere modeling is the concept of anisotropicity, where the atomic electron distribution is not spherical. This has been shown not only in a calculated electron density study[35] but also in crystallographic studies of specific atomic interactions, such as sulfur−sulfur[40] and chlorine−chlorine.[41] Also, in the case of weak bonding interactions it has been stated that the use of the sum of van der Waals radii as an indicator of interaction should be discarded, mainly due to the lack of precision in van der Waals radii, especially for metals.[42]

Derivative fused-sphere surfaces such as the solvent-accessible and solvent-excluded surfaces are sometimes seen to resolve another issue of fused-sphere surfaces, where the regions of overlap between two spheres do not provide a smooth transition from the surface of the first atom to the second, but rather the transition occurs through a nondifferentiable cusp point. This is more similar to what happens in real electron densities, which by their fuzzy natures have smooth transitions between atoms in bonding and overlap regions. However, the derivative fused-sphere models accomplish this not by modeling the molecule itself but rather a specific molecule-solvent interaction that will change with a change in solvent sphere choice. It has also been suggested that particular regions of the modeled molecule will interact differently with a chosen solvent than will other regions, and so variable solvent sphere radii need to be used in relation to a single molecule to give a more reasonable derivative fused-sphere model.[18]

Alternative simple models have been proposed to replace fused-sphere models. Most of these alternatives involve the addition of Gaussian-based functions instead of hard fused-spheres to model the atoms.[3,43−47] The reasons for proposing these models center on "scientific accuracy"[43] or convenience,[44] as Gaussian based functions are easily integrated or differentiated, and the products of Gaussians have useful coalescence properties.[45] As the addition of Gaussian functions does not lead to cusp regions in between atoms, such representations often match the solvent-excluded surfaces very well, especially if the Gaussians are chosen to have a radial distribution that matches closely a predetermined set of hard sphere van der Waals radii.

Pacios has also proposed and used a simplified representation of electron densities that could be used as an alternative to hard-sphere models[48−52] that are not based upon Gaussian functions but rather a parametrized radial distribution function arrived at from exponential function descriptions of the core and valence electrons.

While molecular shape-effects may become manifested in a whole variety of ways (for example, in solute−solvent interactions as well as in various statistical treatment of solutions), the evident fact remains: the shape of a molecule is the shape of the actual material making up the molecule-the atomic nuclei and the electronic density cloud. Since there is nothing else there, and the nuclei are buried within the electron density, the shape of a molecule is determined by the electron density cloud at the fuzzy, peripheral regions of molecules. Therefore, the shape of the electron density is the shape of the molecule. Evidently, any aspect of molecular shape must also ultimately depend on the electron density. This follows from the Hohenberg−Kohn theorem,[53] which states that the ground-state electron density fully determines the Hamiltonian and, hence, any other property of the molecule. Therefore, even for shape problems related to molecular interactions, it is natural to deal with electron density, since the very interactions must occur between and ultimately must depend upon the electron densities of the interacting molecules. For any interactions among a finite number of molecules, the molecular assembly can be regarded as a supermolecule, and all statements valid for the electron density shape characterization of individual molecules automatically also apply to the entire assembly, even if technically such computations might become more complex. Nevertheless, the principles are the same: electron density provides the ultimate shape representation.

The holographic electron density theorem provides an even stronger statement of the use of electron density as the means to define molecular shape: for real, boundaryless molecules in the electronic ground state, it has been proven that any positive volume part of the electron density of a molecule must also contain *all* of the information that is contained in the complete electron density of the molecule.[54] As an extension of the Hohenberg−Kohn theorem, this leads to the conclusion that any finite electron density piece (whether defined as a piece from a single molecule compared to the complete electron density of the molecule or defined as the nearly complete electron density of an individual molecule which is treated as a finite piece of the supermolecular electron density which describes a finite set of interacting molecules) must also contain all the information that would be found in complete ground-state electron density that determines all properties of the system.[55]

To show that fused-sphere models (either hard-sphere or Gaussian-based) are inadequate in modeling the fundamental nature of the molecule (its electron density), a set of 46 small molecules was tested. For each molecule, a hard-sphere representation, a Gaussian-based fuzzy fused-sphere representation, and several calculated electron density representations from basis sets of differing size were created. The similarities between these representations was measured by looking at the equivolume symmetric volume differences between the surface representations as well as the shape similarity of the distribution of electron density between the fuzzy fused sphere and the calculated representations. The results of these similarity comparisons show that, at best, hard-sphere models represent well the electron density of sigma-bonding-only molecules as calculated with a minimal basis set. Increases in basis set size, which are associated with a providing a better modeled representation of actual molecular electron density, also increase the differences seen between the calculated electron density representations and the fused-sphere models. A spherical representation of an atom is shown to be inadequate in modeling pi bonding systems and atoms with lone pairs. As small molecules are poorly represented, the inference that larger molecules will be even more poorly represented is a logical one. With advances in parallel computing and linear scaling electron density methods,[56−59] a movement away from fused-sphere models and toward calculated electron density-based models should be considered.

## Theory

Two means of measuring the similarity of molecular representative surfaces were used: symmetric volume differences and the shape group method.[60−66] In the symmetric volume difference method the similarity of two equal-volume enclosing surfaces expressed on the same molecular configuration $K$ ($G_K$ and $G'_K$) is measured by the volume enclosed by the intersection of surfaces $G_K$ and $G'_K$ ($V_{G\cap G'}$) subtracted from either volume $V_G$ or $V_{G'}$. Because the surfaces enclose equal volumes, the measure is symmetric: $V_G - V_{G\cap G'} = V_{G'} - V_{G\cap G'}$. Surface similarity is expressed by volume difference. Two surfaces with a small volume difference are more similar than two surfaces with a greater volume difference. The relative symmetric volume difference can be expressed as the ratio between the symmetric volume difference and the volume enclosed by one of the surfaces. Figure 1 shows a graphical example of the symmetric volume difference method. It should be noted in the figure that a two-dimensional slice of a three-dimensional model is shown. Therefore, while the areas enclosed by the surfaces in the figure may not be equal, over the entire three-dimensional model, the enclosed volumes are equal.

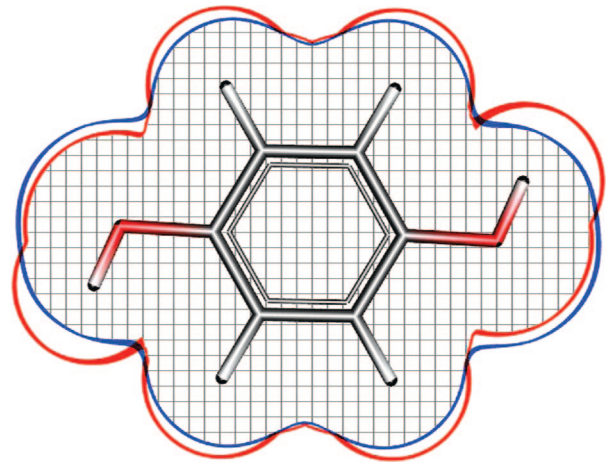


***Figure 1.*** An example of the symmetric volume difference measurement of similarity. The volume enclosed by the intersection of the red and blue surfaces (the patterned region) is subtracted from the volume enclosed by either the red or blue surface. Since the surfaces are chosen to enclose equal volumes, this defined volume difference is the same for either the red or the blue surface.

The shape group method provides consistent description of the shapes of molecules based on electron density representations. These descriptions can be compared to provide a single numerical measure of the similarity of two electron density representations.

Since electron density representations are fuzzy, the shape group method uses molecular isodensity contours (MIDCOs) to allow for the use of discrete mathematics. In three-dimensional space, a MIDCO $G(a)$ based upon molecular configuration K is defined for the fuzzy electron density representation $\rho(\mathbf{r})$ such that each point in the MIDCO has the same electron density value $a$: $G(a) = [\mathbf{r} \in \mathscr{R}^3: \rho(\mathbf{r}) = a]$, $a \in \mathscr{R}^3$. Because the boundary value $a$ can vary continuously throughout the space, a continuum of MIDCOs exist.

The shape characterization of a given MIDCO is performed based upon the local relative curvature properties of all the points that lie on the MIDCO surface.[66] Mathematically, the local curvature at the surface point of interest can be fully described by defining two orthogonal vectors that form the basis of the plane tangent to the surface point. These two vectors, combined with the electron density gradient vector, define a local coordinate frame for the three-dimensional space with the origin at the surface point of interest. For some specific choice of tangent plane vectors (the eigenvectors), the $2 \times 2$ Hessian matrix description of the tangent plane will have the eigenvalues $h_1$ and $h_2$ that describe the surface curvature relative to the tangent plane.

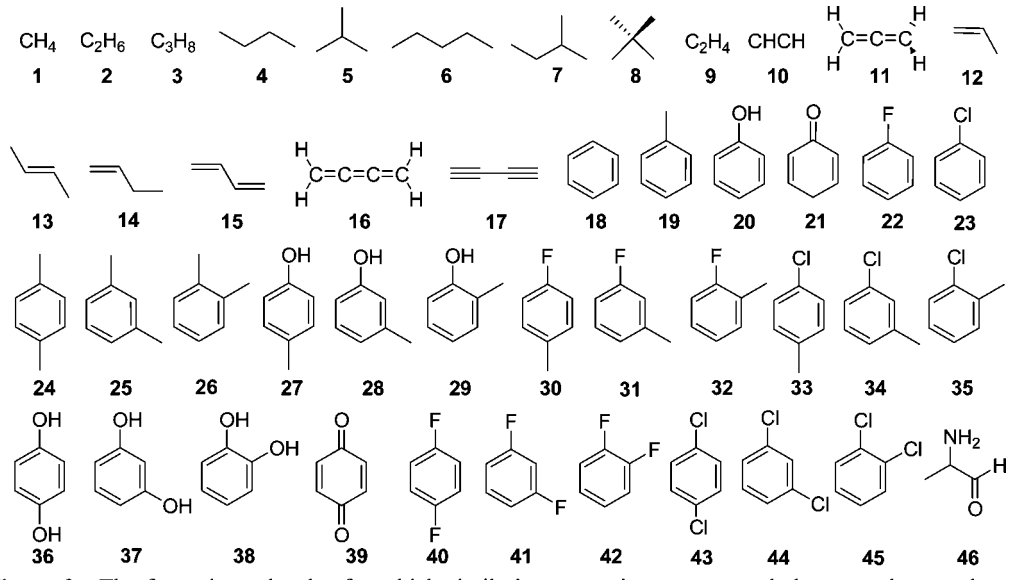



***Figure 2.*** The 46 molecules for which similarity comparisons were made between the van der Waals, fuzzy fused-sphere, and calculated electron density surfaces.

The local relative curvature is measured compared to a reference curvature defined by a parameter $b$ expressed as a sphere with radius $1/|b|$. A positive value of $b$ results in a sphere that curves toward the interior of the surface, a negative value of $b$ curves toward the exterior of the surface, and a zero value for $b$ represents the tangent plane to the point of interest. The local relative curvature is found by a direct comparison of the eigenvalues of the Hessian matrix to the value of $b$. The surface point is locally convex compared to the reference curvature if $h_1 \leq h_2 < b$ and can be said to belong to domain $D_2(b)$. The point is locally concave and belongs to domain $D_0(b)$ if $b \leq h_1 \leq h_2$. Finally, the point is locally saddle (domain $D_1(b)$) if $h_1 < b \leq h_2$.

When all of the surface points of a MIDCO are assigned to curvature domains, the shape of the MIDCO relative to the reference curvature can be described. If all locally convex points are truncated from the surface, the remaining MIDCO points define one or more distinct surface pieces. Each piece may be closed or have one or more holes. Through the use of homology groups[60] each piece of the truncated surface can be assigned three Betti numbers $B_p$ ($p = 0$, 1, or 2) based upon the incidence of oriented point, line, and area pieces to one another. If $B_0 = B_2$, then by the Poincaré index theorem for a two-dimensional surface the piece is a closed surface and has no holes. If $B_0 \neq B_2$, then the piece is open and is topologically equivalent to a sphere with a hole in it. All such open disconnected pieces in the truncated surface share a common hole, which serves to describe their disconnection from each other. Additional holes may also be present in a piece. The number of such holes is given by the first Betti number $B_1$. Therefore, the total number of holes for the piece is given by the first Betti number $B_1$ plus the common hole to all pieces, leading to $B_1+1$ holes. The shape of the MIDCO is described by the *shape ID vector*, which is the ordered list of $B_1+1$ values for all the pieces.

Since the ordered lists of the shape ID vectors can be unwieldy to compare, the description of the shape of the MIDCO can be further encoded by defining the *shape ID number* $c'(a,b)$ based upon a prime number encoding scheme of the ID vector. If the shape ID vector is ordered from the largest magnitude Betti number (B1) to the smallest (Bn), then $c'(a,b) = 2^{B1+1} \times 3^{B2+1} \times 5^{B3+1} \times ... \times P_n^{Bn+1}$.

The shape ID numbers are easily compared, but are only valid for comparison between MIDCOs of the same isodensity $a$ and reference curvature $b$. Since there are infinitely many choices for these variables, different values will result in different shape descriptions of the same molecule. Total shape characterization of a molecule is achieved through a collection of shape codes at various combinations of variable values. In the current implantation of the method, 41 MIDCOs $G(a)$ are chosen throughout the range $10^{-3}$ $e^-/bohr^3 \leq a \leq 10^{-1}$ $e^-/bohr^3$ and are analyzed at 21 reference curvature values ($10^{-5}$ bohr $\leq |b| \leq 1$ bohr and also $b = 0$ bohr). The complete shape of the molecule is described by an 861 ($41 \times 21$) member *(a,b)-map* matrix of shape codes.

Finally, the measurement of the similarity of electron density representations of molecules is given by the single-valued *shape similarity index* $S(X,Y)$. For molecules X and Y, the *shape equivalence* ($\Delta$) of the molecules at isodensity $a$ and curvature $b$ is one if they have the same shape code, and zero if they do not. The shape similarity index is the sum of the shape equivalence values for all matrix elements of the $(a,b)$-map divided by the total number of elements.

## Computational Approach

For each of the 46 molecules examined (Figure 2), a single geometry optimized at the HF/6−31G** level with *Gaussian03*[67] was used as a basis for all subsequent calculations. van der Waals surfaces were generated on a grid of 0.1 bohr utilizing the atomic radii of Gavezzotti.[16] Use of a grid allowed for direct comparison of the van der Waals surface to the grid-based electron density representations in the study.

Electron density representations were calculated on a 0.1 bohr grid utilizing several different basis sets with the Hartree−Fock methodology using an in-house program and

***Table 1.*** Similarity Comparisons for Sigma-Bonded Molecules **1**-**8**[a]

| | | STO-3G | | | | 3−21G | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $V_{VDW}$[b] | CED-VDW[c] | FFS-VDW[c] | FFS-CED[c] | shape[d] | CED-VDW[c] | FFS-VDW[c] | FFS-CED[c] | shape[d] |
| **1** | 189.6 | 2.93% | 2.46% | 1.61% | 0.85 | 3.48% | 2.58% | 1.81% | 0.90 |
| **2** | 294.2 | 2.63% | 2.05% | 1.16% | 0.91 | 3.45% | 2.63% | 1.21% | 0.89 |
| **3** | 404.1 | 2.60% | 2.02% | 1.08% | 0.95 | 3.40% | 2.78% | 1.10% | 0.86 |
| **4** | 519.1 | 2.69% | 2.11% | 1.10% | 0.95 | 3.59% | 2.98% | 1.21% | 0.87 |
| **5** | 514.2 | 2.67% | 2.16% | 1.04% | 0.94 | 3.52% | 2.96% | 1.11% | 0.86 |
| **6** | 626.1 | 2.67% | 2.14% | 1.02% | 0.91 | 3.55% | 3.02% | 1.14% | 0.85 |
| **7** | 624.2 | 2.61% | 2.14% | 0.98% | 0.94 | 3.46% | 2.96% | 1.05% | 0.85 |
| **8** | 624.8 | 2.85% | 2.42% | 1.13% | 0.92 | 3.77% | 3.17% | 1.32% | 0.93 |

| | | 6−31G** | | | | cc-pVDZ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $V_{VDW}$[b] | CED-VDW[c] | FFS-VDW[c] | FFS-CED[c] | shape[d] | CED-VDW[c] | FFS-VDW[c] | FFS-CED[c] | shape[d] |
| **1** | 189.6 | 3.16% | 2.51% | 0.69% | 0.87 | 3.75% | 2.97% | 1.02% | 0.69 |
| **2** | 294.2 | 2.77% | 2.57% | 0.84% | 0.87 | 3.43% | 3.53% | 1.41% | 0.62 |
| **3** | 404.1 | 2.77% | 2.75% | 0.89% | 0.86 | 3.43% | 3.78% | 1.47% | 0.59 |
| **4** | 519.1 | 2.89% | 2.95% | 0.97% | 0.89 | 3.62% | 4.06% | 1.48% | 0.65 |
| **5** | 514.2 | 2.88% | 2.97% | 0.84% | 0.89 | 3.57% | 3.96% | 1.34% | 0.61 |
| **6** | 626.1 | 2.88% | 3.01% | 1.02% | 0.86 | 3.60% | 4.10% | 1.55% | 0.62 |
| **7** | 624.2 | 2.82% | 2.97% | 0.88% | 0.86 | 3.51% | 3.95% | 1.33% | 0.63 |
| **8** | 624.8 | 3.18% | 3.24% | 0.88% | 0.87 | 3.81% | 4.09% | 1.31% | 0.66 |

[a] CED − calculated electron density, FFS − fuzzy fused-sphere, VDW − van der Waals. [b] $V_{VDW}$ − van der Waals surface volume in $bohr^3$/molecule. [c] Symmetric volume difference of the two specified surfaces expressed as a percentage of $V_{VDW}$. [d] Shape group method similarity between CED and FFS.

the full population analysis performed by *Gaussian03*. Additionally, electron density representations of the individual atoms in their ground states were calculated on a 0.08 bohr grid in the same manner. These atomic representations were then added together in a 0.1 bohr grid to give the fuzzy fused-sphere electron density representations. Use of a higher resolution grid for the atomic fragments served to reduce the error that would be caused by nuclear positions that did not sit exactly on any of the face-centered cubic positions of the grid cells. In all cases, the placement and orientation of the molecule within the grid was defined based on the standard orientation of the *Gaussian03* output, with the origin of the space placed exactly in the center of a grid cell.

The volumes enclosed by the van der Waals surfaces were calculated numerically. Electron density surfaces for both the calculated and fuzzy fused- sphere representations enclosing the same volume as the van der Waals surface were then found through an iterative process of adjusting the isosurface value and numerically calculating the enclosed volume. Because all molecular representations were created in the same grid space, the symmetric volume difference between surfaces was also calculated numerically.

Shape similarity comparisons between fuzzy fused-sphere and calculated electron density representations of molecules were carried out with a suite of in-house programs.

## Results and Discussion

The molecules studied were subdivided into groups based upon their bonding. Alkanes **1**-**8** display only sigma bonding, while molecules **9**-**17** and **46** show both pi and sigma bonding. The remaining molecules contain an aromatic ring, with the exception of **21** and **39**, though these molecules are included in the aromatic group because of their ring structure and high degree of conjugation.

Table 1 shows the comparison of the van der Waals (VDW), fuzzy fused-sphere (FFS), and calculated electron density (CED) surfaces for the alkane sigma-bonding group of molecules at four different basis set levels using relative symmetric volume differences. Additionally, the shape group method similarity values between the fuzzy fused-sphere and calculated electron density representations of the molecules are given.

In terms of calculated electron densities, those calculated at the HF/STO-3G level are considered to be the least representative of realistic electron densities, both because the Hartree−Fock method does not account for electron correlation, and because the STO-3G basis set is small. The minimal nature of the basis set does not allow for robust representation of molecular orbitals, to the point where the HF/STO-3G calculated electron density can be thought of as fused-atom representations of the molecules with slight distortions seen in the bonding regions. This notion is borne out in the data of Table 1. The fuzzy fused-sphere representations of the molecules have smaller relative symmetric volume difference values compared to the van der Waals surface (FFS-VDW) than do the calculated electron density representations (CED-VDW), indicating the fuzzy fused-sphere representations are more similar to the van der Waals representations. However, the fuzzy fused-sphere and calculated electron density representations are very similar to each other (FFS-CED) because the addition of fuzzy spheres will better model the off-bond portion of the sigma-bonding regions of the molecules compared to discrete fused-sphere representations. This is not surprising, as a disadvantage of van der Waals surfaces is the poor representation of the electron density in the bonding regions where the spheres overlap. Molecular surface representations counteract this disadvantage somewhat and should end up being more similar to the fuzzy fused-sphere representations than to the

***Table 2.*** Similarity Comparisons for Molecules with Sigma and Pi Bonding[a]

| | | STO-3G | | | | 3−21G | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $V_{VDW}$[b] | CED-VDW[c] | FFS-VDW[c] | FFS-CED[c] | shape[d] | CED-VDW[c] | FFS-VDW[c] | FFS-CED[c] | shape[d] |
| **9** | 263.3 | 3.59% | 4.55% | 1.60% | 0.65 | 3.58% | 3.58% | 4.33% | 0.82 |
| **10** | 237.6 | 4.34% | 5.71% | 2.01% | 0.85 | 3.83% | 4.91% | 6.53% | 0.91 |
| **11** | 341.5 | 3.71% | 4.92% | 1.67% | 0.61 | 3.46% | 3.85% | 5.03% | 0.85 |
| **12** | 373.4 | 3.32% | 3.71% | 1.37% | 0.78 | 3.51% | 3.31% | 3.15% | 0.77 |
| **13** | 482.9 | 3.05% | 3.02% | 1.24% | 0.90 | 3.39% | 3.03% | 2.41% | 0.80 |
| **14** | 482.2 | 3.12% | 3.16% | 1.21% | 0.83 | 3.45% | 3.29% | 2.58% | 0.83 |
| **15** | 449.7 | 3.60% | 4.59% | 1.51% | 0.80 | 3.51% | 3.70% | 4.37% | 0.87 |
| **16** | 418.1 | 3.73% | 5.02% | 1.62% | 0.66 | 3.15% | 3.92% | 5.03% | 0.83 |
| **17** | 392.7 | 3.55% | 4.40% | 1.83% | 0.58 | 3.37% | 3.73% | 4.87% | 0.81 |
| **46** | 486.2 | 3.04% | 2.80% | 1.45% | 0.79 | 4.62% | 3.19% | 3.09% | 0.77 |

| | | 6−31G** | | | | cc-pVDZ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $V_{VDW}$[b] | CED-VDW[c] | FFS-VDW[c] | FFS-CED[c] | shape[d] | CED-VDW[c] | FFS-VDW[c] | FFS-CED[c] | shape[d] |
| **9** | 263.3 | 3.00% | 4.07% | 3.89% | 0.74 | 4.43% | 3.63% | 4.84% | 0.71 |
| **10** | 237.6 | 3.69% | 5.49% | 6.28% | 0.84 | 5.70% | 3.88% | 7.73% | 0.86 |
| **11** | 341.5 | 3.02% | 4.41% | 4.57% | 0.76 | 4.43% | 3.32% | 5.58% | 0.74 |
| **12** | 373.4 | 2.97% | 3.54% | 2.75% | 0.77 | 4.15% | 3.80% | 3.61% | 0.61 |
| **13** | 482.9 | 2.88% | 3.16% | 1.89% | 0.64 | 3.90% | 3.60% | 2.52% | 0.58 |
| **14** | 482.2 | 2.88% | 3.43% | 2.19% | 0.74 | 3.96% | 3.93% | 2.89% | 0.61 |
| **15** | 449.7 | 2.99% | 4.12% | 3.88% | 0.75 | 4.39% | 3.80% | 4.84% | 0.71 |
| **16** | 418.1 | 2.78% | 4.61% | 4.66% | 0.62 | 4.14% | 3.19% | 5.68% | 0.68 |
| **17** | 392.7 | 3.18% | 4.18% | 4.87% | 0.65 | 4.69% | 3.00% | 5.75% | 0.82 |
| **46** | 486.2 | 5.36% | 3.32% | 3.89% | 0.64 | 6.70% | 4.13% | 4.43% | 0.61 |

[a] CED − calculated electron density, FFS − fuzzy fused-sphere, VDW − van der Waals. [b] $V_{VDW}$ − van der Waals surface volume in $bohr^3$/molecule. [c] Symmetric volume difference of the two specified surfaces expressed as a percentage of $V_{VDW}$. [d] Shape group method similarity between CED and FFS.

van der Waals surfaces in most cases. However, in cases where this increased similarity would be seen it needs to be remembered that this comes about as a result of a complementary solvent radius choice and not because the molecular surface model is any more adept at realistically representing electron density in bonding regions of the solvated molecule.

The shape group method similarity measure between the fuzzy fused-sphere and calculated electron density representations shows how the two representations would differ if the isosurface bounds were changed, like in a case where different van der Waals radii are used, leading to changes in volume. Effectively, as long as the calculated electron density can be thought of as the slightly distorted addition of atomic representations, the shape similarity between it and the fuzzy fused-sphere representation should be high, while low shape similarity values indicate large distortions of the calculated electron density from the surface created by adding atomic representations. For the HF/STO-3G sigma-bonded molecules, the shape similarity values are high, indicating the calculated electron density is much like a slightly distorted fuzzy fused-sphere representation.

As basis set size is increased, the calculated electron density is based on a larger number of molecular orbitals, and so the bonding and more electron-diffuse regions of molecules are better defined. Because of this a slight increase is seen in the CED-VDW and FFS-VDW values for the basis sets 3−21G, 6−31G**, and cc-pVDZ as compared to the STO-3G values for the same molecules. However, the FFS-CED values are still quite small, indicating high similarity between the fuzzy fused-sphere and calculated electron density representations. Overall, because sigma bonding can be described as cylindrical in nature, the overlap of two fuzzy spheres will model a $\sigma$ bond well both in terms of the cylindrical electron distribution within the bond as well as in terms of the decrease in electron density as the perpendicular distance from the axis of the cylinder is increased.

The shape similarity numbers tell much the same story as the basis set size is increased from STO-3G to 3−21G to 6−31G**. Generally, small decreases in the similarity for all the sigma-bonded molecules are seen as the basis set size is increased, indicating that the calculated electron density is quite similar to added atomic representations throughout a large range of electron density isosurface values. Use of different sets of van der Waals radii should therefore not have a large impact on the comparability of the representations. This cannot be said, however, for the molecular electron densities calculated at the highest basis set level (cc-pVDZ). The small shape similarity values indicate that fuzzy fused-sphere (and by extension, discrete fused-sphere) surfaces are not adept at modeling the electron density through a range of isosurface values. Since the larger basis set should provide a more realistic representation of the electron density than a smaller basis set, this indicates fused-sphere models are not effectively modeling "real" sigma-bonding systems.

Table 2 shows data comparable to that of Table 1 for the set of molecules (**9**-**17**, **46**) with both sigma and pi bonding.

The data in Table 2 show larger relative symmetric volume difference values than seen in Table 1 for both the calculated electron density and fuzzy fused-sphere surfaces compared to the van der Waals surfaces as well as compared to each other. More specifically, as the basis set size is increased, there is a marked increase in the FFS-CED symmetric volume difference values, indicating the calculated electron

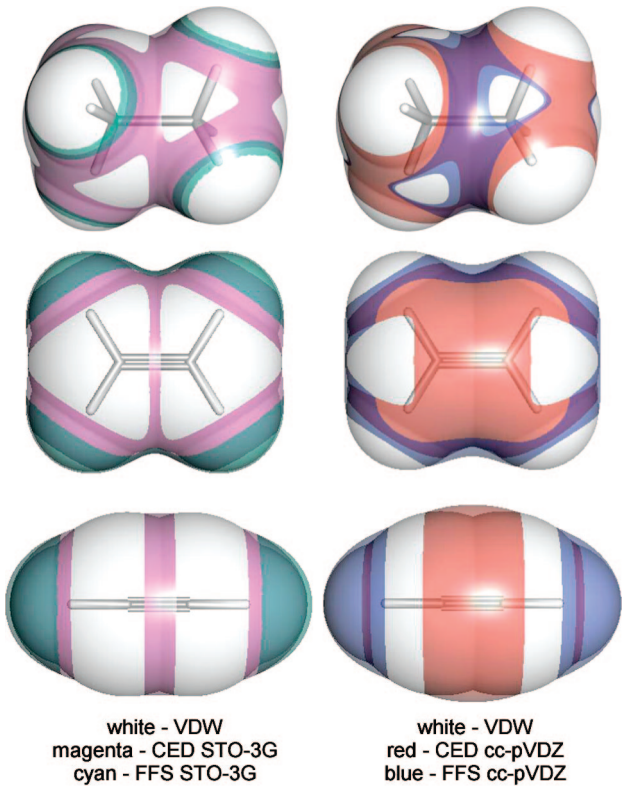



***Figure 3.*** Equivolume van der Waals (VDW), fuzzy fused-sphere (FFS), and calculated electron density (CED) surfaces for ethane (**2**), ethene (**9**), and ethyne (**10**) at the HF/STO-3G and HF/cc-pVDZ levels of theory.

density can no longer be seen as slightly distorted added atomic representations. Pi bonding in the systems is the explanation for this observation. Pi bonding, unlike sigma bonding, is not cylindrical in nature, and so the overlap of spheres cannot represent it as well. This is confirmed as the smallest relative symmetric volume difference values in the table are seen for molecules **12**-**14**, where there are significant sigma-bonded regions of the molecules as well as pi-bonded regions. The values indicate these molecules fall in between those of Table 1 and the more significantly pi-bonded molecules of Table 2, showing that the sigma-bonded regions are somewhat well represented by the fused-sphere models but the pi bonding regions are not.

The notion that fused-sphere models are poor at representing pi bonding is further confirmed by the shape similarity values. In general, they are lower than those seen in Table 1, but more specifically, they are in many cases much smaller for molecules with double bonds. Ethene (**9**) is seen to have shape similarity numbers much smaller than those for ethane (**2**) for the basis sets STO-3G, 3−21G, and 6−31G**. The pi bonding of the double bond is directed in a plane that includes the axis of the $\sigma$ bond and is characterized by fattening of the electron density between the atoms involved in the bonding. Since the cylindrical representation of bonding in fused-sphere models cannot include this directionality, this leads to lowered shape similarity to the calculated electron density. However, the ethyne (**10**) shape similarity numbers increase from those for ethene because the two perpendicular pi bonds do create a cylindrical electron density distribution. This, however, is a fortuitous

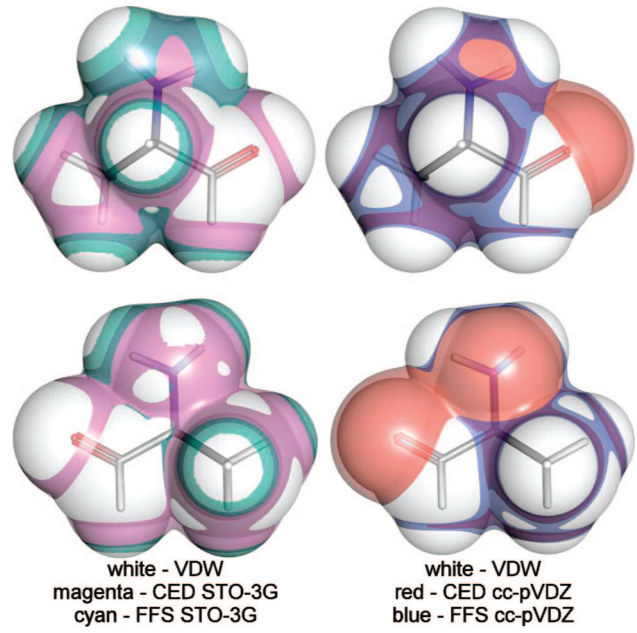



***Figure 4.*** Front and back views of equivolume van der Waals (VDW), fuzzy fused-sphere (FFS), and calculated electron density (CED) surfaces for the alanyl group (**46**) at the HF/STO-3G and HF/cc-pVDZ levels of theory.

occurrence in the fused-sphere models rather than caused by adequate modeling of the electron density, as can be seen in Figure 3.

In Figure 3 equivolume van der Waals, fuzzy fused-sphere, and calculated electron density surfaces are shown for ethane, ethene, and ethyne at the HF/STO-3G and HF/cc-pVDZ levels. For ethane, the main difference between the van der Waals surface and the STO-3G calculated electron density surface is seen in the bonding regions of the molecule, as is to be expected. The FFS-VDW differences are most apparent in the carbon−hydrogen bonding regions, but for the most part, the two surfaces are not very different, as is seen in the symmetric volume difference data. In the HF/cc-pVDZ ethane representations, the calculated electron density shows many of the same differences from the van der Waals surface as for the smaller basis set, but the FFS-VDW comparison is much more complex.

The ethene molecule surfaces are most interesting. The STO-3G basis set, as a minimal basis, is not as effective at modeling pi bonding because of the small number of molecular orbitals, and so the calculated electron density does not show the characteristic pi-bond fattening except in the midpoint of the carbon−carbon bonding region. In the cc-pVDZ CED surface, the pi-bond fattening is quite evident, and so neither the VDW or FFS surface represents well the calculated electron density. The ethyne molecule surfaces also show these same features, but again, because of the cylindrical overlap of the two perpendicular pi bonds in the molecule, the VDW and FFS models more closely match the calculated electron density surfaces.

A molecule of special interest in the study is the hydrogen terminated representation of the alanyl group (**46**), which shows the basic features of amide linked amino acids as they are found in proteins, the most notable of which are the amino group and the carbonyl group. Table 2 contains the

***Table 3.*** Similarity Comparisons for Selected Aromatic Group Molecules[a]

| | | STO-3G | | | | 3−21G | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $V_{VDW}$[b] | CED-VDW[c] | FFS-VDW[c] | FFS-CED[c] | shape[d] | CED-VDW[c] | FFS-VDW[c] | FFS-CED[c] | shape[d] |
| **18** | 557.8 | 3.21% | 4.22% | 1.34% | 0.90 | 2.86% | 3.32% | 4.48% | 0.87 |
| **19** | 668.5 | 3.10% | 3.94% | 1.31% | 0.92 | 2.95% | 3.20% | 3.79% | 0.85 |
| **20** | 599.6 | 3.22% | 4.10% | 1.61% | 0.82 | 3.61% | 3.30% | 4.80% | 0.81 |
| **21** | 602.7 | 3.05% | 3.68% | 1.16% | 0.87 | 3.44% | 3.09% | 3.92% | 0.87 |
| **27** | 710.2 | 3.14% | 3.81% | 1.51% | 0.91 | 3.59% | 3.21% | 4.13% | 0.84 |
| **28** | 711.5 | 3.14% | 3.83% | 1.52% | 0.90 | 3.60% | 3.22% | 4.19% | 0.83 |
| **29** | 710.1 | 3.14% | 3.82% | 1.52% | 0.91 | 3.58% | 3.19% | 4.16% | 0.87 |
| **39** | 610.2 | 3.11% | 3.57% | 1.07% | 0.86 | 3.76% | 3.02% | 3.92% | 0.85 |

| | | 6−31G** | | | | cc-pVDZ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $V_{VDW}$[b] | CED-VDW[c] | FFS-VDW[c] | FFS-CED[c] | shape[d] | CED-VDW[c] | FFS-VDW[c] | FFS-CED[c] | shape[d] |
| **18** | 557.8 | 2.25% | 3.77% | 3.96% | 0.88 | 3.71% | 3.40% | 4.98% | 0.68 |
| **19** | 668.5 | 2.41% | 3.57% | 3.43% | 0.83 | 3.77% | 3.50% | 4.45% | 0.63 |
| **20** | 599.6 | 3.62% | 3.93% | 4.52% | 0.82 | 5.07% | 3.85% | 5.34% | 0.67 |
| **21** | 602.7 | 3.90% | 3.46% | 4.02% | 0.81 | 5.19% | 3.65% | 4.86% | 0.68 |
| **27** | 710.2 | 3.52% | 3.72% | 3.99% | 0.85 | 4.89% | 3.81% | 4.79% | 0.65 |
| **28** | 711.5 | 3.52% | 3.72% | 4.03% | 0.86 | 4.93% | 3.82% | 4.90% | 0.71 |
| **29** | 710.1 | 3.50% | 3.72% | 3.94% | 0.87 | 4.90% | 3.79% | 4.75% | 0.67 |
| **39** | 610.2 | 5.28% | 3.70% | 4.47% | 0.79 | 6.52% | 3.72% | 5.11% | 0.79 |

[a] CED − calculated electron density, FFS − fuzzy fused-sphere, VDW − van der Waals. [b] $V_{VDW}$ − van der Waals surface volume in $bohr^3$/molecule. [c] Symmetric volume difference of the two specified surfaces expressed as a percentage of $V_{VDW}$. [d] Shape group method similarity between CED and FFS.

similarity data for the surfaces of this alanyl group, while Figure 4 shows front and back views of the van der Waals, the HF/STO-3G, and the HF/cc-pVDZ fuzzy fused-sphere and calculated electron density surfaces.

When considered in the group of molecules with both double bonds and significant sigma-bonding-only groups (**12**-**14**), the alanyl group FFS-VDW symmetric volume difference values occur at the low end of the range for STO-3G and then move toward the high end of the range as the basis set size is increased. In Figure 4 it can be seen that the main sources for this increase are the bonding regions of the hydrogen atoms, and the fuzzy-fused sphere overlaps between the groups that are bonded to the central carbon atom.

More importantly, the FFS-CED and the CED-VDW symmetric volume difference comparisons start in the same range of values as molecules **12**-**14** with the minimal basis set, but quickly fall outside of the range, becoming much larger (indicating lesser similarity) when the basis set size is increased. The shape similarity values confirm this trend, as lower similarity is seen as the basis set becomes larger, and the electron density is more realistically modeled.

Figure 4 shows that the STO-3G calculated electron density surface differs most in the bonding and group overlap regions, as has been seen before in ethane. For the cc-pVDZ CED, though, the largest differences are seen around the oxygen atom of the carbonyl group and the nitrogen atom of the amine group. If the electron density is considered in very simple terms these regions of the molecule include sigma-bonding, pi-bonding, and nonbonding electrons. With the double-bonded oxygen atom and its nonbonded electrons, the electron density tends to balloon outside of the fused-sphere representations anisotropically in various directions, while for the nitrogen atom this ballooning tends to occur in the region of the atom where the nonbonding electrons would be found. A sphere of fixed radius, regardless of how the radius is chosen, will not accurately model all three electron types at the same time.

The implication for fused-sphere modeling of polypeptides and proteins is clear. While the electron density sigma-bonding-only side chains of certain amino acids like glycine, alanine, valine, leucine, and isoleucine can somewhat be adequately modeled with fused-sphere representations, the backbone of the protein, with its carbonyl and amine groups, cannot be adequately modeled with fused-sphere representations. Additionally, the side chains of amino acids such as aspartic acid, glutamic acid, lysine, arginine, asparagine, and glutamine most likely face the same fused-sphere modeling difficulties as the backbone groups. Such a notion is supported by a study where parametrization of atomic radii to give agreement between van der Waals and solvent accessible surface-based Poisson−Boltzmann electrostatic free energy calculations show that for a single amino acid the atomic radii need to differ by 2−5% between the van der Waals and solvent accessible surface calculations, but for an approximately 200-residue protein this difference in atomic radii required to achieve agreement increases to over 20%.[8]

Table 3 gives the similarity comparisons for selected aromatic group (**18**-**45**) molecules. In comparison to ethene (**10**), the benzene molecule (**18**) CED-VDW symmetric volume difference values are markedly lower, indicating higher similarity. This higher similarity is also seen in the shape similarity values, with the exception for the HF/cc-pVDZ case, where little change is seen in the relatively poor similarity value. The high level of symmetry leads to this higher similarity, with the ring forming a toroidal distribution of electron density, which acts much like the cylindrical distribution of the electron density in both sigma and triple bonds. Figure 5 shows the surfaces at the HF/STO-3G and HF/cc-pVDZ levels for benzene. As seen before in ethene,

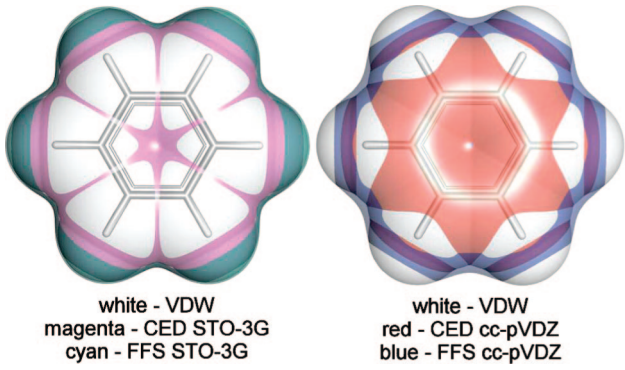



***Figure 5.*** Equivolume van der Waals (VDW), fuzzy fused-sphere (FFS), and calculated electron density (CED) surfaces for benzene (**18**) at the HF/STO-3G and HF/cc-pVDZ levels of theory.

the minimal basis set does not model well the delocalization of the electron density above and below the plane of the ring, leading to the calculated electron density being much like a distorted fused-sphere representation where the most notable differences to the VDW surface are in the midpoints of the bonds. The larger basis set more accurately models this delocalization.

Molecules **21** and **39** in the table are highly conjugated ring compounds that are not aromatic. In terms of the symmetric volume difference values as the basis set size is increased, the trends in the changes in the values as compared to benzene are much like those of the alanyl molecule relative to molecules **12**-**14**, in that sigma-, pi-, and nonbonding electrons are being represented by a single fused-sphere radius, and so as the electron density is better modeled with larger basis sets, the differences become more marked.

## Conclusion

Through the use of symmetric volume difference and the shape group method it has been shown that fused-sphere models and fuzzy fused-sphere models show significant differences to calculated electron densities of small molecules, even for a minimal basis set, where it would be expected the models should show reasonable similarity. Use of more complete basis sets only serve to highlight the increasing difference between fused-sphere models and calculated electron densities, especially as they relate to pi- and nonbonding regions of the molecules. Larger molecules, it can be concluded, will likely be even more poorly represented by fused-sphere or fuzzy fused-sphere models. As such, with advances in computing power, algorithm parallelization and linear scaling electron density methods, the use of calculated electron density models is recommended over fused-sphere models. In cases where fused-sphere based models are the only currently existing models in use for the calculation of a specific property, these results would also indicate that development and exploration of specific property models that are based on calculated electron density representations should be vigorously pursued.

**Acknowledgment.** We thank Drs. Thomas E. Exner and Jürgen Brickmann for the use of the MOLCAD II module[67−70] of the SYBYL molecular modeling package[71] in visualizing our results.

**Supporting Information Available:** Complete table of the similarity comparison data, including fuzzy fused-sphere and calculated electron density isocontour bound values, for all 46 molecules at six basis set levels. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Pauling, L. van der Waals and Nonbonded Radii of Atoms. In *The Nature of the Chemical Bond*, 3rd ed.; Cornell University Press: Ithaca, NY, 1960; pp 257−264.

(2) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379.

(3) Rush, T. S., III; Grant, J. A.; Mosyak, L.; Nicholls, A. *J. Med. Chem.* **2005**, *48*, 1489.

(4) Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144.

(5) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2005**, *1*, 153.

(6) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S., Jr.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765.

(7) Zhu, J.; Shi, Y.; Liu, H. *J. Phys. Chem. B* **2002**, *106*, 4844.

(8) Tjong, H.; Zhou, H.-X. *J. Chem. Theory Comput.* **2008**, *4*, 507.

(9) Gerig, J. T. *J. Org. Chem.* **2002**, *68*, 5244.

(10) Lee, M. S.; Olson, M. A. *J. Phys. Chem. B* **2005**, *109*, 5223.

(11) Miertz, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117.

(12) Harpaz, Y.; Gerstein, M.; Chothia, C. *Structure* **1994**, *2*, 641.

(13) Richards, F. M. *J. Mol. Biol.* **1974**, *82*, 1.

(14) Tsai, J.; Taylor, R.; Chothia, C.; Gerstein, M. *J. Mol. Biol.* **1999**, *290*, 253.

(15) Li, A.-J.; Nussinov, R. *Proteins* **1998**, *32*, 111.

(16) Gavezzotti, A. *J. Am. Chem. Soc.* **1983**, *105*, 5220.

(17) Finney, J. L. *J. Mol. Biol.* **1975**, *96*, 721.

(18) Gerstein, M.; Tsai, J.; Levitt, M. *J. Mol. Biol.* **1995**, *249*, 955.

(19) Pacios, L. F. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1427.

(20) Connolly, M. L. *J. Am. Chem. Soc.* **1985**, *107*, 1118.

(21) Meyer, A. Y. *J. Comput. Chem.* **1988**, *9*, 18.

(22) Pacios, L. F. *J. Mol. Model.* **1995**, *1*, 46.

(23) Liang, J.; Edelsbrunner, H.; Fu, P.; Sudhakar, P. V.; Subramaniam, S. *Proteins* **1998**, *33*, 1.

(24) Liang, J.; Edelsbrunner, H.; Fu, P.; Sudhakar, P. V.; Subramaniam, S. *Proteins* **1998**, *33*, 18.

(25) Estrada, E. *J. Phys. Chem. A* **2002**, *106*, 9085.

(26) Hirano, S.; Toyota, S.; Kato, M.; Toda, F. *Chem. Commun.* **2005**, 3646.

(27) Li, X.-Z.; He, J.-H.; Liao, D.-Z. *Inorg. Chem. Commun.* **2005**, *8*, 939.

(28) Castro, M.; Nicolás-Vázquez, I.; Zavala, J. I.; Sánchez-Viesca, F.; Berros, M. *J. Chem. Theory Comput.* **2007**, *3*, 681.

(29) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441.

(30) Bondi, A. *J. Phys. Chem.* **1966**, *70*, 3006.

(31) O'Keeffe, M.; Brese, N. E. *J. Am. Chem. Soc.* **1991**, *113*, 3226.

(32) Rowland, R. S.; Taylor, R. *J. Phys. Chem.* **1996**, *100*, 7384.

(33) Tsai, J.; Voss, N.; Gerstein, M. *Bioinformatics* **2001**, *17*, 949.

(34) Böhm, H.-J.; Ahlrichs, R. *J. Chem. Phys.* **1982**, *77*, 2028.

(35) Badenhoop, J. K.; Weinhold, F. *J. Chem. Phys.* **1997**, *107*, 5422.

(36) Chauvin, R. *J. Phys. Chem.* **1992**, *96*, 9194.

(37) Deb, B. M.; Singh, R.; Sukumar, N. *J. Mol. Struct: THEOCHEM* **1992**, *259*, 121.

(38) Arteca, G. A.; Grant, N. D. *J. Comput.-Aided. Mol. Des.* **1999**, *13*, 315.

(39) Nag, S.; Banerjee, K.; Datta, D. *New J. Chem.* **2007**, *31*, 832.

(40) Row, T. N. G.; Parthasarathy, R. *J. Am. Chem. Soc.* **1981**, *103*, 477.

(41) Price, S. L.; Stone, A. J.; Lucas, J.; Rowland, R. S.; Thornley, A. E. *J. Am. Chem. Soc.* **1994**, *116*, 4910.

(42) Schiemenz, G. P. *Z. Naturforsch., B: Chem. Sci.* **2007**, *62b*, 235.

(43) Blinn, J. F. *ACM Trans. Graphics* **1982**, *1*, 235.

(44) Friedrichs, M.; Zhou, R.; Edinger, S. R.; Friesner, R. A. *J. Phys. Chem. B* **1999**, *103*, 3057.

(45) Grant, J. A.; Pickup, B. T. *J. Phys. Chem.* **1995**, *99*, 3503.

(46) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. *J. Comput. Chem.* **1996**, *14*, 1653.

(47) Grant, J. A.; Pickup, B. T.; Nicholls, A. *J. Comput. Chem.* **2001**, *22*, 608.

(48) Pacios, L. F. *J. Phys. Chem.* **1991**, *95*, 10653.

(49) Pacios, L. F. *J. Phys. Chem.* **1992**, *96*, 7294.

(50) Pacios, L. F. *J. Phys. Chem.* **1994**, *98*, 3688.

(51) Pacios, L. F. *J. Comput. Chem.* **1995**, *16*, 133.

(52) Pacios, L. F. *J. Comput. Chem.* **2005**, *14*, 410.

(53) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.

(54) Mezey, P. G. *Mol. Phys.* **1999**, *96*, 169.

(55) Mezey, P. G. Topological similarity of molecules and the consequences of the holographic electron density theorem, an extension of the Hohenberg-Kohn theorem. In *Fundamentals of Molecular Similarity*; Carbó-Dorca, R., Girones, X., Mezey, P. G. Eds.; Kluwer Academic/Plenum Publishers: New York, NY, 2001; pp 113−124.

(56) Goh, S. K.; St-Amant, A. *Chem. Phys. Lett.* **1997**, *264*, 9.

(57) Exner, T. E.; Mezey, P. G. *J. Comput. Chem.* **2003**, *24*, 1980.

(58) Skylaris, C.-K.; Haynes, P. D.; Mostofi, A. A.; Payne, M. C. *J. Chem. Phys.* **2005**, *122*, 084119.

(59) Skylaris, C.-K.; Haynes, P. D. *J. Chem. Phys.* **2007**, *127*, 164712.

(60) Mezey, P. G. *Shape in Chemistry: Introduction to Molecular Shape and Topology*; VCH Publishers: New York,1992.

(61) Mezey, P. G. *Int. J. Quant. Chem. Quant. Biol. Symp.* **1986**, *12*, 113.

(62) Mezey, P. G. *J. Comput. Chem.* **1987**, *8*, 462.

(63) Mezey, P. G. *Int. J. Quant. Chem. Quant. Biol. Symp* **1987**, *14*, 127.

(64) Mezey, P. G. *J. Math. Chem.* **1988**, *2*, 299.

(65) Mezey, P. G. *J. Math. Chem.* **1988**, *2*, 325.

(66) Walker, P. D.; Arteca, G. A.; Mezey, P. G. *J. Comput. Chem.* **1993**, *14*, 1172.

(67) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao,O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03 Revision B.05*; Gaussian, Inc.: Wallingford, CT, 2004.

(68) Waldherr-Teshner, M.; Goetze, T.; Heiden, W.; Knoblauch, M.; Volhardt, H.; Brickmann, J. MOLCAD - Computer Aided Visualization and Manipulation of Models in Molecular Science. At Second Eurographics Workshop on Visualization in Scientific Computing; Delft, Netherlands, 1991.

(69) Brickmann, J.; Keil, M.; Exner, T. E.; Marhöfer, R.; Moeckel, G. Molecular Models: Visualization. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. J., III, Schreiner, P. R. Eds.; John Wiley & Sons: Chichester, 1998.

(70) Brickmann, J.; Exner, T. E.; Keil, M.; Marhöfer, R. *J. Mol. Model.* **2000**, *6*, 328.

(71) *Sybyl 6.7*; Tripos, Inc.: 1699 South Hamley Road, St. Louis, MO 63144, 2000.

CT800268C

JCTC Journal of Chemical Theory and Computation

# A Simple Definition of Ionic Bond Order

D. B. Chesnut*

*Department of Chemistry, Duke University, Durham, North Carolina 27708*

Received August 9, 2008

**Abstract:** Taking the square of the bond ionicity, $i_\mu^2(i,j)$, for molecular orbital $\mu$ in the localized orbital representation of Cisolowski and Mixon (*J. Am. Chem. Soc.* **1991**, *113*, 4142−4145) as an ionic bond order, one finds a simple and natural relation between the covalent bond order, $p_{\mathrm{cov},\mu}(i,j)$, and the ionic bond order, $p_{ion,\mu}(i,j)$: $p_{ion,\mu}(i,j) \equiv i_\mu^2(i,j) = 1 - [p_{\mathrm{cov},\mu}(i,j)]/[p_{\mathrm{cov},\mu}^{\max}(i,j)]$ where $p_{\mathrm{cov},\mu}^{\max}(i,j) = t_\mu(i,j)^2$ is the maximum value $p_{\mathrm{cov},\mu}(i,j)$ can attain and where $t_\mu(i,j)$ is the total orbital occupancy of the atoms-in-molecules basins involved. A number of examples and limitations of the method are presented using the B3LYP/6−311+G(d,p) density functional approach.

## I. Introduction

When chemists speak of chemical bonds they usually refer to them as "covalent" or "ionic" or, more generally, "polar" or "nonpolar". When electron pairs are nearly equally shared between two atoms of similar electronegativities, we think of the bond holding the atoms together as a covalent or nonpolar bond. Likewise, when the electronegativity difference is large, such as in NaF, the electron pair is basically held much more by the more electronegative element, and we call the bond ionic or polar. But, of course, pure covalent bonds are rare, and there are in actuality no pure ionic bonds because the electron pair involved will have some spatial component of the less electronegative element (unless the bond is completely broken into ion fragments, in which case there is no bond at all!). More generally, bonds are polar, and the extent of their polarity dictates our qualitative way of discussing them. That is, all bonds have some covalent character, and virtually all bonds between elements of differing electronegativity have some ionic character.

While the general nature of the chemical bond has an implicit understanding among chemists, its quantitative character must be *defined* in terms of reasonable molecular quantities, and there are a multitude of such definitions. Covalent bond orders have been defined in a variety of ways and are useful quantities in that they tend to reflect in a quantitative way the covalent bonding involved. Not only can individual covalent bond orders be defined but also total covalent bond orders for a given atom such as the atomic covalancy of Ángyán, Loos, and Mayer.[1] The covalent bond order of Ángyán et al. and the equivalent single determinant form of the delocalization index of Fradera, Austen, and Bader[2] along with the covalent bond order of Cioslowski and Mixon[3] (CM) can be used to measure such quantities. These bond orders have a strong theoretical foundation since they are based on the (spinless) electron pair density, most appropriate in discussing electron pairs. All these approaches employ the atoms-in-molecules (AIM) approach of Bader[4] in which the electronic space is divided into disjoint basins based on the vector field of the gradient of the electron density, $\rho(\vec{r})$. We focus here on the approach of Cioslowski and Mixon defined in a basis of localized orbitals that allow the bond character to be more easily visualized.

But while covalent bond orders are readily defined, there has, to this point, been no corresponding simple and readily available definition of ionic bond order. In a recent paper Gould et al.[5] suggest a definition for covalent and ionic bond indices that appears most promising, although it is not of immediate transparency. They point out that at the time of their publication the concept of an ionic bond index had not been thoroughly defined on an equal footing with the covalent bond index, and, in contrast to the present method (vide infra), their approach is not limited in its application. The purpose of our paper is to point out that such a bond order can be simply defined that readily follows the basic ideas of bond polarity. The defined ionic bond order is very natural, and its attractiveness lies in the ease with which it

* Corresponding author phone: (919)660-1537; fax: (919)660-1605; e-mail: donald.chesnut@duke.edu.

10.1021/ct800326n CCC: $40.75 
Published on Web 09/18/2008

fits in and complements the corresponding Cisolowski-Mixon covalent bond order.

## II. Theoretical Background

The covalent bond orders of Cioslowski and Mixon[3] are directly related to and more transparently understood by the ideas of the covalent bond order of Ángyán, Loos, and Mayer[1] and the delocalization index of Fradera, Austen, and Bader.[2] The bond orders between AIM atoms (basins) $i$ and $j$ are defined by a decomposition of the spinless pair density, $P_2(\vec{r}_1,\vec{r}_2)$, a decomposition that is due to McWeeny[6,7] who defined $P_2(\vec{r}_1,\vec{r}_2)$ in a way that accentuates the role of correlation by introducing the *correlation factor* $f(\vec{r}_1,\vec{r}_2)$.

$$P_2(\vec{r}_1,\vec{r}_2)=\rho(\vec{r}_1)\rho(\vec{r}_2)[1+f(\vec{r}_1,\vec{r}_2)] \quad (1)$$

In terms of this expression for $P_2(\vec{r}_1,\vec{r}_2)$ one then finds the interbasin pair number, $N_{ij}$, to be given by

$$N_{ij}=\int_{\Omega_i}d\vec{r}_1\int_{\Omega_j}d\vec{r}_2P_2(\vec{r}_1,\vec{r}_2)=\int_{\Omega_i}d\vec{r}_1\int_{\Omega_j}d\vec{r}_2\rho(\vec{r}_1)\rho(\vec{r}_2)[1+f(\vec{r}_1,\vec{r}_2)]=N_iN_j+\int_{\Omega_i}d\vec{r}_1\int_{\Omega j}d\vec{r}_2\rho(\vec{r}_1)\rho(\vec{r}_2)f(\vec{r}_1,\vec{r}_2)=N_iN_j-F_{ij} \quad (2)$$

which defines $F_{ij}$ and holds for all $i,j$, including the case $i=j$. $N_i$ is the (average) electron population of basin $i$. Fradera, Austen, and Bader[3] define $F_{ii}\equiv\lambda_i$ as the *atomic localization index*, and the sum $F_{ij}+F_{ji}=2F_{ij}\equiv\delta_{ij}$ as the *delocalization index*. They do not take the delocalization index as a bond order per se but rather as a measure of the number of electron pairs shared between basins $i$ and $j$.

The nature of the delocalization index is more clearly seen when we express it for a closed shell, single determinant wave function (such as a Hartree−Fock wave function or the Kohn−Sham wave function employed here) where it is given by integration of the exchange density over the two basins involved

$$\delta_{ij}=2F_{ij}=4\sum_{\mu,\nu}^{occ}\langle\mu|\nu\rangle_i\langle\nu|\mu\rangle_j \quad (3)$$

where standard bracket notation is employed, and where the subscripts $i,j$ refer to basins $\Omega_i$, $\Omega_j$. Clearly the delocalization index is nonzero only if (molecular) orbitals span (extend into) both basins $i$ and $j$. It will tend to more readily reflect a "true" bond order when polarization effects are absent but will be more complicated when such effects are present. In the single determinant approach the delocalization index given in eq 3 is *exactly* the topological covalent bond order defined by Ángyán, Loos, and Mayer.[1]

The Cioslowski-Mixon (CM) bond order[3] is found from a decomposition of the total number of electrons in the system. Employing the Bader-Stephens sum rule[8] one can show that

$$N=\sum_{i,j}F_{ij}=\sum_{i,j}\sum_{\nu,\mu}2\langle\nu|\mu\rangle_i\langle\mu|\nu\rangle_j=\sum_{i,j}\sum_{\mu}2\langle\mu|\mu\rangle_i\langle\mu|\mu\rangle_j+\sum_{i,j}\sum_{\nu<\mu}4\langle\nu|\mu\rangle_i\langle\mu|\nu\rangle_j \quad (4)$$

However, the last term in the final expression in eq 4 vanishes due to orbital orthogonality, so that we finally obtain

$$N=\sum_{i,j}\sum_{\mu}2\langle\mu|\mu\rangle_i\langle\mu|\mu\rangle_j=\sum_{i}\sum_{\mu}2\langle\mu|\mu\rangle_i\langle\mu|\mu\rangle_i+\sum_{i<j}\sum_{\mu}4\langle\mu|\mu\rangle_i\langle\mu|\mu\rangle_j \quad (5)$$

The expression for $N$ in eq 5 is quite general for single determinant wave functions.

Cioslowski and Mixon start with the expressions in eq 5, defining the first ($i=j$) term as $N_{atomic}$ and the second ($i<j$) as $N_{diatomic}$. They then perform an orbital localization procedure which maximizes $N_{atomic}$ while maintaining the first order density matrix constant, Cioslowski's *isopycnic* transformation,[9,10] a generalization of the unitary transformation to which it reduces for single determinant wave functions where occupation numbers are unity or zero. Cioslowski and Mixon then *define* a covalent bond order, $p_{\mathrm{cov}}(i,j)$, between AIM basins $i$ and $j$ in this specific representation as

$$p_{\mathrm{cov}}(i,j)=\sum_{\mu}4\langle\mu|\mu\rangle_i\langle\mu|\mu\rangle_j \quad (6)$$

These bond orders generally relate well to conventional ideas of single and multiple covalent bonds, as do those in the more general and invariant approach used by Ángyán, Loos, and Mayer.[1] The more general delocalization index is, as both Fradera et al.[2] and Ángyán, Loos, and Mayer point out, invariant to unitary transformations, while the CM bond order definition, in which only the $\mu=\upsilon$ "diagonal terms" of eq 3 are kept, is not. However, as Ángyán et al. point out, on the basis of the population-localized orbitals Cioslowski and Mixon use, the neglected off-diagonal terms are small and for strictly localizable systems may be negligible. Furthermore, the localization procedure is *not* arbitrary and is well defined for a given basis set. One of its greatest practical advantages is that in its current implementation[11] the various bond orders are readily seen from the individual orbital populations.

Cisolowski and Mixon go on to define the ionicity[12] of molecular orbital $\mu$ (in their representation) involving atoms (basins) $i$ and $j$ as

$$i_\mu(i,j)=\frac{\langle\mu|\mu\rangle_i-\langle\mu|\mu\rangle_j}{\langle\mu|\mu\rangle_i+\langle\mu|\mu\rangle_j} \quad (7)$$

While $i_\mu(i,j)$ does not have a simple relationship to the corresponding covalent bond order $p_{\mathrm{cov},\mu}(i,j)$ for molecular orbital $\mu$, the quantity $p_{ion,\mu}(i,j)\equiv i_\mu^2(i,j)$ does, where $p_{ion,\mu}(i,j)$ is our newly defined ionic bond order for orbital $\mu$. Defining the denominator of eq 7 as $t_\mu(i,j)$, the total occupancy of molecular orbital $\mu$ in basins $i$ and $j$, it is not difficult to show that

$$p_{ion,\mu}(i,j)\equiv i_\mu^2(i,j)=1-\frac{p_{\mathrm{cov},\mu}(i,j)}{t_\mu(i,j)^2} \quad (8)$$

which may also be written as

$$p_{ion,\mu}(i,j)\equiv i_\mu^2(i,j)=1-\frac{p_{\mathrm{cov},\mu}(i,j)}{p_{\mathrm{cov},\mu}^{\max}(i,j)} \quad (9)$$

since $t_\mu(i,j)^2$ is the maximum value $p_{\mathrm{cov},\mu}(i,j)$ can assume for the total occupancies $t_\mu(i,j)=\langle\mu|\mu\rangle_i+\langle\mu|\mu\rangle_j$. $p_{\mathrm{cov},\mu}^{\max}(i,j)$ acts as a scaling factor for the bond order; it will be unity when

virtually all the atomic occupancies of molecular orbital $\mu$ are contained in basins $i$ and $j$ and will also approach this value for a highly localized orbital.

The definition of ionic bond order given in eq 9 (and eq 8) is particularly simple, and its attractivelness lies in the fact that it fits in so naturally with the Cioslowski-Mixon definition of covalent bond order. Given a bonding orbital, the ratio of $p_{\mathrm{cov},\,\mu}(i,j)/p^{\max}_{\mathrm{cov},\,\mu}(i,j)$ varies from (near) zero to unity as the ionic bond order varies from unity to zero. Because of the nature of the atomic orbitals, there will always be some overlap of orbitals into adjacent basins so $p_{\mathrm{cov},\,\mu}(i,j)$ can never actually be zero.

## III. Details of the Calculations

The CM bond orders as well as the general optimizations and minimum-confirming frequency calculations were carried out in the B3LYP/6−311+G(d,p) approach[13,14] with Gaussian 03.[11] In our tables the doubly occupied localized orbitals are listed by increasing negative kinetic energy; accordingly, core orbitals come near the top of the list and valence and lone pair orbitals near the bottom. Our characterization of the orbital type is based on a combination of the energy listing and our basic understanding of the constitution of core, valence, and lone pair orbitals.

## IV. Results and Discussion

For the most part the examples we cite below are straightforward. To the extent a bond is covalent, that is, has a high covalent bond order, it is not ionic, that is, it has a small ionic bond order. Conversely, when a bond has a high ionic bond order, its covalent order will be small. However, before we exhibit a number of examples, we take a brief interlude to talk about some qualitative ideas of chemical bonding.

**IV.1. An Interlude on Chemical Bonding.** The application of the results expressed in eq 9 (or eq 8) requires some restraint as well as some understanding of what chemists mean when they refer to a chemical bond. We will discuss this generally here and show applications of this discussion in the material that follows.

We shall find that the Li 1s core orbital is spread slightly into the hydrogen atomic basin in LiH. We can, therefore, calculate both a (very small) covalent bond order and a corresponding ionic bond nearly equal to unity. But, we normally do not consider this as representing an ionic bond, basically because the two electrons occupying the Li 1s core both "belong" to the Li atom. The oxygen lone pairs in $H_2CO$ have a slight spread into the C atomic basin and even a very small amount into the hydrogen atomic basins. Yet we do not characterize these interactions as constituting highly ionic chemical bonds, again because the two electrons involved in each of the two lone pair orbitals have their origins in the oxygen atom. The orbitals that dominate both CO bonding orbitals have contributions from the two hydrogen atomic basins. Accordingly, there can be defined a covalent bond order between oxygen and each hydrogen in these orbitals as well as an ionic bond order. Such secondary bonding is a result of the inability to completely localize any molecular orbital, but we do not characterize such interactions as covalent or ionic. Do all these interactions contribute to the stability of the molecular species? Yes, otherwise they would not occur as they do. But it is not meaningful to talk about "bonds" in these cases; rather one should talk about stabilizing interactions of special kinds.

***Table 1.*** Atomic Occupancies and Covalent ($p_{cov}$) and Ionic ($p_{ion}$) Bond Orders for LiH and LiF[a]

A. LiH

| orbital | Li | H | orbital type | $p_{cov}$ | $p_{ion}$ |
|---|---|---|---|---|---|
| 1 | 0.9962 | 0.0038 | Li core | (0.0151 | 0.9849) |
| 2 | 0.0556 | 0.9444 | bond orbital | 0.2100 | 0.7900 |

B. LiF

| orbital | Li | F | orbital type | $p_{cov}$ | $p_{ion}$ |
|---|---|---|---|---|---|
| 1 | 0.0000 | 1.0000 | F core | | |
| 2 | 0.0002 | 0.9998 | F lone pair | (0.0008 | 0.9992) |
| 3 | 0.9925 | 0.0075 | Li core | (0.0298 | 0.9702) |
| 4 | 0.0274 | 0.9726 | bond orbital | 0.1066 | 0.8934 |
| 5 | 0.0103 | 0.9897 | F lone pair | (0.0408 | 0.9592) |
| 6 | 0.0103 | 0.9897 | F lone pair | (0.0408 | 0.9592) |

[a] Data for the core and lone pair orbitals are given in parentheses.

It is appropriate to discuss the covalent and ionic character of a bond when the orbital in question is part of the valence space, that is, where electrons from two or more dominant atomic species are involved. In the first few examples we discuss we include in the tables all the interactions to illustrate the points made above. But after this we shall include only the covalent and ionic character of bonds in the valence space.

**IV.2. Examples of Ionic Bonds.** Illustrative of highly ionic bonds are those in the lithium and sodium hydrides and fluorides and the $FHF^-$ anion.

Table 1 exhibits data for LiH and LiF and shows in parentheses contributions from core and lone pair orbitals, cases where, as discussed above, one should strictly speaking not talk about a chemical bond. The singular bonds in LiH and LiF are highly ionic with ionic bond orders of 0.7900 and 0.8934, respectively. Note in the two cases that small but finite covalent bond orders are determined for both core and lone pair orbitals, the latter being quite significant in LiF where three fluorine lone pairs are present. Note also that since we are dealing here with diatomics the sum of the covalent and ionic bond orders sum to unity.

The results for NaH and NaF shown in Table 2 are similar to those of the lithium analogues. Here again we show all calculated bond orders including core and lone pairs where we really cannot talk about chemical bonds. Interestingly, the bond in NaH is of equal covalent and ionic parts (as is KH, data not shown), but the bond in NaF is highly ionic, with an ionic bond order of 0.8453. In this regard since Li, Na, and K all have comparable electronegativities, it is a bit surprising that the covalent bond order of LiH is as small as it is.

Table 3 shows the various contributions to the covalent bond orders in these four molecules. Note that the large contribution by lone pairs in LiF causes the total covalent bond orders in LiH and LiF to be about the same. There is also a noticeable contribution by lone pairs in NaF, but the

***Table 2.*** Atomic Occupancies and Covalent ($p_{cov}$) and Ionic ($p_{ionN}$) Bond Orders for NaH and NaF[a]

A. NaH

| orbital | Na | H | orbital type | $p_{cov}$ | $p_{ion}$ |
|---|---|---|---|---|---|
| 1 | 1.0000 | 0.0000 | Na $n = 1$ core | | |
| 2 | 0.9955 | 0.0046 | Na $n = 2$ core | (0.0183 | 0.9817) |
| 3 | 0.9989 | 0.0011 | Na $n = 2$ core | (0.0044 | 0.9956) |
| 4 | 0.9989 | 0.0011 | Na $n = 2$ core | (0.0044 | 0.9956) |
| 5 | 1.0000 | 0.0000 | Na $n = 2$ core | | |
| 6 | 0.1494 | 0.8506 | bond orbital | 0.5083 | 0.4917 |

B. NaF

| orbital | Na | F | orbital type | $p_{cov}$ | $p_{ion}$ |
|---|---|---|---|---|---|
| 1 | 1.0000 | 0.0000 | Na $n = 1$ core | | |
| 2 | 0.0000 | 1.0000 | F core | | |
| 3 | 0.9926 | 0.0074 | Na $n = 2$ core | (0.0294 | 0.9706) |
| 4 | 0.9982 | 0.0018 | Na $n = 2$ core | (0.0072 | 0.9928) |
| 5 | 0.9982 | 0.0018 | Na $n = 2$ core | (0.0072 | 0.9928) |
| 6 | 1.0000 | 0.0000 | Na $n = 2$ core | | |
| 7 | 0.0008 | 0.9992 | F lone pair | (0.0032 | 0.9968) |
| 8 | 0.0403 | 0.9597 | bond orbital | 0.1547 | 0.8453 |
| 9 | 0.0087 | 0.9913 | F lone pair | (0.0345 | 0.9655) |
| 10 | 0.0087 | 0.9913 | F lone pair | (0.0345 | 0.9655) |

[a] Data for the core and lone pair orbitals are given in parentheses.

***Table 3.*** Contributions to the Total Covalent Bond Orders in LiH, LiF, NaH, and NaF from Core, Lone Pair, and Bond Orbitals

| | LiH | LiF |
|---|---|---|
| Li core | 0.0151 | 0.0298 |
| F lone pairs | | 0.0824 |
| bond | 0.2100 | 0.1066 |
| total | 0.2251 | 0.2188 |

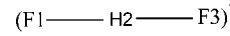
| | NaH | NaF |
|---|---|---|
| Na core | 0.0271 | 0.0438 |
| F lone pairs | | 0.0722 |
| bond | 0.5083 | 0.1547 |
| total | 0.5354 | 0.2707 |

low bond contribution in NaF makes its total covalent bond order about half that of NaH.

The interesting case of the $FHF^-$ anion is illustrated by the data in Table 4. Here all orbital occupancies are listed, but bond orders are only shown for orbitals 9 and 10 which represent the chemical bonds in this system. Note that these are basically two-center two-electron bonds; that is, the CM decomposition does not reveal any three-center bonds which might have been present here. The ionic bond order of the two bonds is quite high showing that one can represent this system basically as F−H+F−.

**IV.3. Some Examples of Typical Covalent Bonds.** Illustrative of typical covalently bonded but polar bonds are $H_2CO$ and $HNCH_2$ whose data are given in Tables 5 and 6. The CH and CO bonds in formaldehyde are shown as orbitals 7,8 and 4,6, respectively. As expected the CH bonds are nearly pure covalent bonds, while those between C and O are polar, exhibiting ionic bond orders of 0.3792 and 0.2313. Because of the polarity of the CO bonds, the covalent character of these sums to 1.3566, considerably less than 2.0 that would be expected for a pure covalent double bond.

***Table 4.*** Atomic Occupancies, Covalent ($p_{cov}$) and Ionic ($p_{ion}$) Bond Orders, and the Covalent Bond Order Matrix for the Linear FHF Anion[a]




A. Orbital Occupancies

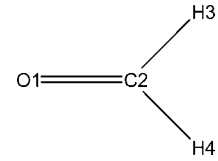
| orbital | F1 | H2 | F3 | orbital type | $p_{cov}$ | $p_{ion}$ |
|---|---|---|---|---|---|---|
| 1 | 0.0000 | 0.0000 | 1.0000 | F3 core | | |
| 2 | 1.0000 | 0.0000 | 0.0000 | F1 core | | |
| 3 | 0.0002 | 0.0001 | 0.9996 | F3 lone pair | | |
| 4 | 0.9996 | 0.0001 | 0.0002 | F1 lone pair | | |
| 5 | 0.0036 | 0.0042 | 0.9922 | F3 lone pair | | |
| 6 | 0.9922 | 0.0042 | 0.0036 | F1 lone pair | | |
| 7 | 0.0036 | 0.0042 | 0.9922 | F3 lone pair | | |
| 8 | 0.9922 | 0.0042 | 0.0036 | F1 lone pair | | |
| 9 | **0.9066** | **0.0726** | 0.0208 | F1H2 bond | 0.2633 | 0.7254 |
| 10 | 0.0208 | **0.0726** | **0.9066** | H2F3 bond | 0.2633 | 0.7254 |

B. Covalent Bond Order Matrix

| | F1 | H2 | F3 |
|---|---|---|---|
| F1 | 9.5806 | | |
| H2 | **0.3033** | 0.0212 | |
| F3 | 0.2110 | **0.3033** | 9.5806 |

[a] The primary bond data are given in bold.

***Table 5.*** Orbital Occupancies, Covalent ($p_{cov}$) and Ionic ($p_{ion}$) Bond Orders for the Indicated Bonds, and the Covalent Bond Order Matrix for $H_2CO$[a]




A. Orbital Occupancies

| orbital | O1 | C2 | H3 | H4 | type | $p_{cov}$ | $p_{ion}$ |
|---|---|---|---|---|---|---|---|
| 1 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | O core | | |
| 2 | 0.0008 | 0.9992 | 0.0000 | 0.0000 | C core | | |
| 3 | 0.9989 | 0.0008 | 0.0002 | 0.0002 | O lone pair | | |
| 4 | **0.8015** | **0.1906** | 0.0040 | 0.0040 | OC bond | 0.6111 | 0.3792 |
| 5 | 0.9448 | 0.0338 | 0.0107 | 0.0107 | O lone pair | | |
| 6 | **0.7292** | **0.2556** | 0.0076 | 0.0076 | OC bond | 0.7455 | 0.2313 |
| 7 | 0.0244 | **0.5104** | **0.4553** | 0.0099 | CH bond | 0.9295 | 0.0033 |
| 8 | 0.0244 | **0.5104** | 0.0099 | **0.4553** | CH bond | 0.9295 | 0.0033 |

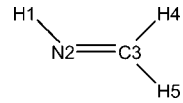
B. Covalent Bond Order Matrix

| | O1 | C2 | H3 | H4 |
|---|---|---|---|---|
| O1 | 8.1318 | | | |
| C2 | **1.5901** | 3.2444 | | |
| H 3 | 0.1214 | **0.9619** | 0.4152 | |
| H4 | 0.1214 | **0.9619** | 0.0368 | 0.4152 |

[a] The primary bond data are indicated in bold.

The covalent bond orders for the two CC bonds in ethylene (data not shown) are 0.9513 and 0.8832 summing to 1.8345, and with zero ionic bond order due to the molecular symmetry. The covalent and ionic bond orders for the CO and OH bonds in ethanol (data not shown) are 0.6959, 0.2726 and 0.6525, 0.3288, respectively.

***Table 6.*** Atomic Occupancies, Covalent ($p_{cov}$) and Ionic ($p_{ion}$) Bond Orders for the Indicated Bonds, and the Covalent Bond Order Matrix for $HNCH_2$[a]




A. Atomic Occupancies of Localized Orbitals

| orbital | H1 | N2 | C3 | H4 | H5 | orbital type | $p_{cov}$ | $p_{ion}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.0000 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | N core | | |
| 2 | 0.0000 | 0.0003 | 0.9997 | 0.0000 | 0.0000 | C core | | |
| 3 | 0.0069 | 0.9749 | 0.0110 | 0.0041 | 0.0031 | N lone pair | | |
| 4 | 0.0051 | **0.7197** | **0.2637** | 0.0078 | 0.0036 | NC bond | 0.7591 | 0.2150 |
| 5 | **0.3139** | **0.6639** | 0.0162 | 0.0023 | 0.0038 | H1N bond | 0.8836 | 0.1281 |
| 6 | 0.0087 | **0.6276** | **0.3412** | 0.0112 | 0.0113 | NC bond | 0.8565 | 0.0874 |
| 7 | 0.0029 | 0.0206 | **0.5139** | 0.0088 | **0.4538** | CH5 bond | 0.9328 | 0.0039 |
| 8 | 0.0019 | 0.0189 | **0.5120** | **0.4590** | 0.0082 | CH4 bond | 0.9400 | 0.0030 |

B. Covalent Bond Orders

| | H1 | N2 | C3 | H4 | H5 |
|---|---|---|---|---|---|
| H1 | 0.1974 | | | | |
| N2 | **0.8976** | 6.6078 | | | |
| C3 | 0.0477 | **1.7834** | 3.4240 | | |
| H4 | 0.0071 | 0.1081 | **0.9820** | 0.4220 | |
| H5 | 0.0106 | 0.0988 | **0.9694** | 0.0318 | 0.4124 |

[a] The primary bond data are indicated in bold.

Methylene imine, $HNCH_2$, is an interesting molecule in that a contribution to N2C3 covalent bond order is made by the nitrogen lone pair. As the data in Table 6 show, both the HN and NC bonds are polar, while the CH bonds are virtually pure covalent.

**IV.4. Two Unusual Cases and Limitations of the Method.** Finally we treat the unusual (hypervalent) $PF_5$ and diborane molecules. Table 7 contains the dominant bond occupancies for these two cases.

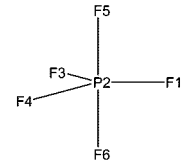
The axial and equatorial bonds in $PF_5$ are very similar, having similar covalent and ionic bond orders and (optimized) bond distances as well, 1.6051 and 1.5709 Å, respectively. Notably there are no three center bonds in the CM localized set of orbitals.

Diborane is a different case and illustrates one of the limitations of specifiying bond orders of either type. While the external BH bonds are two-center dominated (orbital 7), the internal BHB bonds (illustrated by orbital 3) are three-center in character. Our formulas do not easily reconcile how one is to characterize the covalent and ionic character of such bonds. There is obviously covalent character, and the bond, while not polar, is polarized toward the bridging hydrogen. Here we simply take the average of the two two-center fragment data (both the same due to the molecular symmetry) to characterize this bond.

The situation in diborane typifies the general failure of the CM localization process to always provide two-center dominant localized orbitals. Indeed, Cioslowski and Mixon[3] note that unique localized orbitals are generally found when one can associate with the molecule a single, dominant Lewis structure. This is not the case for hightly symmetric molecules like benzene or the cyclopentadinyl anion which do not converge in the CM treatment.

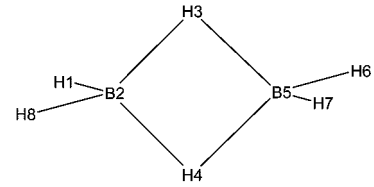
***Table 7.*** Orbital Occupancies and Their Covalent ($p_{cov}$) and Ionic ($p_{ion}$) Bond Orders for the Two Classes of Bonds in $PF_5$ and the External and Internal BH and BHB Bonds in Diborane[a,b]

A. $PF_5$




| | F1 | P2 | F3 | F4 | F5 | F6 | $p_{cov}$ | $p_{ion}$ |
|---|---|---|---|---|---|---|---|---|
| 27 | **0.8629** | **0.1091** | 0.0030 | 0.0030 | 0.0110 | 0.0110 | 0.3766 | 0.6014 |
| 30 | 0.0121 | **0.1003** | 0.0122 | 0.0122 | 0.0002 | **0.8630** | 0.3462 | 0.6269 |

B. $H_2BH_2BH_2$ (diborane)




| | H1 | B2 | H3 | H4 | B5 | H6 | H7 | H8 |
|---|---|---|---|---|---|---|---|---|
| 3 | 0.0136 | **0.1089** | **0.6947** | 0.0332 | **0.1089** | 0.0136 | 0.0136 | 0.0136 |
| 7 | **0.7391** | **0.1980** | 0.0191 | 0.0191 | 0.0038 | 0.0013 | 0.0022 | 0.0176 |

| orbital | $p_{cov}$ | $p_{ion}$ |
|---|---|---|
| 3 | 0.3026[a] | 0.5314[a] |
| 7 | 0.5854 | 0.3334 |

[a] Average of the two-center fragments in orbital 3. [b] The dominant bond data are given in bold.

## V. Summary

Taking the square of the bond ionicity, $i_{\mu}^{2}(i,j)$, for molecular orbital $\mu$ in the localized orbital representation of Cisolowski and Mixon[3] as an ionic bond order, one finds a simple and natural relation between the scaled covalent bond order, $p_{\mathrm{cov},\mu}(i,j)/p_{\mathrm{cov},\mu}^{\max}(i,j)$, and the ionic bond order, $p_{ion,\mu}(i,j)$. Their sum is unity; to the extent that an orbital is covalent, it is not ionic, and to the extent that it is ionic it is not covalent. A number of examples have been presented using the B3LYP/6−311+G(d,p) density functional approach as well as limitations of the method.

**Acknowledgment.** I am indebted to Duke University and the Center for Applied Computational Studies at East Carolina University for providing CPU time that allowed these calculations to be carried out and the Beratan research group for technical and computational assistance.

### References

(1) Ángyán, I.; Loos, M.; Mayer, I. *J. Phys. Chem.* **1994**, *98*, 5244–5248.

(2) Fradera, X.; Austen, M. A.; Bader, R. F. W. *J. Phys. Chem. A* **1999**, *103*, 304–314.

(3) Cioslowski, J.; Mixon, S. T. *J. Am. Chem. Soc.* **1991**, *113*, 4142–4145.

(4) Bader, R. F. *Atoms in Molecules: A Quantum Theory*; Oxford University Press: Oxford, 1994.

(5) Gould, M. D.; Taylor, C.; Wolff, S. K.; Chandler, G. S.; Jayatilaka, D. *Theor. Chem. Acc.* **2008**, *119*, 275–300.

(6) McWeeny, R. *Methods of Molecular Quantum Mechanics*, 2nd ed.; Academic Press: New York, 1989.

(7) McWeeny, R. *Rev. Mod. Phys.* **1960**, *32*, 335–369.

(8) Bader, R. F. W.; Stephens, M. E. J. *Am. Chem. Soc.* **1975**, *97*, 7391–7399.

(9) Cioslowski, J. *J. Math. Chem.* **1991**, *8*, 169–178.

(10) Cioslowski, J. *Int. J. Quant. Chem. Quant. Chem. Symp.* **1990**, *24*, 15–28.

(11) 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.: Pittsburgh, PA, 2003.

(12) Cioslowski, J.; Mixon, S. T. *Inorg. Chem.* **1993**, *32*, 3209–3216.

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

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

CT800326N

JCTC Journal of Chemical Theory and Computation

# DFT Study and Monte Carlo Simulation on the Aminolysis of $XC(O)OCH_3$ (X = $NH_2$, H, and $CF_3$) with Monomeric and Dimeric Ammonias

Xuefei Xia,[†] Chenghua Zhang,[†] Ying Xue,*[,†,‡] Chan Kyung Kim,[§] and Guosen Yan[†]

*College of Chemistry, Key Laboratory of Green Chemistry and Technology in Ministry of Education, Sichuan University, Chengdu 610064, P. R. China, State Key Laboratory of Biotherapy, Sichuan University, Chengdu 610041, P. R. China, and Department of Chemistry, Inha University, Inchon 402-751, Korea*

Received March 24, 2008

**Abstract:** The aminolysis of substituted methylformates ($XC(O)OCH_3$, X = $NH_2$, H, and $CF_3$) in the gas phase and acetonitrile are investigated by the density functional theory B3LYP/6−311+G(d,p) method and Monte Carlo (MC) simulation with free energy perturbation (FEP) techniques. The direct and the ammonia-assisted aminolysis processes are considered, involving the monomeric and dimeric ammonia molecules, respectively. In each case, two different pathways, the concerted and stepwise, are explored. The calculated results show that, for the direct aminolysis, the activation barrier of the concerted path is lower than that of the rate-controlling step of the stepwise process for all three reaction systems. In contrast, for the ammonia-assisted mechanism, the stepwise process is more favorable than the concerted pathway. The substituent effects at the carboxyl C atom of methylformate are discussed. This aminolysis of substituted methylformates is more favored for X = $CF_3$ than for X = H and $NH_2$ in the gas phase for both the direct and the ammonia-assisted processes. Solvent effects of $CH_3CN$ on the reaction of $HC(O)OCH_3 + nNH_3$ ($n$ = 1, 2) are determined by Monte Carlo simulation. The potential energy profiles along the minimum energy paths in the gas phase and in acetonitrile are obtained. It is shown that $CH_3CN$ lowers the energy barriers of all reactions.

## 1. Introduction

As the model for the formation of peptide bonds, the aminolysis of esters involved in the interaction of carbonyl group with nucleophile is under active investigation by using experimental and theoretical methods.[1−15] Up to now, three possible reaction pathways in accordance with the available kinetic results have been generally discussed in literature. The first one is the concerted pathway where the cleavage of C−O bond, the formation of C−N bond, and the transfer of H atom from the N atom at amine to the O atom proceed simultaneously. The second is the stepwise (addition/elimination) mechanism through neutral intermediates. The final way is the stepwise process involving zwitterionic intermediates in the reaction. Solvent effects of acetonitrile as well as the general base catalysis by the amine have been studied.[3,4]

Some authors made efforts to theoretically study the stepwise pathway through zwitterionic intermediates.[13−17] Gorb et al.[16] used three kinds of solvent models to study the mechanism of formamide hydrolysis from *ab initio* calculations and QM/MM molecular dynamics simulations. Their calculations showed that the zwitterionic intermediate is quite easily dissociated and could play a role in the hydrolysis of substituted amides or peptides. When Chalmet et al.[17] reported a theoretical study on the model reaction of ammonia and formic acid, their computations with the continuum model did not predict a stable zwitterionic intermediate, whereas a local energy minimum was found by explicit consideration of four solvent water molecules.

* Corresponding author e-mail: yxue@scu.edu.cn.

[†] Key Laboratory of Green Chemistry and Technology in Ministry of Education, Sichuan University.

[‡] State Key Laboratory of Biotherapy, Sichuan University.

[§] Inha University.

10.1021/ct800099a CCC: $40.75 
Published on Web 09/04/2008

For the reaction of methylformate with ammonia, Ilieva et al. did not succeed in applying the MP2/6−31G(d, p) method to identify zwitterionic transition states and intermediates.[18] They reported that two explicit water molecules are needed to obtain a very shallow minimum. When concerning the formation of a zwitterion between methylformate and ammonia and hydrazine in water, Singleton and Merrigan[19] used the B3LYP/6−31G(d, p) method to calculate various solvated structures involving from four to eleven explicit water molecules; however, they failed to find global minima for these structures. Recently, Sung and his cooperators[20] have studied the structures and stability of zwitterionic complexes in the aminolysis of phenyl acetate with ammonia and pointed out that at least five explicit water molecules are needed to stabilize the zwitterionic intermediate.

Differences in the ester structures,[21−25] amine nature,[26−28] and reaction medium[21,29] can result in changes in the reaction mechanism and the rate determining step. Antonczak et al.[30] first examined electrostatic solvent effects on the hydrolysis of formamide by using a dielectric continuum solvent model. Besides, they particularly studied the water-assisted hydrolysis processes at the MP3/6−31G**//3−21G *ab initio* level. They suggested that electrostatic interactions with the continuum should not significantly modify the energetics of the process. Sordo et al.[31−38] thoroughly investigated the possible mechanisms of the aminolysis reaction systems, such as the aminolysis reaction between the $\epsilon$-amino group of Lysine 199 and benzylpenicillin,[31] the water-assisted aminolysis of 2-azetidinone,[32,35] the aminolysis of monocyclic $\beta$-lactams,[33] the $NH_3$-assisted aminolysis of $\beta$-lactams[34] and penicillins,[37] and the aminolysis of monobactams.[38] When they studied the aminolysis of $\beta$-lactams, they predicted that positively charged ethanolamine molecules can act as bifunctional catalysts.[36] In addition, they also considered water-assisted mechanistic routes and the solvent effects. Ilieva and his co-workers[18] studied the aminolysis of methylformate with the QCISD/6−31G(d, p) and B3LYP/6−31G(d) methods and found that the neutral stepwise and the concerted pathways have very similar activation energy, and the presence of aprotic solvent acetonitrile fully lowers all energy barriers. For the aminolysis of 2-benzoxazolinone with methylamine, Ilieva et al.[39] predicted theoretically the concerted mechanism is most favorable in all three possible pathways at the B3LYP/6−31G(d) level of theory. Zipse et al.[12,13] studied the mechanism of the reaction of methyl acetate with methylamine and obtained the single point MP2/6−31G(d, p) energies through the HF/3−21G and HF/6−31G(d, p) optimized structures. Their results indicated that the stepwise (addition/elimination) pathway is more favorable than the concerted pathway. However, Jin et al.[40] found that the reaction prefers the concerted pathway to the neutral stepwise process in the gas phase and solutions concerning the aminolysis of phenylformate.

To our knowledge, few efforts have been made to investigate systemically the substituent effects on the ester aminolysis by applying a higher level of electron structure theory and solvent effects determined by Monte Carlo simulation from a theoretical point of view. We employed the density functional theory B3LYP/6−31G(d, p) method to study the substituent effects of the leaving groups on aminolysis of p-substituted phenyl acetates with ammonia.[41] In the present work, we aim to examine the effects of the nonleaving group substituents and solvents on the aminolysis of methylformates $XC(O)OCH_3$ (X = H, $NH_2$, and $CF_3$). In each case of the concerted and stepwise pathways, two processes, the direct aminolysis (with monomeric ammonia molecule) and the ammonia-assisted aminolysis (with dimeric ammonia), are considered (see Scheme 1). For the gas-phase reactions, the hybrid density functional theory (B3LYP) has been used in our calculations. In our previous work, we adopted the quantum chemical molecular orbital method and the Monte Carlo (MC) simulation with the free energy perturbation (FEP) technique to study the effects of solvents on the aza-Wittig reaction of iminophosphoranes,[42] the isomerization of imidazolines,[43] and the hydrolysis of N-(2-oxo-1,2-dihydropyrimidinyl) formamide.[44] In this study, the potential energy profiles of the direct and ammonia-assisted aminolysis processes of the compound methylformate in the gas phase and in $CH_3CN$ are obtained. The effect of solvent $CH_3CN$ is studied using the Monte Carlo free energy perturbation method.

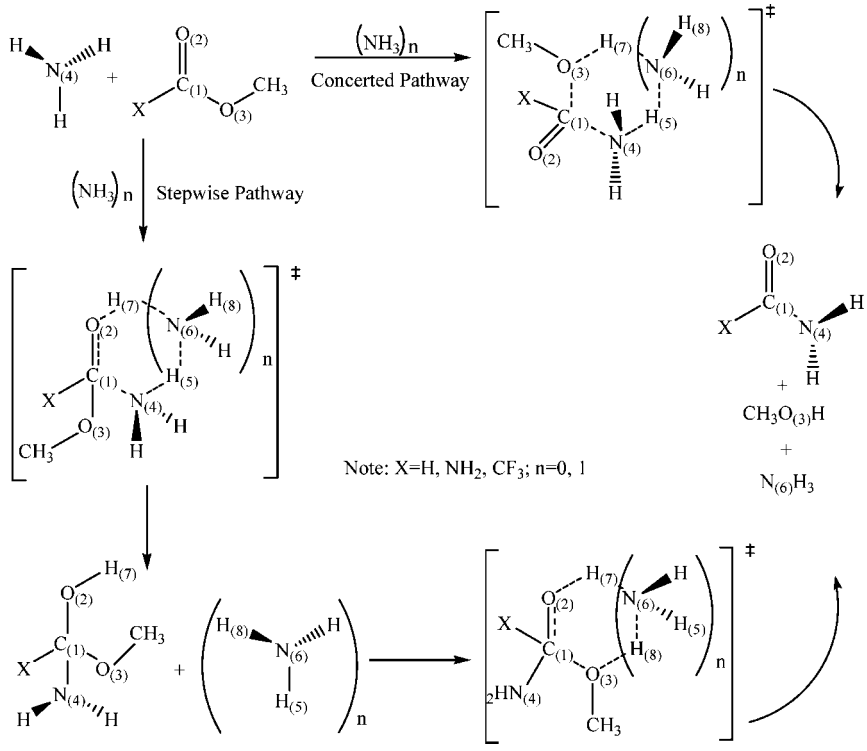
## 2. Computational Details

**2.1. Gas-Phase Calculation.** All calculations were carried out using Gaussian 03 program package.[45] To test the reliability of the theoretical approaches for the calculation of the energetics of reaction systems, we performed calculations for the concerted process of the direct aminolysis of the $HC(O)OCH_3$ molecule with a variety of calculational levels such as HF/6−31G(d, p), B3LYP/6−31G(d, p), B3LYP/6−311+G(d, p), MP2/6−31G (d, p), MP2/6−311+G (d, p), and G2(MP2). Comparison of the results obtained by the methods above indicated that the B3LYP/6−311+G(d, p) level is acceptable and a relatively economical option in this work (see the detailed discussion in the next section). Therefore, the B3LYP/6−311+G(d, p) method was selected for the ensuing calculations on the title reactions.

The geometric structures of all the reactant complexes, product complexes, intermediates, and transition states of aminolysis of $XC(O)OCH_3$ (X = H, $NH_2$, and $CF_3$) reaction systems were optimized at the B3LYP/6−311+G(d, p) level. The harmonic vibrational frequencies of each stationary point were calculated at the same level by diagonalizing the force constant matrix to characterize it as a true minimum with no imaginary frequency or a transition state with only one imaginary frequency. The frequency calculations without scaling also provided the thermodynamic quantities such as the zero-point vibrational energy, thermal correction, enthalpies, Gibbs free energies, and entropies at a temperature of 298.15 K and a pressure of 1.0 atm.

All transition states were checked by intrinsic reaction coordinate (IRC)[46] calculations. In the IRC calculations for the $HC(O)OCH_3 + nNH_3$ ($n$ = 1, 2) reaction systems, the "IRC=tight" option was used to generate the minimum energy path (MEP) in the gas phase. The MEPs at the B3LYP/6−311+G(d, p) level were constructed with a step size of 0.05 amu1/2 bohr. For all points along MEP, the

***Scheme 1.*** Reaction Paths of the Aminolysis of $XC(O)OCH_3$ (X = $NH_2$, H, and $CF_3$) with Monomeric and Dimeric Ammonia Molecules




partial atomic charges were obtained via the Natural Bond Orbital Theory (NBO)[47] at the B3LYP/6−311+G(d, p) level.

**2.2. Monte Carlo Simulation.** $CH_3CN$ was used as solvent to study the solvent effects on the aminolysis of methylformate by the Monte Carlo simulation with statistical perturbation theory.[48] Given a distance $r_{ij}$ between atom $i$ in molecule A and atom $j$ in molecule B, the intermolecular interaction potential function for solute−solvent and solvent−solvent interactions was described by Coulomb and Lennard-Jones terms as shown in eq 1

$$\Delta E_{AB} = \sum\sum \{q_i q_j e^2/r_{ij} + 4\varepsilon_{ij}[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^6]\} \quad (1)$$

The crossing terms $\sigma_{ij}$ and $\varepsilon_{ij}$ in eq 1 were obtained by the combination rules

$$\sigma_{ij} = \sqrt{\sigma_{ii} \times \sigma_{jj}}, \quad \varepsilon_{ij} = \sqrt{\varepsilon_{ii} \times \varepsilon_{jj}} \quad (2)$$

No intramolecular terms were included. The $\epsilon$ and $\sigma$ constants for the solute were taken from the OPLS all-atom parameters of the BOSS 4.2 database.[49] As the partial charge of atom $i$, the $q_i$ was obtained from the gas-phase calculation above. For the solvent $CH_3CN$, the OPLS united-atom models were adopted, and the parameters were also taken from the BOSS 4.2 database.

For the reaction path obtained by DFT studies, the geometries and partial charges along the minimum energy path (MEP) were incorporated into a molecular mechanical potential presented by eq 1 for the reaction system and then applied to calculate free energy changes of solvation along the MEP. The reaction system was immersed in the periodic box containing 390 explicit solvent molecules, which had dimensions of 26.7 Å × 26.7 Å × 40.0 Å. Preferential sampling was applied in the Metropolis algorithm, and the perturbations were performed using double-wide sampling in 51, 95, 51, and 77 windows for the concerted mechanism in the direct aminolysis reaction (designated DC), the stepwise mechanism of the direct aminolysis reaction (DS), the ammonia-assisted aminolysis reaction through the concerted mechanism (AC), and the ammonia-assisted aminolysis reaction through the stepwise mechanism (AS) in solvent $CH_3CN$, respectively. Every simulation included $2 \times 10^6$ configurations for equilibration, followed by $4 \times 10^6$ configurations of averaging in the isothermal, isobaric (NPT) ensemble at 298.15 K and 1 atm. For the solute−solvent and solvent−solvent interactions, a cutoff of 12.0 Å was employed for solvent $CH_3CN$. Finally we added the gas-phase relative energies to the computed free energy changes of solvation for obtaining the total potential energy profile along the reaction path in solution. All Monte Carlo simulation calculations were performed using the BOSS 4.2 program package.[49]

## 3. Results and Discussions

**3.1. Test Calculations.** For an accurate estimation of the energies, the concerted process in the direct aminolysis of $HC(O)OCH_3$ was studied at different computational levels and basis sets up to G2(MP2). The thermodynamic data relative to the reactant complex are listed in Table 1. Because the experimental results on this step are unavailable, it is impossible to carry out the comparison between theoretical and observed activation barriers. However, the theoretically predicted activation data at the G2(MP2) level of theory can serve as benchmark values for comparison (see Table 1). One can see from Table 1 that the HF/6−31G(d, p) method overestimates the activation electronic energies, enthalpies,

***Table 1.*** Activation Energies ($\Delta E^{\ddagger}$), Zero-Point Vibrational Energies ($\Delta E_{ZPV}{}^{\ddagger}$), Enthalpies ($\Delta H^{\ddagger}$), and Gibbs Free Energies ($\Delta G^{\ddagger}$) in the Concerted Pathway of Aminolysis of $HC(O)OCH_3$ at Different Levels of Theory[a]

| methods | $\Delta E^{\ddagger}$ | $\Delta E_{zpv}{}^{\ddagger}$ | $\Delta H^{\ddagger}$ | $\Delta G^{\ddagger}$ |
|---|---|---|---|---|
| HF/6−31G(d, p) | 66.89 | 65.61 | 64.20 | 67.96 |
| B3LYP/6−31G(d, p) | 44.03 | 42.45 | 41.05 | 44.82 |
| B3LYP/6−311+G(d, p) | 44.96 | 43.59 | 42.11 | 46.35 |
| MP2/6−31G(d, p) | 46.26 | 44.63 | 43.22 | 46.92 |
| MP2/6−311+G(d, p) | 45.66 | 44.25 | 42.75 | 46.95 |
| G2(MP2) | | 46.47 | 45.11 | 48.90 |

[a] In kcal/mol.

and Gibbs free energies in comparison with the G2(MP2) values and is poor in describing the aminolysis reactions of methylformate. The best agreement with those reference values comes from $\Delta E_{zpv}{}^{\ddagger}$, $\Delta H^{\ddagger}$, and $\Delta G^{\ddagger}$ values determined at the MP2/6−31G(d, p) level, the computed differences being less than 2.0 kcal/mol. The use of the larger basis set 6−311+G(d, p) in the MP2 approach slightly decreases the activation electronic energy and enthalpy. The results performed at the B3LYP level are quite close to those derived from the MP2 method, and the largest difference of 2.23 kcal/mol appears for the activation electronic energy $\Delta E^{\ddagger}$. The enlarging of the basis set from 6−31G(d, p) to 6−311+G(d, p) improves a bit of the performance of the B3LYP functional. The computed energy barrier with zero-point vibrational energy correction for this concerted step amounts to 43.59 and 46.47 kcal/mol at the B3LYP/6−311+G(d, p) and the G2(MP2) levels of theory, respectively. The corresponding Gibbs free energy changes are also similar, 46.35 kcal/mol for B3LYP/6−311+G(d, p) and 48.90 kcal/mol for G2(MP2), respectively. Therefore, it can be concluded that the B3LYP/6−311+G(d, p) level is suitable to study the title reaction with a good compromise between accuracy and computational cost.

**3.2. Structures of Stationary Points in the Gas Phase.** Here the concerted and stepwise pathways in each of the direct and ammonia-assisted mechanisms were considered for the aminolysis of the parent methylformate, similar to the cases in the previous theoretical study.[18] The calculated results show that these two cases are different in the attack manner, and the proton transfer is the main factor that influences the energy barrier. When the hydrogen atom H at the carbonyl C atom in methylformate was replaced by the $NH_2$ or $CF_3$ group, it was observed that the mechanism and all the critical structures of the reaction systems are similar to those when X = H for either concerted or neutral stepwise pathways. The main optimized geometrical parameters of all transition states along the reaction paths are given in Figure 1 for the reactions of $XC(O)OCH_3$ (X = $NH_2$, H, and $CF_3$) with the monomeric and dimeric ammonia molecules.

For the concerted pathway of the direct aminolysis of methylformate, the reaction involves only one step, in which the creation of the $C_{(1)}$-$N_{(4)}$ bond, the destruction of the $C_{(1)}$-$O_{(3)}$ bond, and the proton $H_{(5)}$ transfer from the ammonia toward $O_{(3)}$ occur in concert. The nucleophilic ammonia molecule attaches to the electrophilic carbon atom $C_{(1)}$, and a proton $H_{(5)}$ transfer from the ammonia molecule toward

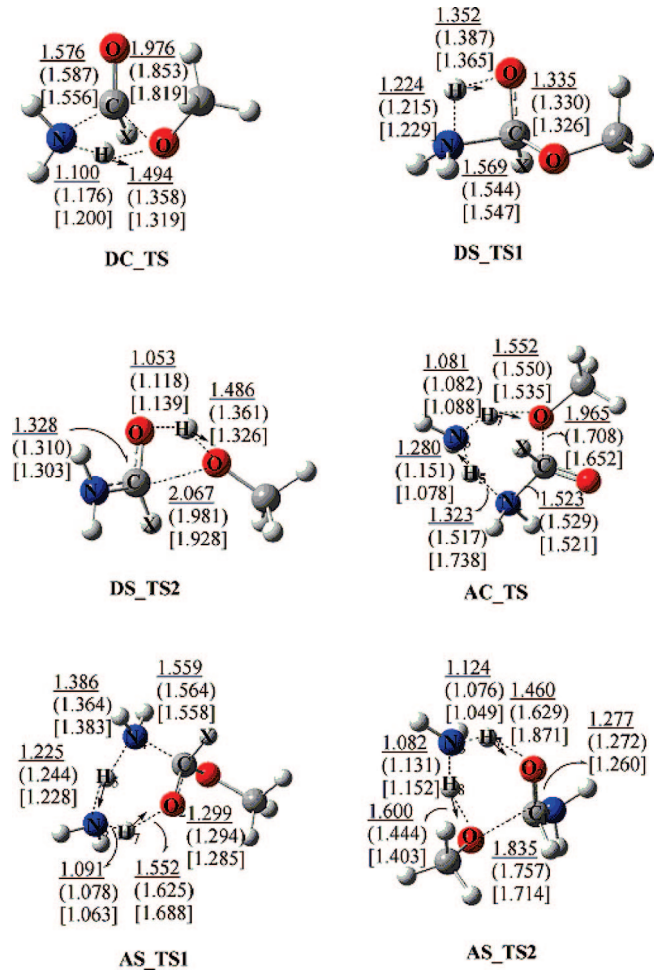



***Figure 1.*** Optimized transition structures along the concerted and stepwise pathways for the aminolysis of $XC(O)OCH_3$, (X = $NH_2$ (underlined), H (in parentheses), and $CF_3$ (in square brackets)). The arrows on the transition states indicate the reaction coordinate. (Bond lengths are in Å.)

the ester oxygen atom $O_{(3)}$ simultaneously occurs. The transition state (designated DC_TS) has a four-membered ring structure constituted by $C_{(1)}$, $O_{(3)}$, $H_{(5)}$, and $N_{(4)}$ atoms. IRC calculations in the reverse and forward directions from the transition state cause producing of the reactant complex (DC_RC) and the product complex (DC_PC). The stepwise pathway for the direct aminolysis of methylformate is mainly an addition/elimination mechanism. For both addition and elimination steps, proton transfers are involved to maintain neutrality in the tetrahedral intermediates formed. Calculated results showed this reaction begins with the addition of the $N_{(4)}$-$H_{(5)}$ bond to the $C_{(1)}$=$O_{(2)}$ double bond and consists of two transition states. The first transition state DS_TS1 has a four-membered ring consisting of $C_{(1)}$, $O_{(2)}$, $H_{(5)}$, and $N_{(4)}$ atoms and involves proton $H_{(5)}$ transfer from $N_{(4)}$ toward the carbonyl oxygen atom $O_{(2)}$. IRC calculations from the forward direction indicated that DS_TS1 converts to the stable intermediate DS_INT, in which the $C_{(1)}$-$N_{(4)}$ bond is formed, and proton $H_{(5)}$ is already transferred to $O_{(2)}$. The second step of the process is an elimination reaction, in which the $C_{(1)}$-$O_{(3)}$ ester single bond is broken, proton $H_5$ transfers from $O_{(2)}$ to $O_{(3)}$, and at the same time the $C_{(1)}$=$O_{(2)}$ bond is

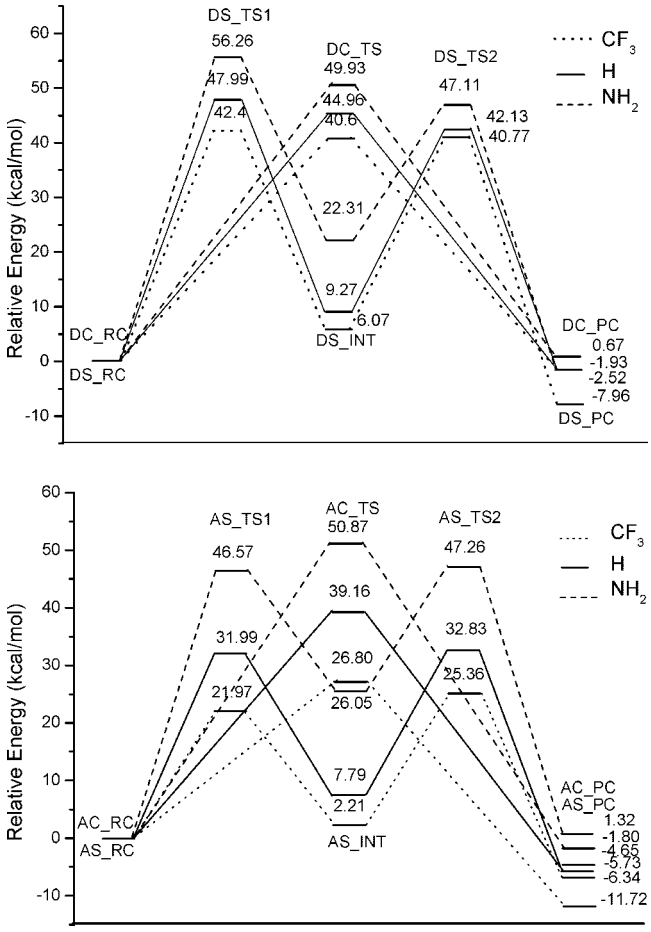



***Figure 2.*** Potential energy profiles for the aminolysis of different substituted methylformates along the concerted and stepwise mechanisms in the gas phase.

simultaneously restored. The transition state DS_TS2 also has a four-membered ring including $C_{(1)}$, $O_{(3)}$, $H_{(5)}$, and $O_{(2)}$ atoms.

The ammonia-assisted aminolysis reaction involves two ammonia molecules: the first ammonia molecule could be considered as the nucleophilic agent, while the second ammonia molecule acts as a catalytic role of facilitating the proton transfer. We also considered two mechanisms: the concerted and the stepwise pathways for this ammonia-assisted reaction of methylformate. From Figure 1, in the concerted process, the transition state AC_TS involves the simultaneous creation of the $C_{(1)}$-$N_{(4)}$ bond, the destruction of the $C_{(1)}$-$O_{(3)}$ bond, and the proton $H_{(5)}$ from ammonia as the nucleophilic agent toward the assisted-ammonia and the other proton $H_{(7)}$ transfer from the assisted-ammonia toward $O_{(3)}$. Similarly, the stepwise case in the ammonia-assisted reaction is coupled with proton transfer to keep neutrality in the hexahedral intermediates formed. As can be seen from Figure 1, the first transition state AS_TS1 contains a six-membered cycle constituted by $C_{(1)}$, $O_{(2)}$, $H_{(5)}$, $N_{(4)}$, $H_{(7)}$, and $N_{(6)}$ atoms and involves the proton $H_{(7)}$ transfer from $N_{(6)}$ toward the carbonyl $O_{(2)}$ and the other proton $H_{(5)}$ from ammonia as the nucleophilic reagent toward the assistant ammonia. Overcoming AS_TS1, the stable intermediate AS_INT is obtained, in which the $C_{(1)}$-$N_{(4)}$ bond is formed and the proton $H_{(7)}$ is already transferred to form a hydroxyl group. The second step of the process is similar to that of DS, associated with the breaking of the $C_{(1)}$-$O_{(3)}$ ester single bond and the simultaneous restoration of the $C_{(1)}$=$O_{(2)}$ bond after proton transfer. In the process of proton transfer, the transfer of $H_{(7)}$ from the hydroxyl group to the assistant ammonia and the transfer of another proton $H_{(8)}$ from the assistant ammonia to $O_{(3)}$ proceed in concert, in which the assistant ammonia acts as a role of proton-transfer catalysis. Compared with Sonia Ilieva's related work on the aminolysis of methylformate,[18] we adjusted the orientations of reaction complexes and transition states and located only one intermediate between AS_TS1 and AS_TS2.

**3.3. Energetics.** The computed relative electronic energies ($\Delta E$), corrected zero-point vibrational energies ($\Delta E_{ZPV}$), enthalpies ($\Delta H$), and Gibbs free energies ($\Delta G$) (relative to the reactant complex) for the fully optimized structures along the DC, DS, AC, and AS processes of aminolysis of $XC(O)OCH_3$ (X = $NH_2$, H, and $CF_3$) are given in the Supporting Information. The potential energy profiles are presented in Figure 2. The thermodynamic data were calculated using the B3LYP/6−311+G(d, p) method at 298.15 K and 1 atm.

In the direct aminolysis reaction, for the concerted pathway (DC), the activation energies increase in the following order: $CF_3$ (40.60 kcal/mol) < H (44.96 kcal/mol) < $NH_2$ (49.93 kcal/mol); while for DS, the activation energy changes agree with the order $CF_3$ (42.40 kcal/mol) < H (47.99 kcal/mol) < $NH_2$ (56.26 kcal/mol) for the addition step, and $CF_3$ (40.77 kcal/mol) < H (42.13 kcal/mol) < $NH_2$ (47.11 kcal/mol) for the elimination step, respectively. So the rate-determining step of the stepwise process is the addition reaction. Compared with the concerted mechanism, the activation energies of the stepwise mechanism are 6.33, 3.03, and 1.80 kcal/mol higher for X = $NH_2$, H, and $CF_3$, respectively. With regard to the parent compound, the electron-withdrawing group $CF_3$ lowers the energy barriers by 4.36 and 5.59 kcal/mol for the concerted and stepwise mechanism, respectively, whereas the electron-donating $NH_2$ group increases the energy barriers by 4.97 and 8.27 kcal/mol. It can be seen that B3LYP/6−311+G(d, p) calculations in the gas phase predict the concerted mechanism for all three aminolysis reactions to be more favorable than the stepwise pathway. These theoretical findings are in qualitative accord with the findings of Ilieva for the aminolysis of methylformate[18] and Yang and Drueckhammer for the aminolysis of methylthionacetate.[14] Latter authors showed that the gas-phase energies of the transition states for stepwise and concerted pathways are very close. More definite conclusions for the preferred mechanism may be made if the general base-catalyzed aminolysis process is considered.

In the ammonia-assisted aminolysis reaction, for the AC path, the activation energies increase in the following order: $CF_3$ (26.80 kcal/mol) < H (39.16 kcal/mol) < $NH_2$ (50.87 kcal/mol); while for AS, the activation energy changes agree with the order $CF_3$ (21.97 kcal/mol) < H (31.99 kcal/mol) < $NH_2$ (46.57 kcal/mol) for the addition step, and $CF_3$ (25.36 kcal/mol) < H (32.83 kcal/mol) < $NH_2$ (47.26 kcal/mol) for the elimination step, respectively. So the rate-determining step of the stepwise process is the elimination reaction.

***Table 2.*** Natural Charges ($q$) and Changes in Charges ($\Delta q$) for the Concerted and the Stepwise Pathways of the Aminolysis of Methylformate at the B3LYP/6-311+G(d, p) Level of Theory[a]

| | Direct Aminolysis | | | | |
|---|---|---|---|---|---|
| | $C_1$ | $O_2$ | $N_4$ | $H_5$ | $O_3$ |
| $q^{DC\_RC}$ | 0.650 | −0.612 | −1.069 | 0.349 | −0.537 |
| $\Delta q^{DC\_TS}$ | −0.087 | −0.036 | 0.208 | 0.125 | −0.216 |
| $\Delta q^{DC\_PC}$ | −0.128 | −0.011 | 0.248 | 0.119 | −0.214 |
| $q^{DS\_RC}$ | 0.649 | −0.612 | −1.069 | 0.370 | −5.373 |
| $\Delta q^{DS\_TS1}$ | −0.057 | −0.224 | 0.232 | 0.105 | 4.730 |
| $\Delta q^{DS\_INT}$ | −0.077 | −0.131 | 0.201 | 0.100 | 4.761 |
| $q^{DS\_INT}$ | 0.572 | −0.743 | −0.868 | 0.469 | −0.612 |
| $\Delta q^{DS\_TS2}$ | 0 | −0.001 | 0.110 | 0.041 | −0.175 |
| $\Delta q^{DS\_PC}$ | −0.042 | 0.107 | 0.073 | 0.013 | −0.157 |

| | Ammonia-Assisted Aminolysis | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $C_1$ | $O_3$ | $N_4$ | $H_5$ | $H_7$ | $N_6$ | $O_2$ | $H_8$ |
| $q^{AC\_RC}$ | 0.636 | −0.575 | −1.089 | 0.391 | 0.377 | −1.080 | −0.588 | 0.355 |
| $\Delta q^{AC\_TS}$ | −0.055 | −0.209 | 0.175 | 0.051 | 0.091 | 0.112 | −0.155 | 0.037 |
| $\Delta q^{AC\_PC}$ | −0.122 | −0.224 | 0.288 | 0.011 | 0.121 | −0.010 | −0.067 | 0.004 |
| $q^{AS\_RC}$ | 0.644 | −0.540 | −1.092 | 0.394 | 0.388 | −1.086 | −0.629 | 0.353 |
| $\Delta q^{AS\_TS1}$ | −0.055 | −0.129 | 0.222 | 0.040 | 0.065 | 0.076 | −0.233 | 0.035 |
| $\Delta q^{AS\_INT}$ | −0.066 | −0.080 | 0.213 | −0.004 | 0.113 | 0.004 | −0.145 | 0.010 |
| $q^{AS\_INT}$ | 0.578 | −0.620 | −0.879 | 0.390 | 0.500 | −1.082 | −0.773 | 0.363 |
| $\Delta q^{AS\_TS2}$ | −0.011 | −0.122 | −0.016 | 0.012 | −0.046 | 0.131 | −0.033 | 0.098 |
| $\Delta q^{AS\_PC}$ | −0.058 | −0.184 | 0.056 | −0.022 | −0.130 | 0.016 | 0.141 | 0.140 |

[a] In electronic charge units.

***Table 3.*** Activation Energies ($\Delta E^{\ddagger}$), Reaction Energies ($\Delta E^{o}$), Intrinsic Barrier ($\Delta E_o^{\ddagger}$), and Thermodynamic Contributions ($\Delta E_{thermo}^{\ddagger}$) as well as Relative Values (in Parentheses) to the Parent Compound for the Concerted Pathway and the Addition/Elimination Steps of the Stepwise Pathway[a]

| | | X | $\Delta E^{\ddagger}$ | $\Delta E^{o}$ | $\Delta E_o^{\ddagger}$ | $\Delta E_{thermo}^{\ddagger}$ |
|---|---|---|---|---|---|---|
| | | | Direct Aminolysis Reaction | | | |
| concerted | | $NH_2$ | 49.93(4.97) | 0.67(2.33) | 49.59(3.80) | 0.34(1.17) |
| | | H | 44.96(0.00) | −1.66(0.00) | 45.79(0.00) | −0.83(0.00) |
| | | $CF_3$ | 40.60(−4.36) | −1.42(0.24) | 41.31(−4.48) | −0.71(0.12) |
| stepwise | I[b] | $NH_2$ | 56.26(8.27) | 22.31(13.04) | 44.40(1.17) | 11.86(7.10) |
| | | H | 47.99(0.00) | 9.27(0.00) | 43.23(0.00) | 4.76(0.00) |
| | | $CF_3$ | 42.40(−5.59) | 6.07(−3.20) | 39.31(−3.92) | 3.09(−1.67) |
| | II[c] | $NH_2$ | 24.80(−8.06) | −24.83(−13.63) | 36.15(−2.11) | −11.35(−5.95) |
| | | H | 32.86(0.00) | −11.2(0.00) | 38.26(0.00) | −5.40(0.00) |
| | | $CF_3$ | 34.70(1.84) | −14.03(−2.83) | 41.42(3.16) | −6.72(−1.32) |
| | | | Ammonia-Assisted Aminolysis Reaction | | | |
| concerted | | $NH_2$ | 50.87(11.71) | −1.80(3.93) | 51.77(9.79) | −0.90(1.92) |
| | | H | 39.16(0.00) | −5.73(0.00) | 41.98(0.00) | −2.82(0.00) |
| | | $CF_3$ | 26.80(−12.36) | −11.72(−5.99) | 32.39(−9.59) | −5.59(−2.77) |
| stepwise | I[b] | $NH_2$ | 46.57(14.58) | 26.05(18.26) | 32.23(4.27) | 14.34(10.31) |
| | | H | 31.99(0.00) | 7.79(0.00) | 27.96(0.00) | 4.03(0.00) |
| | | $CF_3$ | 21.97(−10.02) | 2.21(−5.58) | 20.85(−7.11) | 1.12(−2.91) |
| | II[c] | $NH_2$ | 21.21(−3.83) | −24.73(−12.29) | 32.40(1.45) | −11.19(−5.28) |
| | | H | 25.04(0.00) | −12.44(0.00) | 30.95(0.00) | −5.91(0.00) |
| | | $CF_3$ | 23.15(−1.89) | −8.55(3.89) | 27.26(−3.69) | −4.11(1.80) |

[a] In kcal/mol. [b] Addition step. [c] Elimination step.

Compared with the concerted mechanism, the activation energies of the stepwise mechanism are 3.61, 6.33, and 1.44 kcal/mol lower for X = $NH_2$, H, and $CF_3$, respectively. With regard to the parent compound, the electron-withdrawing group $CF_3$ substantially lowers the energy barriers by 12.36 and 7.47 kcal/mol for the concerted and stepwise mechanism, respectively, whereas the electron-donating $NH_2$ group increases the energy barriers by 11.71 and 14.43 kcal/mol. It can be seen that B3LYP/6−311+G(d, p) calculations in the gas phase for the ammonia-assisted aminolysis reaction predict the stepwise mechanism to be more favorable than the concerted pathway. In the structure AS_TS2, the second ammonia acting as the catalysis role facilities the proton-transfer between $O_{(2)}$ and $O_{(3)}$ and accordingly greatly lowers the activation energy. Compared with the concerted pathway of the direct aminolysis reaction, the stepwise progress of the ammonia-assisted aminolysis is 2.67, 12.13, and 15.24 kcal/mol more favored pathway for X = $NH_2$, H, and $CF_3$, respectively.

The atomic and group charges and their changes by the NBO method for the concerted pathway and addition/elimination steps of the stepwise pathway in the aminolysis of methylformate are shown in Table 2. For the direct aminolysis reaction, the DC_RC and DS_RC have positive

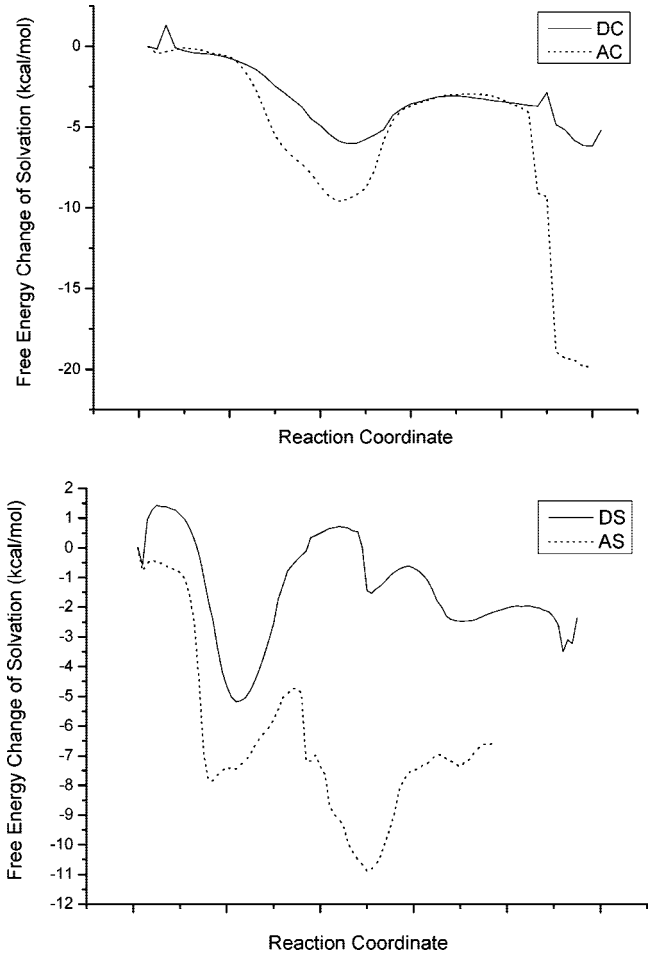



***Figure 3.*** Changes in the free energies of solvation along the reaction coordinate calculated from Monte Carlo simulations.

charges on the $C_{(1)}$ group and negative charges on the $O_{(3)}$ group. In DC_TS and DS_TS1, the positive charges on the $C_{(1)}$ group decrease, while for the $O_{(3)}$ group, it becomes more negative in DC_TS and DS_TS1, suggesting that the rate should increase when the electron-withdrawing group $CF_3$ is the substituent. Therefore, the group $CF_3$ facilitates the reaction and is much more favorable for the aminolysis reaction. Accordingly, we can get the same conclusion when analyzing the ammonia-assisted aminolysis reaction.

**3.4. Substituent Effects.** To analyze the substituent effects on the activation energies of the aminlysis reaction, it is necessary to understand how substituents alter the energies. One useful way is to use the Marcus theory,[50] which was used to analyze the activation energies of reaction pathways.[51] The Marcus equation is given by

$$\Delta G^{\ddagger} = \Delta G_o^{\ddagger} + {}^1\!/_2\Delta G^{\circ} + (\Delta G^{\circ})^2/(16\Delta G_o^{\ddagger}) \qquad (3)$$

where $\Delta G_o^{\ddagger}$ is called the intrinsic barrier representing the barrier of a thermoneutral reaction. $\Delta G^{\ddagger}$ and $\Delta G^{\circ}$ are Gibbs free energy changes of activation and reaction of a nondegenerate reaction, respectively. The third term $(\Delta G^{\circ})^2/(16\Delta G_o^{\ddagger})$ is the correction factor for nonadditivity of the intrinsic and thermodynamic effects. Murdoch advocated that the similar expression in eq 4 can be applied to the nondegenerate reaction and used for some pericyclic reactions and other chemical processes[52,53]

$$\Delta E^{\ddagger} = \Delta E_o^{\ddagger} + {}^1\!/_2\Delta E^{\circ} + (\Delta E^{\circ})^2/(16\Delta E_o^{\ddagger}) \qquad (4)$$

***Table 4.*** Changes in the Free Energies of Solvation ($\Delta G_{sol}$) and Total Free Energy Differences ($\Delta G_{total}$) in the Gas Phase and the Solvent $CH_3CN$ for DC, DS, AC, and AS Pathways of Aminolysis of Methylformate[a]

| | | | gas phase | $CH_3CN$ |
|---|---|---|---|---|
| DC | | $\Delta G_{sol}^{\ddagger}$ | – | −5.50 |
| | | $\Delta G_{total}^{\ddagger}$ | 46.35 | 40.85 |
| | | $\Delta G_{sol}^{o}$ | – | −5.19 |
| | | $\Delta G_{total}^{o}$ | −2.29 | −7.48 |
| DS | I[b] | $\Delta G_{sol}^{\ddagger}$ | – | −3.12 |
| | | $\Delta G_{total}^{\ddagger}$ | 50.37 | 47.25 |
| | | $\Delta G_{sol}^{o}$ | – | 0.02 |
| | | $\Delta G_{total}^{o}$ | 14.91 | 14.93 |
| | II[c] | $\Delta G_{sol}^{\ddagger}$ | – | −2.45 |
| | | $\Delta G_{total}^{\ddagger}$ | 42.76 | 40.31 |
| | | $\Delta G_{sol}^{o}$ | – | −2.37 |
| | | $\Delta G_{total}^{o}$ | −2.84 | −5.21 |
| AC | | $\Delta G_{sol}^{\ddagger}$ | – | −9.19 |
| | | $\Delta G_{total}^{\ddagger}$ | 42.72 | 33.53 |
| | | $\Delta G_{sol}^{o}$ | – | −19.84 |
| | | $\Delta G_{total}^{o}$ | −5.01 | −24.85 |
| AS | I[b] | $\Delta G_{sol}^{\ddagger}$ | – | −7.44 |
| | | $\Delta G_{total}^{\ddagger}$ | 35.90 | 28.46 |
| | | $\Delta G_{sol}^{o}$ | – | −7.16 |
| | | $\Delta G_{total}^{o}$ | 13.42 | 6.26 |
| | II[c] | $\Delta G_{sol}^{\ddagger}$ | – | −9.43 |
| | | $\Delta G_{total}^{\ddagger}$ | 36.86 | 27.43 |
| | | $\Delta G_{sol}^{o}$ | – | −6.60 |
| | | $\Delta G_{total}^{o}$ | −6.67 | −13.27 |

[a] In kcal/mol. [b] Addition step. [c] Elimination step.

In this study, we used eq 4 to separate the intrinsic and thermodynamic contributions of the substituent effects on the activation energies of the aminolysis of $XC(O)OCH_3$ reactions for the concerted mechanism and addition/elimination processes in the stepwise pathway. The intrinsic barrier $\Delta E_o^{\ddagger}$ was calculated using eq 4 with the quantum mechanically calculated values of activation energy $\Delta E^{\ddagger}$ and reaction energy $\Delta E^{\circ}$. Table 3 shows the intrinsic barriers and the thermodynamic contributions for the title reaction systems. For the direct aminolysis reaction, the intrinsic activation energies of the concerted mechanism and the addition step of the stepwise pathway of the parent system are 44.96, 47.99, and 32.86 kcal/mol, respectively. For the concerted mechanism, when X = $CF_3$, the 4.36 kcal/mol decrease in the activation is mainly due to the intrinsic factor, which induces a 4.48 kcal/mol decrease, and with only a 0.12 kcal/mol increase in exothermicity by the group. When X = $NH_2$, the relative high contributions of intrinsic (3.80 kcal/mol) and thermodynamic (1.17 kcal/mol) lead to a small increase in the activation energy. For the addition/elimination steps of the stepwise pathway, the intrinsic contribution is also a dominate factor in the activation energy reductions in the X = $CF_3$ reaction system, while for X = $NH_2$, the relative high contribution of the thermodynamic (7.10 kcal/mol) and the minor contribution of intrinsic (1.17 kcal/mol) lead to a considerable increase in the activation energy. For the ammonia-assisted aminolysis reaction, the intrinsic activation energies of the concerted mechanism and the addition/elimination steps of the stepwise pathway of the parent system are 39.16, 31.99, and 25.04 kcal/mol, respectively.

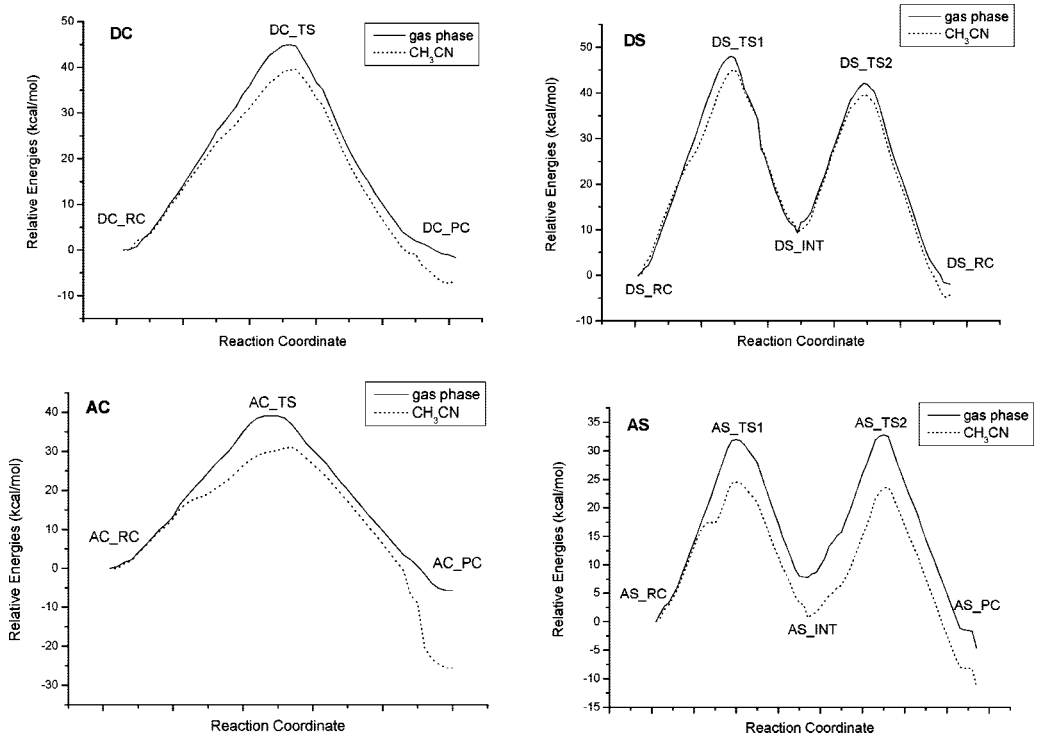



***Figure 4.*** Minimum energy path $E_{MEP(s)}$ for the aminolysis of $HC(O)OCH_3$ in the gas phase and in the solvent $CH_3CN$.

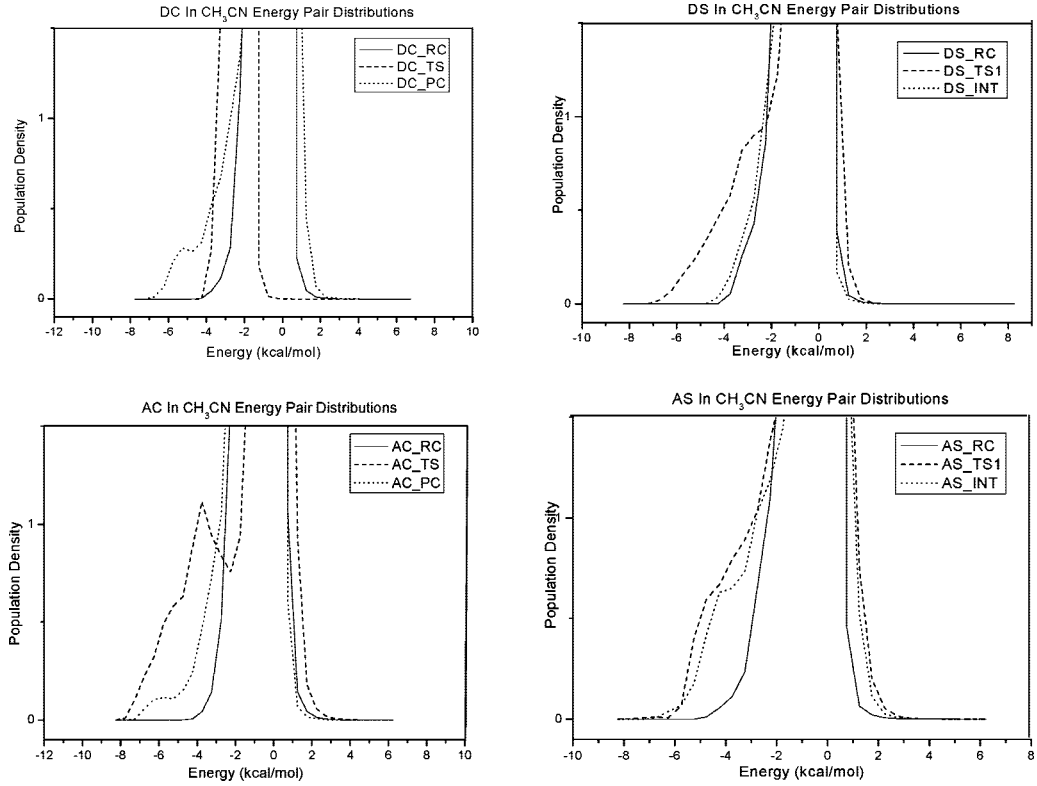



***Figure 5.*** Energy pair distribution of solute−solvent interaction. The ordinate gives the number of solvent molecules coordinated with the solute with the interaction energy shown on the abscissa. The units for the *y*-axis are the number of molecules per kilocalorie per mole.

For the concerted mechanism, when $X = CF_3$, the 12.36 kcal/mol decrease in the activation is mainly due to the intrinsic factor, which induces a 9.59 kcal/mol decrease, and with only a 2.77 kcal/mol increase in exothermicity by the group. When $X = NH_2$, the relative high contributions of intrinsic (9.79 kcal/mol) and thermodynamic (1.92 kcal/mol) lead to a 11.71 kcal/mol increase in the activation energy. For the elimination step of the stepwise pathway, the intrinsic contribution is also a dominate factor in the activation energy reductions in the $X = CF_3$ reaction system, while for $X =$

$NH_2$, the relative high contribution of thermodynamic (−5.28 kcal/mol) and minor contribution of intrinsic (1.45 kcal/mol) lead to a decrease in the activation energy but with a high activation energy of 14.58 kcal/mol for the addition step of the stepwise pathway.

**3.5. Solvent Effects Determined by Monte Carlo Simulation.** The aminolysis of methylformate was studied in $CH_3CN$ using the free energy perturbation method implemented in BOSS 4.2. The theoretical results confirm the conclusion made on the basis of calculations for the gas-phase process. Figure 3 displays the changes in free energies of solvation over the course of the reaction in the solvent $CH_3CN$. The changes in the free energies of solvation and free energy changes in the gas phase and in solution for the activation and DC, DS, AC, and AS reaction procedures are listed in Table 4. For DC, the difference in free energies of solvation between DC_RC and DC_TS in $CH_3CN$ is −5.50 kcal/mol, which indicates that the transition state DC_TS is stabilized in $CH_3CN$ by solvation compared with DC_RC. While for DS, the transition state DS_TS1 is stabilized more by solvation than the reactant complex DS_RC in $CH_3CN$ and has 6.36 kcal/mol smaller free energies of solvation than DS_RC. It can be viewed from Table 4 that free energies of solvation of ammonia-assisted aminolysis reactions are considerably higher than those of the direct aminolysis reactions. For the solution, the calculated free energies of activation of the direct aminolysis by combining the DFT calculation (B3LYP/6−311+G(d, p) level) with the Monte Carlo simulation are 40.85, 47.25, 33.53, and 28.46 kcal/mol for DC, DS, AC, and AS in solvent $CH_3CN$. Figure 4 depicts the relative potential energy profiles $E_{MEP(s)}$ of four systems along the minimum energy path (MEP) in the gas phase and $CH_3CN$. From Table 4 and Figure 4, the solvent effects on the ammonia-assisted aminolysis of the concerted and stepwise pathways by $CH_3CN$ are computed to be more favorable than those on the direct aminolysis of the two pathways. Thus, the calculations predict that a stepwise pathway with a $NH_3$ molecule as a catalyst in the solution is the preferred mechanism. Here comes a question: what factors are responsible for the solvent effects of $CH_3CN$. We consider the development of the electronic charge distribution on going from the reactant complex to the transition state. From Table 2, taking the direct aminolysis as an example, the reactant complex has a positive charge on atoms $C_{(1)}$ and $H_{(5)}$ and a negative charge on atoms $O_{(2)}$, $N_{(4)}$, and $O_{(3)}$. On going to the transition state DC_TS, when atoms $C_{(1)}$ and $H_{(5)}$ have net gains of electrons, atoms $O_{(2)}$, $N_{(4)}$, and $O_{(3)}$ have net losses of electrons. As a result, from DC_RC to DC_TS, the magnitude of charge on each of the atoms involved in the reaction center reduces, and the electronic charge distribution becomes more disperse, thus decreasing the electrostatic interaction with solvents. In contrast, for the ammonia-assisted process, there are net losses of electrons on atoms $C_{(1)}$ and $H_{(7)}$ and net gains of electrons on atoms $N_{(4)}$. On going from DS_RC to DS_TS1, the charges on atoms $C_{(1)}$ and $H_{(5)}$ become more positive, and the charges on atom $N_{(4)}$ become more negative. Such buildup of the electronic charges enhances the electrostatic interaction with solvents. The same conclusion has also been found in the stepwise pathway of both the direct transfer and ammonia-assisted aminolysis reactions.

Figure 5 shows the solute−solvent energy pair distributions for the DC, DS, AC, and AS processes. The plots give the number of solvent molecules on the ordinate that interact with the solute with the interaction energy shown on the abscissa. In the solution, the spikes centered at 0.0 kcal/mol result from the weak interactions between the solute and many distant $CH_3CN$ molecules. In DC, from Figure 5(DC), for DC_RC in $CH_3CN$, there is a bound group of solvent molecules, which forms a band from ca. −10.0 to −2.25 kcal/mol. Integration of the distribution curve for DC_RC up to the end of this band at −2.25 kcal/mol defines 1.557 solvent molecules that interact with $CH_3CN$. Integration of the curves for DC_TS and DC_PC until this limit results in 10.433 and 4.657 $CH_3CN$ molecules. In DS [see Figure 5(DS)], integration of the curves for DS_RC, DS_TS1, and DS_INT up to the end of this band at −2.25 kcal/mol defines 1.603, 4.532, and 2.185 $CH_3CN$ molecules. In the AC system, integration of the curves up to the end of the plateaus at −2.25 kcal/mol reveals 2.333, 6.898, and 5.149 $CH_3CN$ molecules for AC_RC, AC_TS, and AC_PC in $CH_3CN$, as shown in Figure 5(AC). For the AS system [see Figure 5(AS)], the number of $CH_3CN$ molecules that interact with solute is 2.181, 5.878, and 4.950 for AS_RC, AS_TS1, and AS_INT. These differences imply that the stabilization of the transition structure relative to its reactant complex is different due to the effect of $CH_3CN$. The better solvation of the translate state in the ammonia-assisted path can be attributed to an increase in the interaction between solute and solvent molecules, while in the direct aminolysis process the decrease for the interactions with the solvent along the reaction path is responsible for the destabilization of translate state. The catalytic role of the second ammonia molecule affects mostly the proton-transfer processes. Thus six-member rings formed in the transition state structures for the catalyzed processes are more stable than four-member rings in the case of the uncatalyzed aminolysis reactions. This explains the lower energy barriers along the reaction path of the catalyzed process in the gas phase as well as in the presence of solvent $CH_3CN$.

## 4. Conclusion

The aminolysis of $XC(O)OCH_3$ (X = H, $NH_2$, and $CF_3$) was studied using the B3LYP/6−311+G(d, p) level of theory in the gas phase. The solvent effects of $CH_3CN$ on the aminolysis of $HC(O)OCH_3$ were calculated by Monte Carlo simulations. The $NH_3$ catalysis role of the nucleophilic was investigated in detail. The results show that the most favorable pathway of the reaction is through the general base catalyzed neutral stepwise mechanism. The structure and transition vectors of the transition states indicate that the catalytic role of ammonia is realized by facilitating the proton transfer processes. The calculated values correctly reflect that the effect of $CH_3CN$ is more favorable for the ammonia-assisted aminolysis than for the direct aminolysis. The calculated results indicate that the ammonia-assisted pathway is energetically preferred to the direct process in the gas

phase, and the electron-drawing group $CF_3$ facilitates all the aminolysis reaction processes.

**Acknowledgment.** This project has been supported by the National Natural Science Foundation of China (Grant Nos. 20473055 and 20773089) and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (Grant No. 20071108-18-15).

**Supporting Information Available:** Listings of Cartesian coordinates and energies in hartrees at the B3LYP/6−311+G(d, p) level of theory. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Blackburn, G. M.; Jencks, W. P. The mechanism of the aminolysis of methyl formate. *J. Am. Chem. Soc.* **1968**, *90*, 2638.

(2) Jencks, W. P. *Catalysis in Chemistry and Enzymology*; McGraw Hill: New York, 1969.

(3) Jencks, W. P.; Carriuolo, J. General Base Catalysis of the Aminolysis of Phenyl Acetate. *J. Am. Chem. Soc.* **1960**, *82*, 675.

(4) Jencks, W. P.; Gilchrist, M. General Base Catalysis of the Aminolysis of Phenyl Acetate by Primary Alkylamines. *J. Am. Chem. Soc.* **1966**, *88*, 104.

(5) Bruice, T. C.; Donzel, A.; Huffman, R. W.; Butler, A. R. Aminolysis of Phenyl Acetates in Aqueous Solutions. Observations on the Influence of Salts, Amine Structure, and Base Strength. *J. Am. Chem. Soc.* **1967**, *89*, 2106.

(6) Rogers, G. A.; Bruice, T. C. Isolation of a tetrahedral intermediate in an acetyl transfer reaction. *J. Am. Chem. Soc.* **1973**, *95*, 4452.

(7) Rogers, G. A.; Bruice, T. C. Synthesis and evaluation of a model for the so-called charge-relay system of the serine esterases. *J. Am. Chem. Soc.* **1974**, *96*, 2473.

(8) Bruice, T. C.; Benkovic, S. J. Acyl Transfer Reactions Involving Carboxylic Acid Esters and Amides. *Bioorganic Mechanisms*, *Vol. 1*; W. A. Benjamin, Inc.: New York, 1966; Chapter 1.

(9) Bunnett, J. F.; Davis, G. T. The Mechanism of Aminolysis of Esters. *J. Am. Chem. Soc.* **1960**, *82*, 665.

(10) Gresser, M. J.; Jencks, W. P. Ester aminolysis. Structure-reactivity relationships and the rate-determining step in the aminolysis of substituted diphenyl carbonates. *J. Am. Chem. Soc.* **1977**, *99*, 6963.

(11) Williams, A. Concerted mechanisms of acyl group transfer reactions in solution. *Acc. Chem. Res.* **1989**, *22*, 387.

(12) Zipse, H.; Wang, L.; Houk, K. N. Polyether catalysis of ester aminolysis-A computational and experimental study. *Liebigs Ann.* **1996**, *10*, 1511.

(13) Wang, L.; Zipse, H. Bifunctional catalysis of ester aminolysis-A computational and experimental study. *Liebigs Ann.* **1996**, *10*, 1501.

(14) Yang, W.; Drueckhammer, D. G. Computational Studies of the Aminolysis of Oxoesters and Thioesters in Aqueous Solution. *Org. Lett.* **2000**, *2*, 4133.

(15) Oie, T.; Loew, G. H.; Burt, S. K.; Binkley, J. S.; McElroy, R. D. Quantum chemical studies of a model for peptide bond formation: formation of formamide and water from ammonia and formic acid. *J. Am. Chem. Soc.* **1982**, *104*, 6169.

(16) Gorb, L.; Asensio, A.; Tuñón, I.; Ruiz-López, M. F. The Mechanism of Formamide Hydrolysis in Water from *Ab Initio* Calculations and Simulations. *Chem. Eur. J.* **2005**, *11*, 6743.

(17) Chalmet, S.; Harb, W.; Ruiz-Lopez, M. F. Computer Simulation of Amide Bond Formation in Aqueous Solution. *J. Phys. Chem. A* **2001**, *105*, 11574.

(18) Ilieva, S.; Galabov, B.; Schaefer, H. F. Computational Study of the Aminolysis of Esters. The Reaction of Methylformate with Ammonia. *J. Org. Chem.* **2003**, *68*, 1496.

(19) Singleton, D. A.; Merrigan, S. R. Resolution of Conflicting Mechanistic Observations in Ester Aminolysis. A Warning on the Qualitative Prediction of Isotope Effects for Reactive Intermediates. *J. Am. Chem. Soc.* **2000**, *122*, 11035.

(20) Sung, D. D.; Koo, I. S.; Yang, K.; Lee, I. DFT studies on the structure and stability of zwitterionic tetrahedral intermediate in the aminolysis of esters. *Chem. Phys. Lett.* **2006**, *426*, 280.

(21) Lee, I.; Sung, D. D. Theoretical and physical aspects of stepwise mechanisms in acyl-transfer reactions. *Curr. Org. Chem.* **2004**, *8*, 557.

(22) Gresser, M. J.; Jencks, W. P. Ester aminolysis. Structure-reactivity relationships and the rate-determining step in the aminolysis of substituted diphenyl carbonates. *J. Am. Chem. Soc.* **1977**, *99*, 6963.

(23) Castro, E. A.; Stander, C. L. Nonlinear Broensted-type plot in the pyridinolysis of 2, 4-dinitrophenyl benzoate in aqueous ethanol. *J. Org. Chem.* **1985**, *50*, 3595.

(24) Castro, E. A.; Valdiva, J. L. Linear free-energy relationship in the pyridinolysis of 2, 4-dinitrophenyl p-chlorobenzoate in aqueous ethanol solution. *J. Org. Chem.* **1986**, *51*, 1668.

(25) Castro, E. A.; Steinfort, G. B. Kinetics and mechanism of the pyridinolysis of 2, 4-dinitrophenyl *p*-nitrobenzoate. *J. Chem. Soc., Perkin Trans. 2* **1983**, *2*, 453.

(26) Castro, E. A.; Leandro, L.; Quesieh, N.; Santos, J. G. Kinetics and Mechanisms of the Reactions of 3-Methoxyphenyl, 3-Chlorophenyl, and 4-Cyanophenyl 4-Nitrophenyl Thionocarbonates with Alicyclic Amines. *J. Org. Chem.* **2001**, *66*, 6130.

(27) Castro, E. A.; Galvez, A.; Leandro, L.; Santos, J. G. Kinetic and Mechanistic Investigation of the Aminolysis of 3-Methoxyphenyl 3-Nitrophenyl Thionocarbonate, 3-Chlorophenyl 3-Nitrophenyl Thionocarbonate, and Bis (3-nitrophenyl) Thionocarbonate. *J. Org. Chem.* **2002**, *67*, 4309.

(28) Um, I. H.; Lee, S. E.; Kwon, H. J. Effect of Amine Nature on Reaction Mechanism: Aminolyses of *O*-4-Nitrophenyl Thionobenzoate with Primary and Secondary Amines. *J. Org. Chem.* **2002**, *67*, 8999.

(29) Oh, H. K.; Kim, S. K.; Cho, I. H.; Lee, H. W.; Lee, I. Kinetics and mechanism of the aminolysis of aryl phenyldithioacetates in acetonitrile. *J. Chem. Soc., Perkin Trans. 2* **2000**, *2*, 2306.

(30) Antonczak, S.; Ruiz-Lopez, M. F.; Rivail, J. L. Ab Initio Analysis of Water-Assisted Reaction Mechanisms in Amide Hydrolysis. *J. Am. Chem. Soc.* **1994**, *116*, 3912.

(31) Díaz, N.; Suárez, D.; Sordo, T. L.; Merz, K. M., Jr. A Theoretical Study of the Aminolysis Reaction of Lysine 199 of Human Serum Albumin with Benzylpenicillin: Consequences for Immunochemistry of Penicillins. *J. Am. Chem. Soc.* **2001**, *123*, 7574.

(32) Díaz, N.; Suárez, D.; Sordo, T. L. Theoretical Study of the Water-Assisted Aminolysis of *â*-Lactams: Implications for the Reaction between Human Serum Albumin and Penicillins. *J. Am. Chem. Soc.* **2000**, *122*, 6710.

(33) Díaz, N.; Suárez, D.; Sordo, T. L. Importance of a Synperiplanar Stepwise Mechanism through Neutral Intermediates in the Aminolysis of Monocyclic *â*-Lactams: A Theoretical Analysis. *J. Org. Chem.* **1999**, *64*, 9144.

(34) Díaz, N.; Suárez, D.; Sordo, T. L. $NH_3$-Assisted Ammonolysis of *â*-Lactams: A Theoretical Study. *J. Org. Chem.* **1999**, *64*, 3281.

(35) Díaz, N.; Suárez, D.; Sordo, T. L. Ammonolysis and Aminolysis of b-Lactams: A Theoretical Study. *Chem. Eur. J.* **1999**, *5*, 1045.

(36) Díaz, N.; Suárez, D.; Sordo, T. L.; Me'ndez, R.; Villacorta, J. M. A Combined Theoretical and Experimental Research Project into the Aminolysis of $\beta$-Lactam Antibiotics: The Importance of Bifunctional Catalysis. *Eur. J. Org. Chem.* **2003**, 4161.

(37) Díaz, N.; Suárez, D.; Sordo, T. L. Theoretical Study of Amine-Assisted Aminolysis of Penicillins - The Kinetic Role of the Carboxylate Group. *Eur. J. Org. Chem.* **2001**, 793.

(38) Díaz, N.; Suárez, D.; Sordo, T. L. Theoretical Study of Ammonolysis of Monobactams: Kinetic Role of the N-Sulfonate Group. *Helv. Chim. Acta* **2002**, *85*, 206.

(39) Ilieva, S.; Galabov, B.; Musaev, D. G.; Morokuma, K. Computational Study of the Aminolysis of 2-Benzoxazolinone. *J. Org. Chem.* **2003**, *68*, 3406.

(40) Jin, L.; Wu, Y.; Xue, Y.; Guo, Y.; Xie, D. Q.; Yan, G. S. Theoretical Studies on the Aminolysis of Phenyl Formate. Mechanism and Solvent Effect. *Acta Chim. Sin.* **2006**, *64*, 873.

(41) Yi, G. Q.; Zeng, Y.; Xia, X. F.; Xue, Y.; Kim, C. K.; Yan, G. S. The substituent effects of the leaving groups on the aminolysis of phenyl acetates: DFT studies. *Chem. Phys.* **2008**, *345*, 73.

(42) Xue, Y.; Kim, C. K. Effects of Substituents and Solvents on the Reactions of Iminophosphorane with Formaldehyde: *Ab Initio* MO Calculation and Monte Carlo Simulation. *J. Phys. Chem. A* **2003**, *107*, 7945.

(43) Xue, Y.; Kim, C. K.; Guo, Y.; Xie, D. Q.; Yan, G. S. DFT study and Monte Carlo simulation on proton transfers of 2-amino-2-oxazoline, 2-amino-2-thiazoline, and 2-amino-2-imidazoline in the gas phase and in water. *J. Comput. Chem.* **2005**, *26*, 994.

(44) Wu, Y.; Xue, Y.; Xie, D. Q.; Kim, C. K.; Yan, G. S. Theoretical Studies on the Hydrolysis Mechanism of N-(2-oxo-1,2-dihydro-pyrimidinyl) Formamide. *J. Phys. Chem. B* **2007**, *111*, 2357.

(45) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 03, Revision D. 01*; Gaussian, Inc.: Pittsburgh, PA, 2005.

(46) Fukui, K. Formulation of the reaction coordinate. *J. Phys. Chem.* **1970**, *74*, 4161.

(47) Reed, A. E.; Curtiss, L. A.; Weinhold, F. Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint. *Chem. Rev.* **1988**, *88*, 899.

(48) Jorgensen, W. L.; Ravimohan, C. Monte Carlo simulation of differences in free energies of hydration. *J. Chem. Phys.* **1985**, *83*, 3050.

(49) Jorgensen, W. L.; Tirado-Rives, J. *J. Comput. Chem.* **2005**, *26*, 1689–1700.

(50) Marcus, R. A. On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer. *J. Chem. Phys.* **1956**, *24*, 966.

(51) Yoo, H. Y.; Houk, K. N. Theory of Substituent Effects on Pericyclic Reaction Rates: Alkoxy Substituents in the Claisen Rearrangement. *J. Am. Chem. Soc.* **1997**, *119*, 2877.

(52) Murdoch, J. R. Relationship between More O'Ferrall plots and Marcus rate theory. Overriding orbital-symmetry constraints on chemical reactions. *J. Am. Chem. Soc.* **1983**, *105*, 2660.

(53) Murdoch, J. R. A simple relationship between empirical theories for predicting barrier heights of electron-, proton-, atom-, and group-transfer reactions. *J. Am. Chem. Soc.* **1983**, *105*, 2159.

CT800099A

JCTC Journal of Chemical Theory and Computation

# Mechanism for the Substitution of an Aqua Ligand of $UO_2(OH_2)_5^{2+}$ by Chloride

François P. Rotzinger*

*Institut des Sciences et Ingénierie Chimiques (ISIC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Station 6, CH-1015 Lausanne, Switzerland*

Received April 16, 2008

**Abstract:** Geometry and energy of the reactant ($UO_2(OH_2)_5 \cdot Cl^+$), the transition state ($UO_2(OH_2)_5 \cdots Cl^{+\ \ddagger}$), and the product ($UO_2Cl(OH_2)_4 \cdot OH_2^+$) of the title reaction have been computed with complete active space SCF (geometries and vibrational frequencies) and multiconfiguration quasi-degenerate second-order perturbation theory (total energies). Hydration was treated using the polarizable continuum model. The two investigated active spaces, (12/11) and (12/12), produce the same results. In contrast to the water exchange reaction on $UO_2(OH_2)_5^{2+}$, which proceeds via the associative (A) mechanism (which is a two step reaction involving an intermediate with an increased coordination number, $UO_2(OH_2)_6^{2+}$), water substitution by chloride follows the associative interchange ($I_a$) mechanism (which does not proceed via any intermediate). In this case, structure and imaginary mode of the transition state are not straightforward criteria for the attribution of the substitution mechanism, since they are both typical for the A pathway. The $I_a$ mechanism was derived from the computed intrinsic reaction coordinate, which showed that no intermediate (for example $UO_2Cl(OH_2)_5^+$) exists as a local minimum on the potential energy surface. The activation free enthalpy is 31 kJ mol$^{-1}$. As for the water exchange reaction, the dissociative mechanism is unlikely to operate because of its higher free activation enthalpy (by ≈25 kJ mol$^{-1}$).

## Introduction

The activation enthalpy ($\Delta H^{\ddagger}$) and free enthalpy ($\Delta G^{\ddagger}$) of the water exchange reaction 1 on the uranyl(VI) aqua ion have been measured with variable-temperature $^{17}O$ NMR techniques.[1]

$$UO_2(OH_2)_5^{2+} + H_2\mathbf{O} \rightarrow UO_2(\mathbf{O}H_2)(OH_2)_4^{2+} + H_2O \quad (1)$$

Substitution mechanisms are classified[2] as associative (A), dissociative (D), or concerted (I), whereby the concerted mechanism might have associative or dissociative character, which is denoted as $I_a$ or $I_d$, respectively. The A and the D mechanisms proceed via intermediates with an increased or reduced coordination number, whereas the concerted pathways ($I_a$, I, or $I_d$) do not involve any intermediate. According to recent ab initio and DFT computations, the associative (A) mechanism is favored over the dissociative (D) mechanism by a rather modest energy of ≈15−25 kJ mol$^{-1}$.[3,4] Criteria for the distinction of the concerted ($I_a$) from the stepwise (A) substitution mechanism were presented.[3] The lifetime of the intermediate $UO_2(OH_2)_6^{2+}$ was estimated as ≈1−6 ps.[3]

In this article, the substitution of an aqua ligand of the uranyl(VI) aqua ion by chloride (reaction 2) was investigated using the same ab initio techniques as for reaction 1.[3]

$$UO_2(OH_2)_5^{2+} + Cl^- \rightarrow UO_2Cl(OH_2)_4^+ + H_2O \quad (2)$$

Static electron correlation was treated with complete active space SCF (CAS-SCF) and multiconfiguration quasi-degenerate second-order perturbation theory (MCQDPT2)[5,6] on the basis of the previously used (12/11) active space involving 12 electrons in 11 molecular orbitals (MOs) and a larger (12/12) active space used in recent studies of Hagberg et al.[7] and van Besien et al.[8] The computational methods, together with the approximations and limitations, are the same as in the previous study[3] and will not be reiterated.

* Corresponding author e-mail: francois.rotzinger@epfl.ch.

10.1021/ct8001305 CCC: $40.75 
Published on Web 09/12/2008

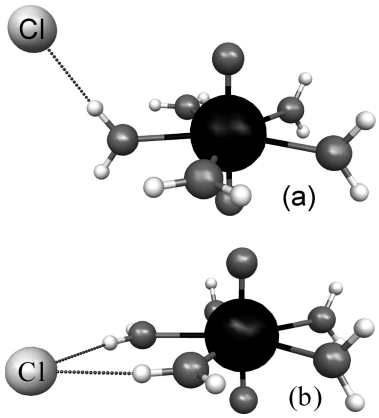



***Figure 1.*** Perspective view of the reactant ion-pair $UO_2(OH_2)_5 \cdot Cl^+$ (CAS-SCF(12/11)-PCM geometry): (a) less stable isomer and (b) stable isomer.

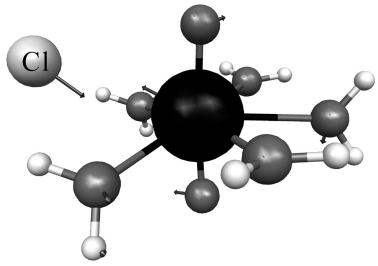



***Figure 2.*** Perspective view and imaginary mode (55.6i $cm^{-1}$) of the transition state $UO_2(OH_2)_5 \cdots Cl^{+\ \ddagger}$ (CAS-SCF(12/11)-PCM geometry).

For reaction 2, no kinetic data are available, but its stability constant ($K_{Cl}$) has been estimated in several studies.[9−13] In an X-ray absorption fine structure (XAFS) spectroscopic study,[9] $K_{Cl}$ was determined as ≈0.2−0.3 $M^{-1}$ (25 °C). According to other work,[9−12] cited in the XAFS article, $K_{Cl}$ is also smaller than 1 $M^{-1}$, 0.7−0.9 $M^{-1}$. In contrast, the dissociation constant determined spectroscopically by Hefley and Amis[13] is 2.28 × $10^{-2}$ M (in water at 25 °C and $I$ = 1.238 M), from which $K_{Cl}$ = 43.9 $M^{-1}$ is derived.

## Computational Details

All of the calculations were performed using the GAMESS[14] programs. For uranium, the relativistic effective core potential (ECP) basis set of Hay and Martin[15] was used, in which the 1s - 5s, 2p - 5p, 3d - 5d, and 4f shells are included in the relativistic core, and the 6s, 7s, 6p, 7p, 6d, and 5f shells are represented by a (10s, 8p, 2d, 4f) basis set contracted to [3s, 3p, 2d, 2f]. For O and H, the 6−31G(d) basis set[16,17] was used ($\alpha_d$ = 1.20[18]), and for chlorine, the ECP basis set of Stevens et al.[19] supplemented with a d polarization function ($\alpha_d$ = 0.65[18]) was used. Figures 1, 2, and 4 were generated with MacMolPlt.[20]

The Hessians, the total energies, and the thermodynamic variables ($\Delta H^{\ddagger}$, $\Delta H$, $\Delta S^{\ddagger}$, $\Delta S$, $\Delta G^{\ddagger}$, and $\Delta G$) at 25 °C were computed as described.[3] Hydration was treated using the polarizable continuum model (PCM)[21,22] as reported previously.[3] Geometries and vibrational frequencies were calculated at the CAS-SCF(12/11)-PCM and CAS-SCF(12/12)-PCM levels, and the total energies were computed with

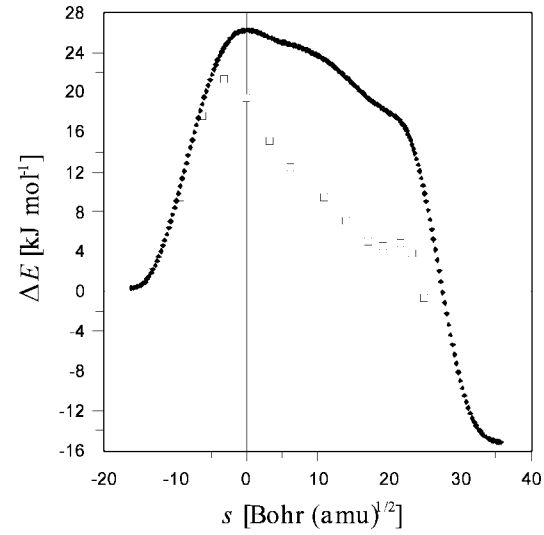



***Figure 3.*** Intrinsic reaction coordinate of reaction 4 (CAS-SCF(12/11)-PCM geometries: CAS-SCF(12/11)-PCM energies (♦) and MCQDPT2(12/11)-PCM energies (□).

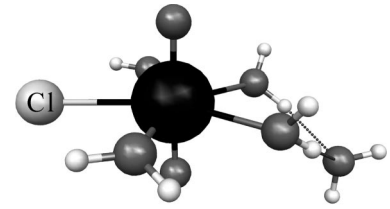



***Figure 4.*** Perspective view of the product $UO_2Cl(OH_2)_4 \cdot OH_2^+$ (CAS-SCF(12/11)-PCM geometry).

MCQDPT2(12/11)-PCM and MCQDPT2(12/12)-PCM.[3] The atomic coordinates of the investigated species are given in Tables S1−S4 (Supporting Information).

The transition state was located by maximizing the energy for the imaginary U···Cl stretching mode via Eigen-mode following, whereby along all of the other modes, the energy was minimized. The intrinsic reaction coordinate (IRC), which is the steepest descent path or the minimum energy path, was computed on the basis of the second-order Gonzalez-Schlegel method.[23]

## Results

**Active Space for the CAS-SCF and MCQDPT2 Calculations and Model for Reaction 2.** Configuration interaction singles-doubles CISD calculations on $UO_2(OH_2)_5^{2+}$ indicated that static electron correlation should be treated at least via a (12/11) active space,[24] which has been used for the study of reaction 1.[3] Preferably, active spaces are chosen to be composed of the corresponding bonding and antibonding pairs of MOs, which gives rise to the same number of electrons and orbitals for closed shell systems. This principle has been applied in the CASPT2(12/12) studies of Hagberg et al.[7] and van Besien et al.[8] In this study, it will be shown that there is no loss of accuracy, when the $\sigma^*$(U═O) MO with a much lower occupation[3] than the other antibonding $\sigma^*$(U═O) MO is excluded from the active space.

For the investigation of reaction 2 with quantum chemical methods, (2) is decomposed into reaction 3 describing the

***Table 1.*** Selected Bond Lengths (Å) of the Uranyl(VI) Complexes Involved in Reaction *2*

| | active space | U=O | U−O | U···O | U−Cl or U···Cl | H···Cl or H···O |
|---|---|---|---|---|---|---|
| $UO_2(OH_2)_5 \cdot Cl^+$ [a] | (12/11) | 1.761, 1.762 | 2.460, 2.519, 2.499, 2.498, 2.518 | | 4.884 | 2.127 |
| $UO_2(OH_2)_5 \cdot Cl^+$ [b] | (12/11) | 1.762, 1.762 | 2.502, 2.502, 2.504, 2.499, 2.504 | | 4.883 | 2.237, 2.239 |
| $UO_2(OH_2)_5 \cdot Cl^+$ [b] | (12/12) | 1.771, 1.771 | 2.502, 2.502, 2.505, 2.499, 2.504 | | 4.884 | 2.236, 2.239 |
| $UO_2(OH_2)_5 \cdots Cl^{+\ddagger}$ | (12/11) | 1.758, 1.761 | 2.572, 2.572, 2.549, 2.541, 2.550 | | 3.283 | 2.455, 2.455 |
| $UO_2(OH_2)_5 \cdots Cl^{+\ddagger}$ | (12/12) | 1.767, 1.770 | 2.571, 2.569, 2.548, 2.539, 2.549 | | 3.290 | 2.458, 2.457 |
| $UO_2Cl(OH_2)_4 \cdot OH_2^+$ | (12/11) | 1.763, 1.763 | 2.524, 2.524, 2.512, 2.512 | 4.143 | 2.820 | 1.869, 1.873 |
| $UO_2Cl(OH_2)_4 \cdot OH_2^+$ | (12/12) | 1.772, 1.772 | 2.525, 2.525, 2.512, 2.513 | 4.152 | 2.819 | 1.870, 1.870 |

[a] Less stable isomer. [b] Stable isomer.

***Table 2.*** Thermodynamic Activation and Reaction Energies and Entropies

| active space | $\Delta E^{\ddagger}$ ($\Delta E$) [kJ mol$^{-1}$] | $\Delta H^{\ddagger}$ ($\Delta H$) [kJ mol$^{-1}$] | $\Delta S^{\ddagger}$ ($\Delta S$) [J K$^{-1}$ mol$^{-1}$] | $\Delta G^{\ddagger}$ ($\Delta G$) [kJ mol$^{-1}$] |
|---|---|---|---|---|
| | | Reaction 4, $I_a$ Mechanism | | |
| (12/11) | 19.1 (−24.9) | 18.4 (−23.3) | −42.3 (−25.7) | 31.0 (−15.7) |
| (12/12) | 19.9 (−23.7) | 19.2 (−22.2) | −43.0 (−24.1) | 32.0 (−15.0) |
| | | Reaction 1, A Mechanism[a] | | |
| (12/11) | 25.6 (20.5)[b] | 22.9 (20.2)[b] | −23.4 (−2.4)[b] | 29.9 (20.9)[b] |
| | | Reaction 1, D (or $I_d$) Mechanism[a] | | |
| (12/11) | 52.6 (47.9)[c] | 50.0 (48.7)[c] | −24.0 (−14.9)[c] | 57.1 (53.2)[c] |

[a] Reference 3. [b] Intermediate $UO_2(OH_2)_6^{2+}$. [c] Intermediate $UO_2(OH_2)_4 \cdot OH_2^{2+}$.

formation of the ion-pair $UO_2(OH_2)_5 \cdot Cl^+$ ($K_{IP,Cl}$) and reaction 4 representing the substitution of an aqua ligand by $Cl^-$ ($K_{Cl}'$).

$$UO_2(OH_2)_5^{2+} + Cl^- \rightarrow UO_2(OH_2)_5 \cdot Cl^+ \quad (3)$$

$$UO_2(OH_2)_5 \cdot Cl^+ \rightarrow UO_2(OH_2)_5 \cdots Cl^{+\ddagger} \rightarrow UO_2Cl(OH_2)_4 \cdot OH_2^+ \quad (4)$$

The ion-pair formation constant ($K_{IP,Cl}$) can be estimated on the basis of the Fuoss equation;[25] it amounts to 0.84 M$^{-1}$ at 25 °C and $I = 1.238$ M$^{-1}$ (the U···Cl distance is 4.88 Å for both isomers of $UO_2(OH_2)_5 \cdot Cl^+$, Table 1). It should be noted that the Fuoss equation takes into account entropy effects approximately. The thermodynamic values and the equilibrium constant ($K_{Cl}'$) for reaction 4 were determined via quantum chemical calculations. Thus, the equilibrium constant for reaction 2, $K_{Cl}$, is available from eq 5.

$$K_{Cl} = K_{IP,Cl} K_{Cl}' \quad (5)$$

**Reactant $UO_2(OH_2)_5 \cdot Cl^+$.** For this ion-pair, there are two isomers: in the less stable one, the chloride ion forms a single hydrogen bond with $UO_2(OH_2)_5^{2+}$ (Figure 1a), and in the more stable isomer, $Cl^-$ forms two hydrogen bonds with the uranyl(VI) cation (Figure 1b). The MCQDPT2(12/11)-PCM energy difference is 13.0 kJ mol$^{-1}$, and the free enthalpy difference is virtually equal, 12.8 kJ mol$^{-1}$. Selected bond lengths of all investigated uranyl(VI) complexes are reported in Table 1. The less stable isomer will not be considered further, since activation and reaction energies have to be based on the global minimum. The MCQDPT2(12/11)-PCM and MCQDPT2(12/12)-PCM energies together with the pertinent thermodynamic values ($\Delta H^{\ddagger}$, $\Delta H$, $\Delta S^{\ddagger}$, $\Delta S$, $\Delta G^{\ddagger}$, and $\Delta G$) for reaction 4 are summarized in Table 2.

The increase of the active space from (12/11) to (12/12) leads to an elongation of the U=O bonds by 0.009 Å for all species (Table 1). The activation or reaction energies and entropies are equal within ≈1 kJ mol$^{-1}$ and ≤2 J K$^{-1}$ mol$^{-1}$, respectively (Table 2).

**Transition State $UO_2(OH_2)_5 \cdots Cl^{+\ddagger}$.** Its geometry, involving an elongated U···Cl bond, but U−O bonds as in the reactant (Table 1), would suggest that this is a transition state for the A mechanism (as that for reaction 1[3]). The $H_2O$ ligand trans to the entering $Cl^-$ ion will leave, although it has the shortest U−O bond (which will be referred to as U−$O_{trans}$). The imaginary mode (Figure 2) represents the entry of the $Cl^-$ ion into the first coordination sphere and a small out-of-plane motion of the trans $H_2O$ ligand. Thus, also the imaginary mode might suggest the A mechanism. However, this substitution reaction 4 proceeds via the $I_a$ pathway, since intermediates are absent from the computed (CAS-SCF(12/11)-PCM level) intrinsic reaction coordinate (IRC) *s* (Figure 3).

The transition state is formed via a common reaction coordinate (Figure 3, negative *s* values), which represents the shortening of the U···Cl bond that occurs concerted with the rearrangement of the $H_2O$ ligands. The IRC calculation (started from the transition state) yields the stable reactant isomer. This proves that the transition state is indeed formed from the reactant in the global minimum.

Product formation (Figure 3, positive *s* values) involves two stages, shortening of the U···Cl bond followed by the elongation of the U−$O_{trans}$ bond. In the range of $s \approx 0-20$ Bohr (amu)$^{1/2}$, there are several inflection points, and the energy does not diminish strongly at the CAS-SCF(12/11)-PCM level (Table 3). Afterward, at $s \gtrsim 22$ Bohr (amu)$^{1/2}$, the U−$O_{trans}$ bond length increases strongly during the steep energy drop leading to the product. In the first stage, the U···Cl bond shortens rapidly to ≈3 Å (at $s \approx 0-6$ Bohr (amu)$^{1/2}$), while the U−$O_{trans}$ bond is elongated only slightly (up to $s \approx 11$ Bohr (amu)$^{1/2}$). In the range of $s \approx 6-11$ Bohr (amu)$^{1/2}$ both, the U−Cl and the U−$O_{trans}$ bonds, change only marginally, and their sum is minimal. The species in this *s* range has a geometry that would be typical for a $UO_2Cl(OH_2)_5^+$ intermediate. Since, however, in this *s* range of ≈6−11 Bohr (amu)$^{1/2}$, no local minimum is present on the potential energy surface (PES), an intermediate for

***Table 3.*** Bond Parameters That Change in a Pronounced Manner during the Transformation of the Transition State into the Product

| $s$ [Bohr (amu)$^{1/2}$] | U−Cl [Å] | U−$O_{trans}$ [Å] | ∠(O=U−$O_{trans}$) [°] |
|---|---|---|---|
| 0[a] | 3.283 | 2.541 | 86.5 |
| 6.225 | 3.024 | 2.591 | 83.8 |
| 10.944 | 2.969 | 2.637 | 78.5 |
| 16.998 | 2.914 | 2.726 | 71.3 |
| 21.513 | 2.876 | 2.940 | 65.5 |

[a] Transition state $UO_2(OH_2)_5\cdots Cl^{+\ddagger}$.

the A mechanism does not exist. In the present case, the local minimum is absent most likely, since the reaction is asymmetric ($H_2O$ is substituted by $Cl^-$), and since the free reaction enthalpy is negative.

The IRC represents the minimum energy pathway. Thus, the substitution of the $H_2O$ ligand trans to the entering $Cl^-$ is the most facile process. The elimination of one of the other four $H_2O$ ligands does not correspond to a minimum energy path, and, hence, such processes are unlikely to be competitive with the elimination of the trans $H_2O$ ligand. If another isomer for the transition state $UO_2(OH_2)_5\cdots Cl^{+\ddagger}$ existed, it would not be possible to exclude that a pathway giving rise to the elimination of one of the nontrans $H_2O$ ligands would take place. However, within the present model, there is only one isomer for the transition state because of the high symmetry of the $UO_2(OH_2)_5^{2+}$ ion.

In the determination of the CAS-SCF(12/11)-PCM geometries and frequencies, dynamic electron correlation was neglected. Therefore, these data are approximate. The MCQDPT2(12/11)-PCM technique produces the most accurate total energies with minimal computational efforts for the present system, but it should be remembered that they are based on approximate geometries. Thus, the difference between the MCQDPT2(12/11)-PCM and the CAS-SCF(12/11)-PCM energies (Figure 3) is due to dynamic electron correlation. At $s \approx 18.5$ Bohr (amu)$^{1/2}$, there is a very shallow local minimum on the MCQDPT2(12/11)-PCM PES. Its depth amounts to <0.3 kJ mol$^{-1}$ which means that the lifetime ($\tau_I$) of this intermediate would be <0.2 ps. Since $\tau_I$ is smaller than the duration of the vibration ($\tau_{vib}$) leading to the product ($\tau_{vib} \approx 0.3-0.4$ ps[3]), this "intermediate" does not exhibit a significant lifetime, and, therefore, it is irrelevant for the reactivity and the reaction mechanism.[3] Hence, also on the basis of the MCQDPT2(12/11)-PCM energies, the $I_a$ mechanism is attributed to reaction 4.

The imaginary mode (Figure 2) reflects the uncommon transition state structure: it describes the entry of the $Cl^-$ ion without concerted elongation of the U−$O_{trans}$ bond of the leaving ligand. For the $I_a$ as well as the I and $I_d$ mechanisms, the imaginary mode represents usually the *concerted* motions of the entering and leaving ligands.[26] The U−$O_{trans}$ bond stretching component is missing because in the transition state, this bond is not weakened; this process takes place later, at $s \gtrsim 17$ Bohr (amu)$^{1/2}$.

The thermodynamic activation free enthalpy of reaction 4 for the $I_a$ pathway is virtually equal to that of reaction 1 via the A mechanism (Table 2).

**Product $UO_2Cl(OH_2)_4\cdot OH_2^+$.** It exhibits a U−Cl bond of 2.82 Å (Table 1 and Figure 4), which is too long by 0.1 Å compared with the experimental value of 2.71−2.72 Å.[9] As pointed out previously,[3,24] this error arises from the neglect of dynamic electron correlation. The product is more stable by 15 kJ mol$^{-1}$ than the reactant ion-pair (Table 2).

## Discussion

**Comparison with Experimental Data.** As already mentioned in the Introduction, the experimental data[9−13] for $K_{Cl}$ are controversial. On the basis of Hefley and Amis's $K_{Cl}$ value of 43.9 M$^{-1}$ (in water at 25 °C and $I = 1.238$ M),[13] and $K_{IP,Cl} = 0.84$ M$^{-1}$ based on the Fuoss equation[25] (25 °C and $I = 1.238$ M), $K_{Cl}' = 52.3$ is estimated. $\Delta G$ based on this $K_{Cl}'$ value is −9.8 kJ mol$^{-1}$, which agrees well with $\Delta G$ computed for reaction 4 (Table 2). According to the other experimental data,[9−12] $\Delta G$ for (4) would be slightly positive ($\approx$2 kJ mol$^{-1}$). All of the experimental data lie within the computational accuracy, which is $\leqslant$10−15 kJ mol$^{-1}$. The approximations and limitations of the model and the computational methods have already been discussed.[3]

**Comparison with Other Computed Data.** Very recently, Bühl et al.[27] studied the stability of chloro complexes of $UO_2(OH_2)_5^{2+}$ as well as their coordination numbers with density functional theory (DFT). For reaction 2, they obtained $\Delta G = -27.2$ and −18.4 kJ mol$^{-1}$, respectively, using the BLYP and B3LYP functionals and PCM hydration. On the basis of Car−Parrinello MD simulations based on the BLYP functional, they computed $\Delta A = 9.6$ kJ mol$^{-1}$. Apart from the BLYP-PCM result, their computed data agree with experiment and the present MCQDPT2-PCM results.

The B3LYP geometries of the $UO_2(OH_2)_5^{2+}$ ion in aqueous solution, for example,[3] are more accurate than the CAS-SCF geometries, but the uranyl(VI)−ligand bond lengths remain too long in comparison with experiment.[3,27] In spite of the worse geometries of CAS-SCF compared with B3LYP, high-level ab initio energies are more accurate than DFT energies.[3,28] The above-discussed DFT calculations[3,27,28] were performed with commonly used functionals. It will be interesting to see computational results on such systems which are realized with, for example, the very recent novel and promising functionals developed by Zhao and Truhlar.[29]

**Substitution Mechanism of Reaction 4.** The water exchange reaction 1 proceeds most likely via the A mechanism, which involves the $UO_2(OH_2)_6^{2+}$ intermediate.[3] Since its lifetime ($\tau_I$) is short, only slightly longer than the duration of the vibration ($\tau_{vib}$) leading to the product, the attribution of the A mechanism cannot be definitive.[3] It was shown in this Results section that structure and imaginary mode (Table 1 and Figure 2) of the transition state for reaction 4 are typical for the A mechanism but that due to the absence of any intermediate on the intrinsic reaction coordinate (Figure 3) reaction 4 proceeds via the $I_a$ mechanism. As shown in the Results section, this substitution reaction proceeds in two steps: the first one is the formation of a $UO_2Cl(OH_2)_5^+$ species via shortening of the U$\cdots$Cl bond after the transition state. In the second step, the U−$O_{trans}$ bond is broken, which leads to the product. Since there is no local minimum on

this PES (Figure 3), this two-stage process has to be classified as $I_a$. This example shows that structure and imaginary mode of the transition state are not sufficient criteria for the determination of the reaction mechanism; it is necessary to find all of the stationary points on the PES between the reactant and the product. Compared with the water exchange reaction 1 exhibiting $H_2O$ as entering ligand, the $Cl^-$ ion causes a change of the mechanism from A to $I_a$, most likely because of the asymmetry and the exergonicity of reaction 4.

The activation energy for the D mechanism is (by definition) independent of the entering ligand. In both reactions 1 and 4, a $H_2O$ ligand is eliminated from the $UO_2(OH_2)_5^{2+}$ cation. Hence, the activation energy for water substitution *via the D mechanism* would be equal for these two reactions. However, a *small* difference in activation energies arising from differences in the environment of the cation, $Cl^-$ being present in the second coordination sphere for (4) and absent for (1), can be expected. This difference will never amount to the sizable $I_a$-D activation energy difference of $\approx$25 kJ mol$^{-1}$. The conclusion that reaction 4 follows the $I_a$ mechanism is safe. For (1) and (4), the activation free enthalpies for the D pathway would be approximately equal and higher by $\approx$25 kJ mol$^{-1}$ than those for the A and $I_a$ mechanisms (Table 2). Thus, as for reaction 1, the dissociative mechanism is unfavorable for (4).

The hydration of the $Cr(NH_3)_5Cl^{2+}$ complex follows the $I_a$ mechanism,[30,31] whereby the corresponding transition state structure is typical for the $I_a$ mechanism, although this reaction is also asymmetric: the bonds of both ligands, $Cl^-$ and $H_2O$, which are involved in the substitution reaction, are longer than in the reactant or the product. This is the reason why such transition states are denoted as $Cr(NH_3)_5\cdots(Cl)(OH_2)^{2+\ddagger}$, for example. The imaginary mode describes the *concerted* formation and breaking of the bonds of the entering and leaving ligands, respectively.[26] Compared with such usual transition states for the $I_a$ (and $I_d$) mechanism,[26] the transition state of (4) is atypical, since it does not exhibit two elongated bonds. A typical transition state for $I_a$ (with two elongated bonds) would be denoted as $UO_2(OH_2)_4\cdots(Cl)(OH_2)^{+\ddagger}$. Hence, because of the absence of an elongated U$\cdots$O bond, this $I_a$ transition state is described as $UO_2(OH_2)_5\cdots Cl^{+\ddagger}$.

**Note Added after ASAP Publication.** This article was released ASAP on September 12, 2008, with minor errors in the first column of Table 1. The correct version was posted on September 20, 2008.

**Supporting Information Available:** Atomic coordinates of the species involved in eq 4 (CAS-SCF(12/11)-PCM and CAS-SCF(12/12)-PCM geometries) (Tables S1−S4). This material is available free of charge via the Internet at http://pubs.acs.org.

## References


(1) Farkas, I.; Bányai, I.; Szabó, Z.; Wahlgren, U.; Grenthe, I. *Inorg. Chem.* **2000**, *39*, 799.

(2) Merbach, A. E. *Pure Appl. Chem.* **1982**, *54*, 1479.

(3) (a) Rotzinger, F. P. *Chem. Eur. J.* **2007**, *13*, 800. (b) Corrigendum.

(4) Bühl, M.; Kabrede, H. *Inorg. Chem.* **2006**, *45*, 3834.

(5) Nakano, H. *J. Chem. Phys.* **1993**, *99*, 7983.

(6) Nakano, H. *Chem. Phys. Lett.* **1993**, *207*, 372.

(7) Hagberg, D.; Karlström, G.; Roos, B. O.; Gagliardi, L. *J. Am. Chem. Soc.* **2005**, *127*, 14250.

(8) Van Besien, E.; Pierloot, K.; Görller-Walrand, C. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4311.

(9) Allen, P. G.; Bucher, J. J.; Shuh, D. K.; Edelstein, N. M.; Reich, T. *Inorg. Chem.* **1997**, *36*, 4676.

(10) Day, R. A.; Powers, R. M. *J. Am. Chem. Soc.* **1954**, *76*, 3895.

(11) Bednarczyk, L.; Fidelis, I. *J. Radioanal. Nucl. Chem.* **1978**, *45*, 325.

(12) Awasthi, S. P.; Sundaresan, M. *Indian J. Chem.* **1981**, *20A*, 378.

(13) Hefley, J. D.; Amis, E. S. *J. Phys. Chem.* **1960**, *64*, 870.

(14) (a) 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. (b) Gordon, M. S.; Schmidt, M. W. 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; pp 1167−1189.

(15) Hay, P. J.; Martin, R. L. *J. Chem. Phys.* **1998**, *109*, 3875.

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

(17) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1971**, *54*, 724.

(18) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571.

(19) Stevens, W. J.; Basch, H.; Krauss, M. *J. Chem. Phys.* **1984**, *81*, 6026.

(20) Bode, B. M.; Gordon, M. S. *J. Mol. Graphics Modell.* **1998**, *16*, 133.

(21) Tomasi, J. *Theor. Chem. Acc.* **2004**, *112*, 184.

(22) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

(23) Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2154.

(24) Rotzinger, F. P. *Chem. Eur. J.* **2007**, *13*, 10298.

(25) Fuoss, R. M. *J. Am. Chem. Soc.* **1958**, *80*, 5059.

(26) Rotzinger, F. P. *Chem. Rev.* **2005**, *105*, 2003.

(27) Bühl, M.; Sieffert, N.; Golubnychiy, V.; Wipff, G. *J. Phys. Chem. A* **2008**, *112*, 2428.

(28) Rotzinger, F. P. *J. Phys. Chem. B* **2005**, *109*, 1510.

(29) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157.

(30) Guastalla, G.; Swaddle, T. W. *Can. J. Chem.* **1973**, *51*, 821.

(31) Rotzinger, F. P. *Inorg. Chem.* **1999**, *38*, 5730.


CT8001305

JCTC Journal of Chemical Theory and Computation

# Energy Analysis of Zn Polycoordination in a Metalloprotein Environment and of the Role of a Neighboring Aromatic Residue. What Is the Impact of Polarization?

Benoit de Courcy,[†] Jean-Philip Piquemal,*[,‡,§] and Nohad Gresh*[,†]

*Laboratoire de Pharmacochimie Moléculaire et Cellulaire, U648 INSERM, UFR Biomédicale, Université Paris Descartes, 45, rue des Saints-Pères, 75006 Paris, France UPMC Univ Paris 06, UMR 7616, Laboratoire de Chimie Théorique, case courrier 137, 4 place Jussieu, F-75005, Paris, France, and CNRS, UMR 7616, Laboratoire de Chimie Théorique, case courrier 137, 4 place Jussieu, F-75005, Paris, France*

Received May 29, 2008

**Abstract:** We analyze the intermolecular interaction energies stabilizing the complex of ethanol in the binding site of alcohol dehydrogenase Zn-metalloenzyme (ADH). In this site Zn(II) is ligated by two cysteine and one imidazole residue and by the ethanol substrate. Ethanol is stacked over a phenylalanine residue. The system has been studied by means of SIBFA (Sum of Interactions Between Fragments Ab initio computed) polarizable molecular mechanics (PMM) supplemented by quantum chemical (QC) computations at various levels of theory. The nonadditivities of the QC interaction energies can be traced back by energy-decomposition analyses and are essentially due to polarization, charge-transfer, and electron correlation energies. These contributions can be reproduced by PMM computations. Interestingly, the polarization energy associated with the presence of the benzene ring in the ADH complex is canceled due to many-body/nonadditivity effects. Therefore this ring does not contribute to stabilization prior to including electron correlation/dispersion effects in the QC calculations or in the absence of the PMM dispersion energy contribution. When these effects are taken into account, the stabilization it contributes is in the 3−9 kcal/mol range, reflecting the need for an accurate reproduction of all components of the interaction energy by PMM.

## Introduction

Cation-$\pi$ interactions constitute a widely encountered determinant in molecular recognition. In proteins, they mostly involve the electron-rich Trp, Phe, and Tyr residues and the cationic Arg or Lys residues [reviewed in ref 1]. A novel motif was put forth by Zaric et al.,[2] in which Trp or Phe could indirectly interact with a metal cation, by means of a stacking interaction with a metal ligand. Examples from X-ray crystallography are provided by metalloproteins having Cu(II),[3] Mg(II),[4] Fe(III),[5] or Zn(II)[6] cofactors. This has led us to analyze the energetical factors stabilizing such complexes. We consider here the recognition site of the Zn-metalloprotein alcohol dehydrogenase (ADH), which catalyzes the oxidation of alcohol to aldehyde, and whose crystal structure was published in ref 6. In this site, Zn(II) is bound by four residues, namely two anionic ones, Cys46 and Cys174, and two neutral ones, His67 and the ethanol substrate. Ethanol is stacked over a Phe residue, Phe93.

In this contribution, we propose to address the following points: (1) What are the magnitudes of the intermolecular interaction energies and of their individual contributions

* Corresponding author e-mail: jpp@lct.jussieu.fr (J.-P.P.) and nohad.gresh@univ-paris5.fr (N.G.).

[†] Université Paris Descartes.

[‡] UPMC Univ Paris 06, UMR 7616.

[§] CNRS, UMR 7616.

10.1021/ct800200j CCC: $40.75 
Published on Web 09/13/2008

within the tetracoordinated Zn complex and the amount of additional stabilization contributed by the Phe residue? (2) What is the extent of nonadditivity in the complex and could possibly nonadditivity modulate Phe binding? (3) To what an extent could the magnitudes of the binding energies be affected by the level of the quantum-chemical (QC) computations, and, in the perspective of computations on large proteins, could polarizable molecular mechanics (PMM) such as the SIBFA (Sum of Interactions Between Fragments Ab initio computed) approach satisfactorily match the QC results?

## Procedure

**QC Computations.** We used the Restricted Variational Space Analysis (RVS)[7] to deconvolute Hartree−Fock (HF) intermolecular interaction energies, denoted $\Delta E$(RVS), into four separate contributions: Electrostatic/Coulomb ($E_{\text{Coul.}}$) and exchange-repulsion ($E_{\text{exch}}$) at first-order (denoted as $E_1=E_{\text{Coul.}}+E_{\text{exch}}$) and polarization ($E_{\text{pol}}$) and charge-transfer ($E_{\text{ct}}$) at second-order (denoted as $E_2=E_{\text{pol}}+E_{\text{ct}}$). These computations were done using the CEP 4−31G(2d) basis set.[8] Contributions of correlation/dispersion to the total intermolecular interaction energies, $\Delta E$(MP2), were computed by the MP2 procedure[9] using the following approximation:

$$\delta E(\text{MP2}) = \Delta E(\text{MP2}) - \Delta E(\text{HF})$$

$$= \Delta E(\text{correlation}) \sim \Delta E(\text{dispersion}) \qquad (1)$$

This representation should enable evaluation of the efficiency of the SIBFA $E_{\text{disp}}$ component since explicit evaluation of dispersion energy by means of Symmetry Adapted Perturbation theory (SAPT[10]) is limited due to the size of the considered systems. We are also aware that electron correlation can affect also the other components of the energy.[10,11]

These computations were done with the GAMESS package.[12] It is to be noted that the RVS procedure as coded in GAMESS removes the Basis Set Superposition Error (BSSE[13]). Thus the reported $\Delta E$(RVS) values are BSSE-corrected at both bi- and multimolecular complexes. The RVS BSSE with the CEP 4−31G(2d) basis set are small, namely, ∼5 kcal/mol out of 624. Small relative BSSE values of <1.5% of the interaction energy were previously reported for polycoordinated complexes of a Zn(II) cation.[14] On the other hand, in the computation of $\delta E$(MP2), the values of $\Delta E$(MP2)-$\Delta E$(HF) are BSSE-uncorrected. Therefore $\delta E$(MP2) embodies those BSSE effects that appear at the MP2 level, and these have larger magnitudes.[15]

Additional DFT computations have been performed using the B3LYP[16] functional. They used the CEP 4−31G(2d) basis set as well as the 6−311G** and LACV3P** basis sets.[17] This latter is equivalent to the 6−311G** basis set on nonmetal atoms. These computations were done with the Gaussian 03 package,[18] except for LACV3P**, where the Jaguar 6.5 software[19] was used. lMP2 computations based on the approach developed by Saebo et al.,[20] as implemented in Jaguar, are also provided at the LACV3P** basis set level and compared to corresponding Jaguar HF values. The BSSE corrections were not done for the 6−311G** and LACV3P** basis sets. Table 2 shows the nonadditivity of CEP 4−31G(2d) BSSE to be small, not exceeding 1.3 kcal/mol. In as much as BSSE has small nonadditivities with other, more extended basis sets as well, this should not affect the analyses of nonadditivity trends.

***Table 1.*** Intermolecular Interaction Energies (kcal/mol) in Complexes *a* and *b*, Which Model the Recognition Site of ADH without and with, Respectively, the Presence of the Benzene Ring[a]

| | complex *a* without Phe93 | | complex *b* with Phe93 | |
|---|---|---|---|---|
| | ab initio | SIBFA | ab initio | SIBFA |
| $E_{\text{Coul.}}/E_{\text{MTP}}{}^{*}$ | −661.6 | −657.9 | −664.4 | −662.4 |
| $E_{\text{exch.}}/E_{\text{rep}}{}^{*}$ | 177.4 | 168.9 | 180.2 | 173.5 |
| $E_1$ | −484.2 | −489.0 | −484.2 | −488.9 |
| $E_{\text{pol}}$(HF)/$E_{\text{pol}}$ | −85.5 | −94.5 | −84.5 | −93.2 |
| $E_{\text{pol}}$(RVS)/$E_{\text{pol}}{}^{*}$ | −113.2 | −121.7 | −111.8 | −121.1 |
| $E_{\text{ct}}$(RVS) | −48.5 | | −48.6 | |
| BSSE | −5.1 | | −5.6 | |
| $E_{\text{ct}}{}^{*}/E_{\text{ct}}$ | −43.4 | −41.3 | −43.0 | −41.3 |
| $E_2$ | −156.6 | −163.0 | −154.8 | −162.4 |
| **$\Delta E$** | **−618.2** | **−624.9** | **−617.3** | **−623.4** |
| $\delta E$(MP2)/$E_{\text{disp}}$ | −48.4 | −65.9 | −57.8 | −70.5 |
| **$\Delta E_{\text{tot}}$** | **−666.6** | **−690.8** | **−675.1** | **−693.9** |

[a] See text for definition. $E_{\text{pol}}$(HF) = $\Delta E$ − $E_1$ − $E_{\text{ct}}$(RVS). [This procedure enables us to evaluate a Morokuma-like polarization energy as the KM approach does not converge, as discussed in the text.] $E_{\text{ct}}{}^{*}$ = $E_{\text{ct}}$(RVS) − BSSE. $E_2$(HF) = $E_{\text{pol}}$(RVS) + $E_{\text{ct}}{}^{*}$. $E_2$(SIBFA) = $E_{\text{pol}}{}^{*}$ + $E_{\text{ct}}$. $\Delta E$(SIBFA) = $E_1$ + $E_{\text{pol}}$ + $E_{\text{ct}}$.

**Polarizable Molecular Mechanics Computations.** We have used the SIBFA polarizable force field. Within the SIBFA procedure,[21] the intermolecular interaction energy is computed as a sum of five separate contributions: penetration corrected multipolar electrostatics,[21c] $E_{\text{MTP}*}$; anisotropic short-range repulsion,[21d] $E_{\text{rep}}$; polarization, $E_{\text{pol}}$; charge-transfer, $E_{\text{ct}}$; and dispersion, $E_{\text{disp}}$. Details on the formulation and calibration of these contributions are given in ref 22. The molecular fragments making up the binding site are methanethiolate, imidazole, benzene, and ethanol. They belong to the SIBFA library of fragments. In keeping with our previous studies, the distributed multipoles[23] and polarizabilities[24] are those derived from their HF molecular orbitals computed with the CEP 4−31G(2d) basis set.

Energy minimizations on the internal coordinates used the Merlin package.[25] Because the X-ray structure shows an unrealistically short distance between Zn(II) and the cysteinate ($Cy^-$) 174 S atom, of about 2.0 Å instead of 2.2−2.3, we have optimized the structure in two steps. First, we relaxed the position of Zn(II) inside the cavity and then relaxed simultaneously the Zn(II) cation, the conformation of the ethanol hydroxyl end, and the methanethiolate group representing Cy-174. For this group, the first H atom lying along the $C_\alpha$-$C_\beta$ bond and used to anchor the methanethiolate moiety was not relaxed. The energy-minimized structure is shown in Figure 1a.

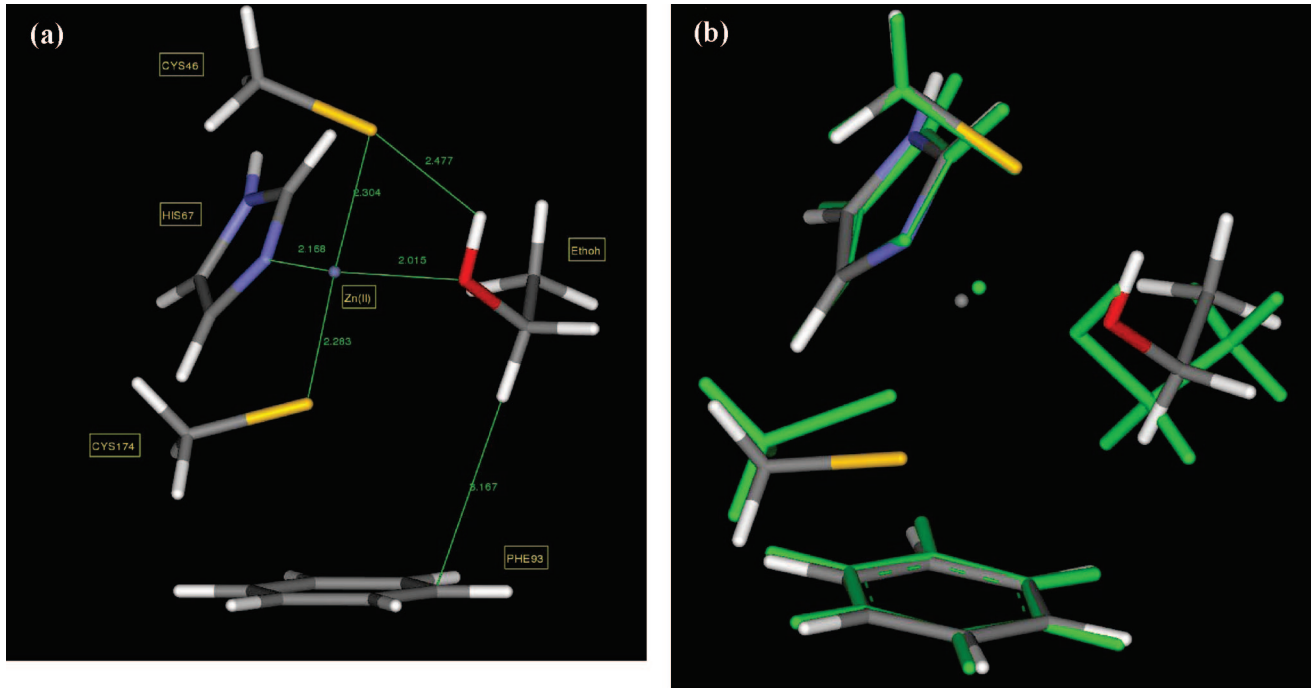



***Figure 1.*** a) Representation of the energy-minimized structure of the recognition site of ADH and b) superimposition of the energy-minimized structures with the PDB structure (in green).

***Table 2.*** Values (kcal/mol) of the RVS/CEP 4-31G(2d) and SIBFA Contributions of the Bimolecular Complexes, of Their Sums and Values of Their Nonadditivities, and of the 6-311G** HF and MP2[a]

| | RVS | SIBFA | RVS | SIBFA | RVS | SIBFA | RVS | SIBFA | RVS | SIBFA | RVS | | | SIBFA | RVS | SIBFA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $E_{Coul.}$ | $E_{MTP^*}$ | $E_{exch.}$ | $E_{rep}$ | $E_1$ | | $E_{pol}$ (RVS) | $E_{pol^*}$ | $E_{pol}$ (HF) | $E_{pol}$ | $E_{ct}$ | BSSE | $E_{ct^*}$ | $E_{ct}$ | $\Delta E$ | |
| $Cy^-$/Imh | 7.6 | 6.1 | 4.2 | 8.8 | 11.8 | 14.9 | −3.2 | −3.7 | −3.1 | −3.6 | −0.6 | −0.4 | −0.2 | 0.0 | 8.1 | 11.3 |
| $Cy^-$/$Cy^-$ | 80.8 | 80.2 | 1.3 | 2.0 | 82.1 | 82.2 | −5.7 | −5.9 | −5.1 | −5.2 | 0.1 | −0.3 | 0.4 | 0.0 | 77.1 | 77.0 |
| $Cy^-$/Zn(II) | −312.9 | −310.1 | 52.4 | 48.8 | −260.5 | −261.3 | −76.3 | −76.2 | −79.6 | −79.6 | −51.1 | −1.3 | −49.8 | −51.2 | −391.3 | −392.1 |
| $Cy^-$/Ethoh | −2.3 | −2.0 | 11.3 | 9.8 | 9.0 | 7.8 | −3.8 | −4.5 | −3.8 | −4.8 | −1.6 | −0.5 | −1.1 | −1.9 | 3.6 | 1.1 |
| Imh/$Cy^-$ | 12.9 | 11.9 | 1.9 | 3.4 | 14.8 | 15.3 | −3.0 | −2.9 | −2.8 | −2.8 | −0.1 | −0.4 | 0.3 | 0.0 | 11.9 | 12.5 |
| Imh/Zn(II) | −83.2 | −83.9 | 20.7 | 20.9 | −62.5 | −63.0 | −61.6 | −52.0 | −63.2 | −53.7 | −16.5 | −0.5 | −16.0 | −15.0 | −142.1 | −131.7 |
| Imh/Ethoh | 3.1 | 1.3 | 1.6 | 3.0 | 4.7 | 4.3 | −0.4 | −0.4 | −0.5 | −0.4 | −0.2 | −0.4 | 0.2 | 0.0 | 4.0 | 3.9 |
| $Cy^-$/Zn(II) | −317.9 | −308.4 | 53.9 | 49.7 | −264.0 | −258.7 | −76.7 | −79.6 | −79.9 | −83.1 | −51.2 | −1.5 | −49.7 | −51.0 | −395.1 | −392.8 |
| $Cy^-$/Ethoh | 9.8 | 7.9 | 2.1 | 2.2 | 11.9 | 10.1 | −2.4 | −2.3 | −2.3 | −2.2 | −0.3 | −0.4 | 0.1 | 0.0 | 9.3 | 7.9 |
| Ethoh/Zn(II) | −61.5 | −60.9 | 22.6 | 20.3 | −38.9 | −40.6 | −48.0 | −44.5 | −49.2 | −45.7 | −10.6 | −0.6 | −10.0 | −8.1 | −98.7 | −94.5 |
| sum | −663.6 | −657.9 | 172.0 | 168.9 | −491.6 | −489.0 | −281.1 | −272.0 | −289.5 | −281.1 | −132.1 | −6.3 | −125.8 | −127.3 | −913.2 | −897.4 |
| complex *a* (without Phe93) | −661.6 | −657.9 | 177.4 | 168.9 | −484.2 | −489.0 | −113.2 | −121.7 | −85.5 | −94.5 | −48.5 | −5.1 | −43.4 | −41.3 | −618.2 | −624.9 |
| $\delta E_{nadd}$ | 2.0 | 0.0 | 5.4 | 0.0 | 7.4 | 0.0 | 167.9 | 150.3 | 204.0 | 186.6 | 83.6 | 1.2 | 82.4 | 86.0 | 295.0 | 272.5 |
| $Cy^-$/Benz | 1.0 | 1.1 | 0.0 | 0.0 | 1.0 | 1.1 | −0.3 | −0.5 | −0.3 | −0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 0.6 |
| Imh/Benz | 0.5 | 0.3 | 0.4 | 0.4 | 0.9 | 0.7 | −0.1 | −0.1 | −0.1 | −0.1 | −0.1 | −0.2 | 0.1 | 0.0 | 0.8 | 0.6 |
| Benz/$Cy^-$ | 0.9 | 0.1 | 2.4 | 3.6 | 3.3 | 3.7 | −3.2 | −3.7 | −3.2 | −3.7 | −0.5 | −0.3 | −0.2 | 0.0 | −0.4 | −0.1 |
| Benz/Zn(II) | −5.1 | −5.5 | 0.0 | 0.0 | −5.1 | −5.5 | 59.6 | −6.9 | 4.9 | −6.9 | −69.7 | 0.0 | −69.7 | 0.0 | −69.9 | −12.3 |
| Benz/Ethoh | −0.1 | −0.5 | 0.4 | 0.6 | 0.3 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | −0.1 | −0.1 | 0.0 | 0.0 | 0.2 | 0.0 |
| sum | −666.4 | −662.4 | 175.2 | 173.5 | −491.2 | −488.9 | −225.2 | −283.1 | −288.1 | −292.2 | −202.5 | −6.9 | −195.6 | −127.3 | −981.9 | −908.6 |
| complex *b* (with Phe93) | −664.4 | −662.4 | 180.2 | 173.5 | −484.2 | −488.9 | −111.8 | −120.1 | −84.5 | −93.2 | −48.6 | −5.6 | −43.0 | −41.3 | −617.3 | −623.4 |
| $\delta E_{nadd}$ | 2.0 | 0.0 | 5.0 | 0.0 | 7.0 | 0.0 | 113.4 | 163.0 | 203.6 | 199.0 | 153.9 | 1.3 | 152.6 | 86.0 | 364.6 | 285.2 |

[a] The IMP2 and MP2 energy gain, $\delta E$(IMP2) and $\delta E$(MP2), respectively, are also reported. The corresponding SIBFA values are recast for ease of comparison.

Its superimposition with the X-ray structure is represented in Figure 1b.

## Results and Discussion

Table 1 reports the intermolecular interaction energies and their contributions in two Zn-tetracoordinated complexes, without and with, respectively, the involvement of the Phe93 side chain. These latter are denoted as complexes *a* and *b*, respectively. These energies were obtained at both QC/CEP 4−31G(2d) and SIBFA levels. Table 2 gives the values of the RVS intermolecular interaction energies in all bimolecular complexes as well as their individual contributions. The corresponding SIBFA interaction energies (without $E_{disp}$) are given in comparison. The values of nonadditivities, $\delta E_{nadd}$, are given as the difference between the summed bimolecular interaction energies and the value in the polycoordinated complex *a* or *b*.

$$\delta E_{nadd} = \Delta E_{poly(many\text{-}body)} - \sum \Delta E_{bimol(2\text{-}body)} \qquad (2)$$

***Table 3.*** Values (kcal/mol) of the CEP 4-31G(2d) RVS, MP2, and DFT Bimolecular Interaction Energies, of Their Sums, and Values of Their Nonadditivities[a]

| | RVS | SIBFA | MP2 | SIBFA | MP2 | MP2 | DFT | SIBFA |
|---|---|---|---|---|---|---|---|---|
| | $\Delta E$ | | $\delta E$(MP2) | $E_{disp}$ | $\Delta E$ (RVS) + $\delta$E(MP2) | $\Delta E$ | $\Delta E$ | $\Delta E_{tot}$ |
| $Cy^-$/Imh | 8.1 | 11.3 | −5.9 | −3.0 | 2.3 | 1.8 | 5.3 | 8.3 |
| $Cy^-/Cy^-$ | 77.1 | 77.0 | −2.7 | −3.2 | 74.3 | 73.9 | 75.5 | 73.8 |
| $Cy^-$/Zn(II) | −391.3 | −392.1 | −20.6 | −18.0 | −411.9 | −413.3 | −445.5 | −410.1 |
| $Cy^-$/Ethoh | 3.6 | 1.1 | −5.8 | −4.2 | −2.2 | −2.7 | −0.4 | −3.1 |
| Imh/$Cy^-$ | 11.9 | 12.5 | −3.8 | −2.0 | 8.1 | 6.3 | 10.0 | 10.5 |
| Imh/Zn(II) | −142.1 | −131.7 | −12.2 | −7.3 | −154.3 | −154.8 | −175.9 | −139.0 |
| Imh/Ethoh | 4.0 | 3.9 | −3.7 | −1.6 | 0.4 | −1.4 | 2.9 | 2.3 |
| $Cy^-$/Zn(II) | −395.1 | −392.8 | −20.4 | −18.3 | −415.5 | −417.1 | −449.6 | −411.1 |
| $Cy^-$/Ethoh | 9.3 | 7.9 | −3.2 | −1.8 | 6.1 | 5.6 | 7.3 | 6.1 |
| Ethoh/Zn(II) | −98.7 | −94.5 | −8.8 | −6.5 | −107.5 | −108.0 | −126.7 | −101.0 |
| sum | −913.3 | −897.4 | −87.1 | −65.8 | −1000.4 | −1009.7 | −1097.2 | −963.3 |
| complex *a* (without Phe93) | −618.2 | −624.9 | −48.4 | −65.8 | −666.6 | −673.1 | −676.4 | −690.8 |
| $\delta E_{nadd}$ | 295.1 | 272.5 | 38.7 | 0.0 | 333.8 | 336.6 | 420.8 | 272.5 |
| $Cy^-$/Ben | 0.7 | 0.6 | −0.8 | −0.1 | −0.1 | −0.1 | 0.6 | 0.5 |
| Imh/Ben | 0.8 | 0.6 | −3.6 | −0.8 | −2.8 | −3.0 | 0.6 | −0.2 |
| Ben/$Cy^-$ | −0.4 | −0.1 | −5.8 | −2.7 | −6.1 | −7.7 | −2.5 | −2.7 |
| Ben/Zn(II) | −69.9 | −12.3 | −47.6 | −0.2 | −117.5 | −117.6 | −140.3 | −12.5 |
| Ben/Ethoh | 0.2 | 0.0 | −2.5 | −0.8 | −2.3 | −3.5 | 0.6 | −0.7 |
| sum | −981.9 | −908.6 | −147.3 | −70.5 | −1129.3 | −1141.5 | −1238.2 | −978.9 |
| complex *b* (with Phe93) | −617.3 | −623.4 | −57.8 | −70.5 | −675.1 | −683.2 | −676.7 | −693.9 |
| $\delta E_{nadd}$ | 364.6 | 285.2 | 89.5 | 0.0 | 454.2 | 458.3 | 561.5 | 285.0 |

[a] The MP2 energy gain, $\delta E$(MP2), is also reported. The corresponding SIBFA values are given along with their QC counterparts.

Positive $\delta E_{nadd}$ values indicate anticooperativity. Table 3 regroups the intermolecular QC interaction energies at the HF level as well as at correlated levels, together with their SIBFA counterparts. Thus the CEP 4−31G(2d) are recast at HF and MP2 levels and complemented with the DFT results. The 6−311G** calculations are given at the HF, DFT, and MP2 levels, while the LACV3P** results are given at the HF, DFT, and lMP2 levels. Since we wish to compare trends, all QC computations were single-point computations done at the SIBFA-energy-minimized geometries.

In Table 1 two values of $E_{pol}$ are given. $E_{pol}$(RVS) is the value of the summed monomer polarization energies at the RVS level and, as in ref 26, is compared to $E_{pol*}$(SIBFA), obtained prior to the iterative inclusion of the effects of the induced dipoles on the field. The Kitaura-Morokuma[27](denoted as KM) procedure strongly overestimates the polarization energy in the presence of strong electric fields such as those generated by metal cations due to a lack of fulfillment by the Pauli principle [see refs 11, 26c, and 28 and references therein]. Indeed, in that case, the repulsive exchange-polarization term is neglected as the wave function is not fully antisymmetrized. Therefore we have indirectly derived a value for $E_{pol}$(HF) after completion of the SCF cycles. Thus, from the converged interaction energy $\Delta E$(HF), we subtracted the summed values of $E_1$ and $E_{ct}$. For such an evaluation, both $\Delta E$(HF) and $E_{ct}$ are uncorrected for BSSE effects[13] for consistency. $E_{pol}$(HF) is then compared to $E_{pol}$(SIBFA), derived at the end of the iterative process on the induced dipoles (see Table 1 for details). The comparisons between $E_{pol}$(HF) and $E_{pol}$(RVS), on the one hand, and between $E_{pol}$(SIBFA) and $E_{pol*}$(SIBFA), on the other hand, give insight into the contribution of induced dipoles to anticooperativity. It is seen that such a contribution has closely similar values from both QC and SIBFA calculations, namely in the 27.3−27.9 kcal/mol range. $E_{ct*}$ denotes the value of $E_{ct}$ after the BSSE correction. $\Delta E$(RVS) and $\Delta E$(SIBFA) denote the total QC and PMM intermolecular interaction energies prior to, respectively, the MP2 procedure and without the $E_{disp}$ contribution. $\Delta E$(MP2) and $\Delta E_{tot}$(SIBFA) denote respectively the corresponding values after the MP2 procedure and with the $E_{disp}$ contribution.

**RVS Results.** Table 1 shows that for both *a* and *b* complexes, a very close agreement between RVS and SIBFA obtains, consistent with previous studies.[14,22,26a,b] It bears on both the total energies and their individual contributions. The magnitude of $\Delta E$(SIBFA) is larger than that of $\Delta E$(RVS) by less than 1.5%. The trends in energy contributions upon including benzene are similar in the RVS and SIBFA approaches. In the context of each methodology, $E_{Coul.}/E_{MTP*}$ and $E_{exch}/E_{rep}$ increase in magnitude by similar amounts. $E_1$ is seen to undergo a virtually null change with both approaches. $E_{pol}$ decreases in magnitude by less than 1.5 kcal/mol out of 100, while $E_{ct}$ is lowered by a negligible amount (< 0.4 kcal/mol). Thus the values of both $\Delta E$(RVS) and $\Delta E$(SIBFA) are modestly (<1.5 kcal/mol out of 620) decreased in magnitude by the involvement of the benzene ring. This could imply that, prior to including electron correlation/dispersion effects, indirect cation-$\pi$ interactions involving benzene would weakly destabilize the complex rather than stabilize it. However, analysis of nonadditivity as reported below (Table 2) shows the present results to be only due to the anticooperativities of $E_{pol}$ and $E_{ct}$: these result from the neutralization of the fields exerted on benzene by Zn(II), on the one hand, and by the two cysteinates, on the other hand.

**MP2 Results.** In complex *a*, $\delta E$(MP2) is smaller in magnitude than $E_{disp}$, namely −48.4 kcal/mol as compared to −65.9, owing to nonadditivity at the MP2 level (see below). This results in $\Delta E_{tot}$(SIBFA) now being larger in magnitude than $\Delta E$(MP2) by 3.5% instead of 1.5% at the RVS level. As shown below, larger relative energy differ-

ences can actually be found between the QC $\Delta E$ values depending upon the basis sets and the handling of correlation. This could be a concern owing to the large magnitudes of the absolute binding energies. In this connection, we have recently investigated the complexes formed between competing inhibitors and protein targets, such as Zn-metalloenzymes $\beta$-lactamase[29a] and phosphomannoisomerase.[29b] In model complexes extracted from the inhibitor-protein complexes, we found that, as in the present study, the SIBFA $\Delta E$ values differed from the CEP 4−31G(2d) target values by relative amounts of 2−3%, and slightly larger relative errors were observed between LACVP3** and CEP 4−31G(2d) $\Delta E$(QC) values. Nevertheless, upon comparing the relative stabilities of several competing complexes for a given model site, the $\Delta E$(QC) values from the two basis sets displayed parallel evolutions, and the $\Delta E$(SIBFA) values very closely reproduced their trends and the energy ranking of the competing complexes. While these studies should be extended to other molecular recognition problems, such results indicate that a correct reproduction of relative energy differences and trends could be expectable from PMM. The corresponding values in complex *b* are −57.8 and −70.5 kcal/mol. Upon comparing the values of $\Delta E$(MP2) and of $\Delta E_{tot}$(SIBFA) in complexes *a* and *b*, it is seen that the benzene ring contributes −8.5 and −3.1 kcal/mol by MP2 and SIBFA computations, respectively. In the superoxide dismutase (SOD) binding site, a Trp residue interacts with the Fe(III) cofactor through a water molecule.[5] A QC study showed it to contribute by a larger amount (10 kcal/mol) to the stabilization energy[2] than computed here for ADH. However, a different balance of effects could come into play in the SOD site, since in contrast to ADH, the field exerted on the ring by a trivalent metal cation could now be incompletely neutralized by the anionic charges of the iron-coordinating Asp residue and azide molecule. The values of $\Delta E$ are very large since the present calculations are in the gas phase. Extrapolation to the actual ethanol-ADH complex would require the inclusion of the entire protein and perform energy balances taking into account the solvation energy of the complex, on the one hand, and the separate desolvation energies of the protein and the substrate prior to complex formation, on the other hand. Inclusion of the latter terms results in a considerable reduction of the magnitudes of the resulting binding energies. Accounting for the protein and ligand conformational energy rearrangement further reduces their magnitudes. Such energy balances have been reported concerning the complexation of inhibitors to the Zn-metalloproteins phosphomannoisomerase[29b] and the second Zn-finger of the HIV-1 nucleocapsid.[30] They resulted in binding energies in the range of −20 kcal/mol. Inclusion of entropy effects should further reduce their magnitudes. Nevertheless the trends in $\Delta E$ contributions regarding the effect of Phe93 should be conserved in the model site compared to the entire protein.

## Analysis of Nonadditivity

**Complex *a.*** *1) QC Results.* Nonadditivity in several polycoordinated Zn(II) complexes was previously analyzed in parallel by RVS and SIBFA.[26a,b] Consistent with these studies, as shown in Table 2 $E_{pol}$ is found here to be the most anticooperative contribution, with $\delta E_{nadd}$ amounting to 168 and 204 kcal/mol for $E_{pol}$(RVS) and $E_{pol}$(KM), respectively. $E_{ct}$ is also strongly anticooperative, with $\delta E_{nadd}$ in the 83−87 kcal/mol range, while $E_{Coul.}$ and $E_{exch}$ have very modest anticooperativities (2 and 5 kcal/mol, respectively). A small anticooperativity of the BSSE correction (1.3 kcal/mol out of 6) can be noted. Table 3 shows that the energy gain due to the MP2 procedure, $\delta E$(MP2), has a significant anticooperativity. It amounts to 38.7 kcal/mol and is in the same range as found in related polycoordinated Zn(II) complexes.[26a,b] Since such a value comes on top of the nonadditivities of $E_{pol}$ and $E_{ct}$ at the RVS level, it should stem mainly from the increases of the relative weights of both $E_{pol}$ and $E_{ct}$ due to correlation.[11] Further SIBFA studies are planned using correlated multipoles and polarizabilities [see for an example ref 21d]. They should allow quantification of the extent to which correlation affects the anticooperativities of $E_{pol}$ and $E_{ct}$ in the context of molecular mechanics. For this purpose, a preliminary recalibration of $E_{pol}$ and $E_{ct}$ on such monoligated Zn(II) complexes will be necessary and is outside the scope of this work. In a study of the complexes of nucleic acid base pairs with divalent metal cations, it was recalled[31] that the nonaddivity of the actual dispersion term appears only at the MP3 level[32] and is therefore not accounted for in the present calculations.

*2) SIBFA Results.* $E_{pol}$ and $E_{ct}$ are presently the only nonadditive SIBFA contributions. The values of $\delta E_{nadd}$(SIBFA) for $E_{pol}$ and $E_{ct}$ are consistent with the RVS ones. Those of $E_{pol}$ and $E_{pol*}$ are somewhat smaller than the corresponding RVS ones, but such underestimations are found to compensate for some corresponding underestimations of $E_{pol}$(SIBFA) with respect to $E_{pol}$(QC). This occurs notably in one binary complex, that of Zn(II) with imidazole at the 2.16 Å Zn−N distance. As commented on in ref 26, $\delta E_{nadd}$ is larger when in the QM computations $E_{pol}$ is derived at the outcome of the SCF procedure, and when in the SIBFA computations, it is computed after iterative inclusion of the induced dipoles. The value of $\delta E_{nadd}$ for $E_{ct}$ is very close to the corresponding RVS ones. This indicates that, compared to our previous calibration, the $E_{ct}$ recalibration reported in ref 14 affords an improved control of its large nonadditivity in Zn(II) complexes.[26a,b]

**Complex *b.*** Quantifying nonadditivity by RVS/CEP 4−31G(2d) is prevented by the fact that at large separation, the bimolecular Zn(II)-benzene complex diverges asymptotically toward an open-shell state where an electron is transferred to the dication. In the present complex, the distance between Zn and the centroid of benzene is 5.4 Å. The RVS analysis gives a value of $E_{ct}$ of −69.7 kcal/mol and an artifactual positive $E_{pol}$ value of 59.6. Interposing the ethanol ligand, as in the trimolecular benzene-ethanol-Zn(II) complex, recovers a negative $E_{pol}$ of −25.5 kcal/mol and a reduced $E_{ct}$ value of −43.3 kcal/mol (unpublished). A fortiori, completion of the Zn(II) coordination shell recovers meaningful values of both contributions. The SIBFA computations show that the separate values of $E_{pol}$ in the bimolecular complexes of benzene with Zn(II) and with one cysteinate are significant despite the distances of separation, amounting to −6.9 and −3.8 kcal/mol, respectively, but that the actual

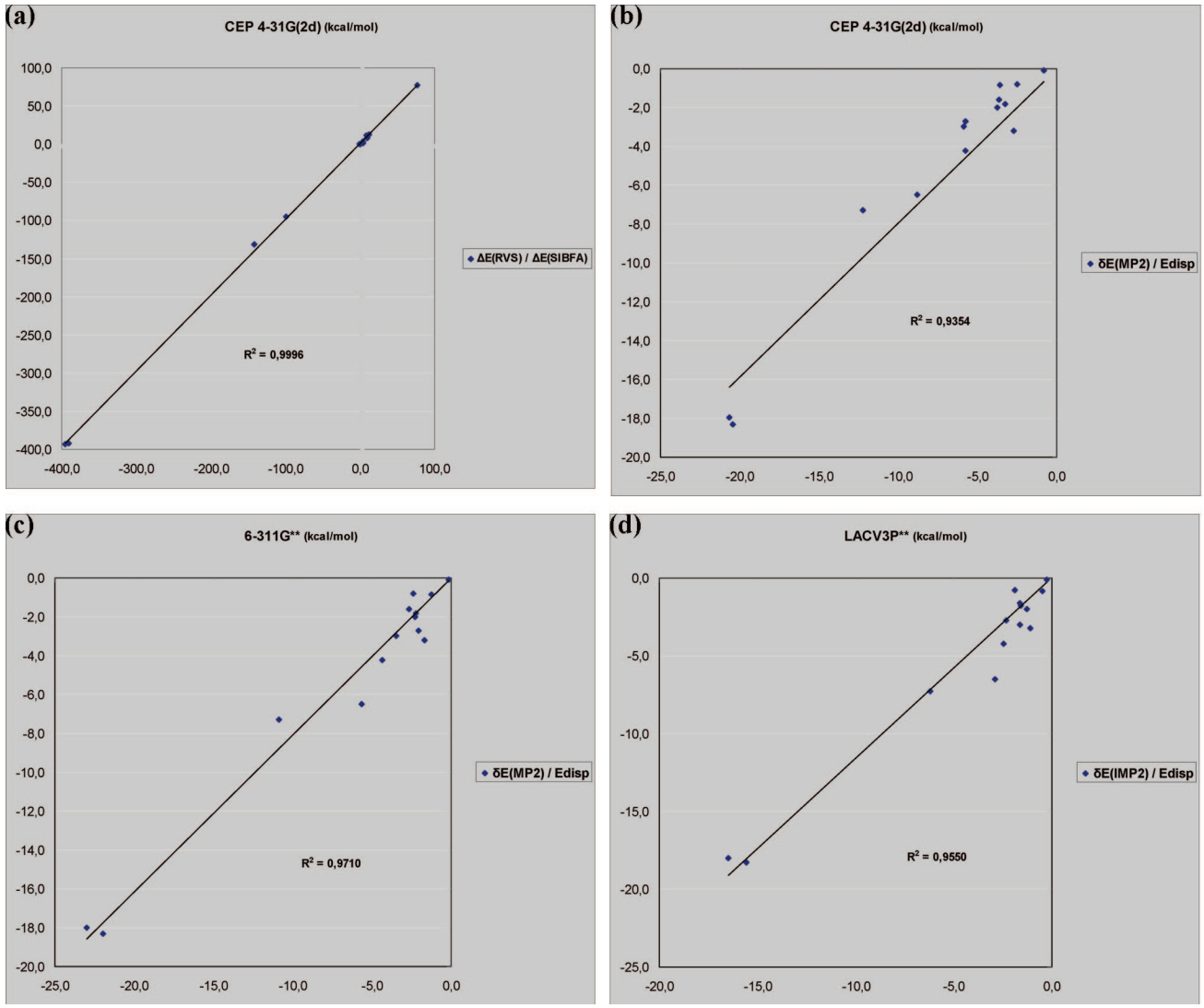



***Figure 2.*** Correlation between the SIBFA and QC interaction energies in all bimolecular complexes except for Zn(II)-benzene. a) $\Delta E$(SIBFA) and $\Delta E$(RVS/HF); b) $E_{disp}$(SIBFA) and $\delta E$(MP2)/CEP4−31G(2d); c) $E_{disp}$(SIBFA) and $\delta E$(MP2)/6−311G**; and d) $E_{disp}$(SIBFA) and $\delta E$(lMP2)/LACV3P**.

increase of $E_{pol}$ upon passing from complex *a* to *b* is negligible owing to the anticooperativity of $E_{pol}$. Thus in the context of SIBFA, this should leave $E_{disp}$ as the sole energy contribution stabilizing complex *b* over complex *a*.

We have reported in Figure 2a the correlation between $\Delta E$ (RVS) and $\Delta E$(SIBFA) bearing on all bimolecular complexes, except Zn(II)-benzene. The $r^2$ correlation coefficient is 0.9996. We have similarly evaluated the correlation between $E_{disp}$(SIBFA) and $\delta E$(MP2). $E_{disp}$(SIBFA) is an approximation to the real dispersion, since electron correlation affects also electrostatic and induction terms [see ref 11 and references therein]. Therefore a less satisfactory correlation has to be expected, especially as BSSE effects are in general not negligible. A reasonable $r^2$ of 0.9354 nevertheless is obtained with $\delta E$(MP2)/CEP 4−31G(2d) (Figure 2b), which actually increases to 0.97010 concerning $\delta E$(MP2)/6−311G** (Figure 2c). The $r^2$ value with respect to $\delta E$(lMP2)/LACV3P** is 0.9550 (Figure 2d). At this point it is recalled that the calibration of $E_{disp}$(SIBFA) was performed[33] on the basis of SAPT computations; however, SAPT can become intractable upon increasing the size of the molecular complexes. Alternatively, MPn ($n = 3$ or 4) or CCSD(T) computations could be used for $E_{disp}$(SIBFA) recalibration on model dimeric complexes. There could be two means to improve the representation of correlation in SIBFA. One is a simple rescaling on the basis of such more extended correlated computations. The second is, as mentioned above, the use of correlated multipoles and polarizabilities:[21d] these could provide the contributions of correlation to $E_{MTP}$ and $E_{pol}$. A rescaling of $E_{disp}$ should be done subsequently to provide the actual contribution of the van der Waals component.

*Effects of the Level of the QC Computations (Tables 3 and 4).* This analysis was performed in order to evaluate, on the one hand, the sensitivity of $\delta E_{nadd}$ to the level of the QC computation, and, on the other hand, the amount of stabilization due to correlation in both complexes *a* and *b*. In complex *b*, the 'van der Waals' component should be amplified, which would lead to $\Delta E$ underestimation by DFT [see ref 34 and references therein]. On the other hand, complex *a* is predominantly stabilized by electrostatic interactions, so that all QC procedures could be expected to

***Table 4.*** Values (kcal/mol) of the LACV3P** HF, IMP2, and DFT Bimolecular Interaction Energies, of Their Sums, and Values of Their Nonadditivities and Corresponding Values of the 6-311G** HF and MP2[a]

| | LACV3P** | | | | | | | 6−311G** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | HF | SIBFA | IMP2 | SIBFA | IMP2 | DFT | SIBFA | HF | SIBFA | MP2 | SIBFA | MP2 | SIBFA |
| | $\Delta E$ | | $\delta E$(IMP2) | $E_{disp}$ | $\Delta E$ | | $\Delta E_{tot}$ | $\Delta E$ | | $\delta E$(MP2) | $E_{disp}$ | $\Delta E$ | $\Delta E_{tot}$ |
| $Cy^-$/Imh | 8.4 | 11.3 | −1.6 | −3.0 | 6.7 | 6.9 | 8.3 | 8.3 | 11.3 | −3.5 | −3.0 | 4.8 | 8.3 |
| $Cy^-$/$Cy^-$ | 78.4 | 77.0 | −1.1 | −3.2 | 77.3 | 77.3 | 73.8 | 78.4 | 77.0 | −1.7 | −3.2 | 76.7 | 73.8 |
| $Cy^-$/Zn(II) | −378.5 | −392.1 | −16.5 | −18.0 | −395.0 | −407.0 | −410.1 | −368.4 | −392.1 | −23.0 | −18.0 | −391.4 | −410.1 |
| $Cy^-$/Ethoh | 4.3 | 1.1 | −2.5 | −4.2 | 1.9 | 0.4 | −3.1 | 4.3 | 1.1 | −4.3 | −4.2 | 0.0 | −3.1 |
| Imh/$Cy^-$ | 12.6 | 12.5 | −1.3 | −2.0 | 11.4 | 11.4 | 10.5 | 12.5 | 12.5 | −2.3 | −2.0 | 10.2 | 10.5 |
| Imh/Zn(II) | −137.7 | −131.7 | −6.2 | −7.3 | −143.9 | −155.9 | −139.0 | −133.7 | −131.7 | −10.9 | −7.3 | −144.6 | −139.0 |
| Imh/Ethoh | 3.6 | 3.9 | −1.6 | −1.6 | 1.9 | 2.6 | 2.3 | 3.5 | 3.9 | −2.6 | −1.6 | 0.9 | 2.3 |
| $Cy^-$/Zn(II) | −382.5 | −392.8 | −15.6 | −18.3 | −398.1 | −410.9 | −411.1 | −372.3 | −392.8 | −21.9 | −18.3 | −394.2 | −411.1 |
| $Cy^-$/Ethoh | 10.2 | 7.9 | −1.6 | −1.8 | 8.6 | 8.2 | 6.1 | 10.1 | 7.9 | −2.2 | −1.8 | 7.9 | 6.1 |
| Ethoh/Zn(II) | −97.2 | −94.5 | −2.9 | −6.5 | −100.1 | −112.2 | −101.0 | −93.5 | −94.5 | −5.7 | −6.5 | −99.2 | −101.0 |
| sum | −878.5 | −896.7 | −50.8 | −65.8 | −929.3 | −979.3 | −962.5 | −850.9 | −962.5 | −78.1 | −65.8 | −929.0 | −962.5 |
| complex *a* (without Phe93) | −607.9 | −624.9 | −20.8 | −65.8 | −628.8 | −629.5 | −690.8 | −594.7 | −624.9 | −35.4 | −65.8 | −630.1 | −690.8 |
| $\delta E_{nadd}$ | 270.6 | 272.0 | 30.0 | 0.0 | 300.5 | 349.9 | 271.9 | 256.2 | 271.9 | 42.7 | 0.0 | 298.9 | 271.9 |
| $Cy^-$/Ben | 0.8 | 0.6 | −0.2 | −0.1 | 0.5 | 0.6 | 0.5 | 0.8 | 0.6 | −0.2 | −0.1 | 0.6 | 0.5 |
| Imh/Ben | 0.5 | 0.6 | −0.5 | −0.8 | | 0.4 | −0.2 | 0.5 | 0.6 | −1.3 | −0.8 | −0.7 | −0.2 |
| Ben/$Cy^-$ | 0.3 | −1.5 | −2.3 | −2.7 | −2.1 | −0.9 | −2.7 | 0.3 | −1.5 | −2.0 | −2.7 | −2.6 | −2.7 |
| Ben/Zn(II) | −63.5 | −12.3 | −30.7 | −0.2 | −94.1 | −114.4 | −12.5 | −60.3 | −12.3 | −44.1 | −0.2 | −105.1 | −12.5 |
| Ben/Ethoh | 0.0 | 0.7 | −1.9 | −0.8 | −1.9 | −0.1 | −0.7 | 0.0 | 0.7 | −2.4 | −0.8 | −1.1 | −0.7 |
| sum | −940.4 | −908.5 | −908.5 | −70.5 | −1026.9 | −1093.7 | −978.1 | −909.5 | −978.1 | −128.0 | −70.5 | −1037.9 | −978.1 |
| complex *b* (with Phe93) | −608.0 | −623.4 | −24.1 | −70.5 | −632.1 | −630.1 | −693.9 | −595.0 | −623.4 | −40.3 | −70.5 | −635.3 | −693.9 |
| $\delta E_{nadd}$ | 332.4 | 285.3 | 285.3 | 0.0 | 394.8 | 463.5 | 284.4 | 314.5 | 284.4 | 87.7 | 0.0 | 402.6 | 284.4 |

[a] The IMP2 and MP2 energy gains, $\delta E$(IMP2) and $\delta E$(MP2), respectively, are also reported. The corresponding SIBFA values are recast for ease of comparison.

show similar trends. The 6−311G** and LACV3P** computations differ by the use of an effective large core pseudopotential on Zn(II) in the latter, while a full electron basis set is used on the cation in the 6−311G** calculations. Table 3 reports the values of the total interaction energies and those of all bimolecular complexes at the CEP 4−31G(2d) level, and Table 4 reports the corresponding values at the LACV3P** and 6−311G** levels.

**Complex *a*.** At the HF level, the magnitudes of the interaction energies are along the sequence CEP 4−31G(2d) > LACV3P** > 6−311G**. The 10 out of −620 kcal/mol energy difference between the first two basis sets is the same as the corresponding one previously computed for the complex of Zn(II) with two cysteinates and two imidazoles that represented a Zn-finger Zn binding site.[14] The magnitudes of anticooperativity effects follow the same trend as the $\Delta E$ values. Such larger $\delta E_{nadd}$ values with the CEP 4−31G(2d) basis set translate the larger relative weights of the summed second-order contributions with respect to the summed first-order ones, occurring with this basis set compared to the LACV3P and 6−311G** ones. It could be due to the presence in this basis of two diffuse 3d polarization AOs on the heavy atoms.

In Table 3 two columns of MP2 values are given, namely columns 5 and 6 of results. The first column gives the results after addition of the MP2 energy gain, $\delta E$(MP2), to $\Delta E$-(RVS), i.e., after BSSE correction at the HF level. The second column gives the corresponding results after the addition of $\delta E$(MP2) to $\Delta E$(HF), namely, without the BSSE(HF) correction. Thus the energy values are slightly smaller than in the preceding column. The penultimate column gives the DFT results, and the last column recasts the SIBFA ones. At the MP2 level, the total interaction energies are larger with the CEP 4−31G(2d) than with the 6−311G** basis set, but the $\delta E_{nadd}$ values are close, 38.7 and 42.7 kcal/mol, respectively. The larger magnitudes of $\Delta E$(MP2) with the CEP 4−31G(2d) basis set stem from their larger magnitudes in the separate monoligated Zn(II) complexes. It is noted that for these complexes at optimized Zn-ligand distances the values of $\Delta E$(MP2) using the CEP 4−31G(2d) set[35] are found to be very close to the large basis set computations recently published by Rayon et al. on a series of representative Zn-ligand complexes.[36] These computations used aug-cc-pVTZ basis sets with both MP2 and CCSD(T) methods.

At the DFT level, the CEP 4−31G(2d) basis set has a larger $\delta E_{nadd}$ than the LACV3P** basis set (420.8 versus 349.9 kcal/mol), the 71 kcal/mol difference being amplified with respect to the corresponding HF $\delta E_{nadd}$ value which amounted to 24.5 kcal/mol. With both CEP 4−31G(2d) and LACV3P** basis sets, the DFT computations are seen to overestimate the Zn-monoligand interaction energies. In this connection, recent analyses[11] of DFT intermolecular interaction energies with the Constrained Space Orbital Variation procedure[37] linked these overestimations to a strong increase of polarization, charge-transfer contributions, and Zn(II) polarizability as compared to the corresponding HF values. Overestimations of $\Delta E$ were recently also noted in the case of Zn(II) complexes with anionic ligands,[36] while new functionals are being developed and evaluated.[38] Using B3LYP and the CEP 4−31G(2d) basis set, the DFT larger $\delta E_{nadd}$ values compared to MP2 compensate for the larger monoligated $\Delta E$(DFT) values. As a result, the final $\Delta E$(DFT) values come close to the $\Delta E$(MP2) ones, −676.4 as compared to −666.6 kcal/mol.

At the lMP2 level, and with the LACV3P** basis set, $\Delta E$(lMP2) has in the monoligated Zn(II) complexes smaller values than $\Delta E$(DFT) but also a smaller $\delta E_{nadd}$ value: this mutual compensation results in $\Delta E$(lMP2) and $\Delta E$(DFT) being virtually equal in the polycoordinated complex.

**Complex *b*.** We compare here the contribution of the benzene ring to stabilization, as translated by the energy variations upon passing from complex *a* to complex *b*. It will be denoted $\delta E_{a\text{-}b}$. At the HF level, and consistent with the CEP 4−31G(2d) results, both 6−311G** and LACV3P** basis sets indicate the benzene ring to contribute negligibly to the interaction energy. This occurs in spite of the artifactually strong $\Delta E$ value in the 'bimolecular' Zn(II)-benzene complex of −60 to −63.5 kcal/mol, comparable to the corresponding CEP 4−31G(2d) value of −70 kcal/mol value.

At the MP2 level, $\delta E_{a\text{-}b}$/6−311G** amounts to −5.2 kcal/mol. This value is smaller in magnitude than the $\delta E_{a\text{-}b}$/CEP 4−31G(2d) value of −8.5 kcal/mol but closer to the SIBFA value of −3.1 kcal/mol. Concerning the bimolecular complexes involving benzene with ethanol, each cysteinate as well as imidazole, at the HF level, $\Delta E$/CEP 4−31G(2d) is seen to be only slightly more stabilizing than $\Delta E$ computed with the larger 6−311G** basis set. However the corresponding energy differences are enhanced at the MP2 level, regardless of the relative proximity to benzene. This illustrates the need for extended basis sets in order to handle correlation. Concerning the complexes between two conjugated molecules, it was shown by Hobza and Sponer[39] that extrapolation to the complete basis set (CBS) limit is necessary to obtain converged estimates of the MP2 interaction energy as well as stable MP2-CCSD(T) energy differences. The CCSD(T) interaction energies have for such complexes smaller magnitudes than the MP2 ones. The fact that, for the complexes involving benzene, $E_{disp}$(SIBFA) in its present formulation is closer in magnitude to $\delta E$(MP2) computed with the 6−311G** than to the CEP 4−31G(2d) constitutes thus a favorable feature. Furthermore, as previously observed in a series of H-bonded complexes using the CEP 4−31G(2d) basis set,[15] the overestimations of $\Delta E$(MP2)/CEP 4−31G(2d) can stem in part from large BSSE effects at the MP2 level, in marked contrast with the small CEP 4−31G(2d) BSSE magnitudes at the HF level. These caveats are to be noted, while, on the other hand and as above-mentioned, the Zn(II) monoligated interaction energies can be accurately computed at the MP2/CEP 4−31G(2d) basis set.

$\delta E_{a\text{-}b}$ at the DFT level has extremely small magnitudes (<0.3 kcal/mol) with both CEP 4−31G(2d) and LACV3P** basis sets. By contrast, at the lMP2 level, $\delta E_{a\text{-}b}$ (LACV3P**) amounts to −3.3 kcal/mol.

## Conclusions

We have analyzed by SIBFA and QC computations the energetical factors stabilizing the Zn-binding site of alcohol dehydrogenase (ADH), in which Zn(II) is polycoordinated to two cysteinates and one histidine and by the ethanol substrate. A Phe residue is stacked over ethanol. The stabilization energy it contributes was computed to be in the range of 3−9 kcal/mol. However, because of the mutual cancelation of the fields polarizing the benzene ring in the ADH binding site as compared to the separate bimolecular complexes involving it, no stabilization was computed in the context of QC/HF calculations and, concerning the SIBFA procedure, in the absence of the $E_{disp}$ contribution.

Regarding nonadditivity, the present analyses have shown, in unanticipated fashion, some significantly differing behaviors of QC depending upon the level of computations. Thus DFT was found to display much larger anticooperativity than either MP2 in CEP 4−31G(2d) computations or lMP2 in LACV3P** computations. However such larger $\delta E_{nadd}$ values were in both cases found to compensate for the larger DFT magnitudes of the separate Zn-monoligated complexes: with the CEP 4−31G(2d) basis set, this resulted in $\Delta E$(DFT) differing from $\Delta E$(MP2) by small amounts, namely 9.8 kcal/mol out of 670 in complex *a* and 1.6 kcal/mol out of 675 in complex *b*. On the other hand, more conservatively, $\delta E_{nad\text{-}d}$(MP2) was found to have very similar values with either CEP 4−31G(2d) and 6−311G** basis sets, namely 38.7 and 42.7 kcal/mol, respectively, in *a* and 89.5 and 87.7 kcal/mol in *b*. The lMP2 computations with the LACV3P** basis set had smaller corresponding $\delta E_{nadd}$ values of 29.8 and 62.4 kcal/mol. The large $\delta E_{nadd}$ values in complex *b* are due to the artificially strong $E_{ct}$ value in the benzene-Zn(II) complex which are not in direct interaction and the onset of an open-shell state where an electron is transferred to Zn(II).

The present investigation also confirms the accuracy of the SIBFA procedure into reproducing its target QC/CEP 4−31G(2d) interaction energies and a good control of both $E_{pol}$ and $E_{ct}$ nonadditivities. The relative error is 1.5% concerning the HF level. It raises however to 3.5% at the correlated level, because $E_{disp}$ is additive, while $\delta E$(MP2) in polycoordinated Zn(II) complexes is anticooperative.[26a,b] The −3.1 kcal/mol stabilization contributed by the benzene ring appears closer to the 6−311G** than the CEP 4−31G(2d) MP2 value (−5.2 and −8.5 kcal/mol, respectively). In light of the results published by Hobza and Sponer,[39] it is likely that the 6−311G** stabilization energy of −5.2 kcal/mol is closer to the CBS result than the CEP 4−31G(2d) one and that its magnitude is itself an upper bound to the CCSD(T) value.

Polarization is indispensable to reliably compute cation-$\pi$ complexes where an aromatic ring directly interacts with a cationic partner.[40] To our knowledge, the very first evaluation of $E_{pol}$ in such complexes was published in 1980 upon studying the complexes of mono- and tetramethylammonium with the indole ring.[41] However the present study shows that, as concerns the ADH binding site, the contribution of the benzene ring to overall stabilization is not due to polarization. This contribution is canceled out because the dicationic charge, with which benzene interacts indirectly, is neutralized by the two anionic cysteinates. The main contribution then stems from dispersion, as in classical, nonpolarizable force-fields. Thus the present analysis reaffirms the need for a complete separability of the interaction potential[21d,22] in order to

accurately reproduce each of the QC contributions to the binding energy in a diversity of situations (see Table 1).

Finally, the present results also suggest that PMM procedures which can accurately reproduce the results from QC computations could be used as a tool to refine X-ray crystal structures, as was previously demonstrated in the context of quantum chemistry[42] or by the use of distributed multipoles.[43,44] In view of such an evaluation, and as we had done in previous papers,[29] we give as Supporting Information the coordinates of the SIBFA energy-minimized structure. These could also be used to benchmark other polarizable molecular mechanics approaches.

**Acknowledgment.** The computations reported in this work were done in the computer centres of CINES (Montpellier, France), CRIHAN (Rouen, France), and CCRE (Paris, France). We wish to thank Nicole Audiffren from CINES for her invaluable help in the RVS computations. This work was supported in part by la Ligue Nationale contre le Cancer, Equipe Labellisée 2006 (U648).

**Supporting Information Available:** Cartesian coordinates of the energy-minimized structure of complex *b*. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Ma, J. C.; Dougherty, D. A. *Chem. Rev.* **1997**, *97*, 1303.

(2) Zaric, S. D.; Popovic, D. M.; Knapp, E.-W. *Chem. Eur. J.* **2000**, *6*, 21.

(3) Guss, J. M.; Merritt, E. A.; Phizackerley, R. P.; Freeman, H. C. *J. Mol. Biol.* **1996**, *262*, 686.

(4) Bellsolell, L. L.; Prieto, J.; Serrano, L.; Coll, M. *J. Mol. Biol.* **1994**, *238*, 489.

(5) Lah, M. S.; Dixon, M. M.; Pattridge, K. A.; Stallings, W. C.; Fe, J. A.; Ludwig, M. L. *Biochemistry* **1995**, *34*, 1646.

(6) Li, H.; Hallows, W. H.; Punzi, J. S.; Pankiewicz, K. W.; Watanabe, K. A.; Goldstein, B. M. *Biochemistry* **1994**, *33*, 11734.

(7) Stevens, W. J.; Fink, W. *Chem. Phys. Lett.* **1987**, *139*, 15.

(8) Stevens, W. J.; Basch, H.; Krauss, M. *J. Chem. Phys.* **1984**, *81*, 6026.

(9) Pople, J. A.; Binkley, J. S.; Seeger, R. *Int. J. Quantum Chem.* **1976**, *10*, 1.

(10) (a) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887. (b) Langlet, J.; Caillet, J.; Bergès, J.; Reinhardt, P. *J. Chem. Phys.* **2003**, *118*, 6157.

(11) Piquemal, J.-P.; Marquez, A.; Parisel, O.; Giessner-Prettre, C. *J. Comput. Chem.* **2005**, *26*, 1052.

(12) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., Jr *J. Comput. Chem.* **1993**, *14*, 1347.

(13) (a) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553. (b) Cammi, R.; Hoffmann, H. J.; Tomasi, J. *Theor. Chim. Acta* **1989**, *76*, 297.

(14) Gresh, N.; Piquemal, J.-P.; Krauss, M. J. *Comput. Chem.* **2005**, *26*, 1113.

(15) Gresh, N.; Leboeuf, M.; Salahub, D. R. In *Modelling the Hydrogen Bond, ACS Symposium Series 569*; Smith, D. A., Ed.; 1994; 82.

(16) (a) Becke, A. D. *J. Chem. Phys.* **1988**, *88*, 1053. (b) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev.* **1988**, *B37*, 785. (c) Becke, A. *J. Chem. Phys.* **1993**, *98*, 5648.

(17) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299.

(18) 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, 2007.

(19) *Jaguar 6.5*; Schrodinger Inc.: Portland, OR, 2005.

(20) (a) Saebo, S.; Pulay, P. *J. Chem. Phys.* **1987**, *86*, 914. (b) Murphy, R. B.; Beachy, M. D.; Friesner, R. A. *J. Chem. Phys.* **1995**, *103*, 1481.

(21) (a) Gresh, N.; Claverie, P.; Pullman, A. *Theor. Chim Acta* **1984**, *66*, 1. (b) Gresh, N. *J. Comput. Chem.* **1995**, *16*, 856. (c) Piquemal, J.-P.; Gresh, N.; Giessner-Prettre, C. *J. Phys. Chem A* **2003**, *107*, 10353. (d) Piquemal, J.-P.; Chevreau, H.; Gresh, N. *J. Chem. Theory Comput.* **2007**, *3*, 824.

(22) (a) Gresh, N. *J. Chim.-Phys. Chim. Biol.* **1997**, *94*, 1365. (b) Gresh, N.; Guo, H.; Kafafi, S. A.; Salahub, D. R.; Roques, B. P. *J. Am. Chem. Soc.* **1999**, *121*, 7885. (c) Gresh, N. *Curr. Pharm. Des* **2006**, *12*, 2121. (d) Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. *J. Chem. Theory Comput.* **2007**, *3*, 1960.

(23) Vigné-Maeder, F.; Claverie, P. *J. Chem. Phys.* **1988**, *88*, 4934.

(24) Garmer, D. R.; Stevens, W. J. *J. Phys. Chem. A* **1989**, *93*, 8263.

(25) Evangelakis, G. A.; Rizos, J. P.; Lagaris, I. E.; Demetropoulos, I. N. *Comput. Phys. Commun.* **1987**, *46*, 401.

(26) (a) Tiraboschi, G.; Gresh, N.; Giessner-Prettre, C.; Pedersen, L. G.; Deerfield, D. W. *J. Comput. Chem.* **2000**, *21*, 1011. (b) Tiraboschi, G.; Roques, B. P.; Gresh, N. *J. Comput. Chem.* **1999**, *20*, 1379. (c) Piquemal, J.-P.; Chelli, R.; Procacci, P.; Gresh, N. *J. Phys. Chem. A* **2007**, *111*, 8170.

(27) Kitaura, K.; Morokuma, K. *Int. J. Quantum Chem.* **1976**, *10*, 325.

(28) Piquemal, J.-P.; Cisneros, G. A.; Reinhardt, P.; Gresh, N.; Darden, T. A. *J. Chem. Phys.* **2006**, *124*, 104101.

(29) (a) Antony, J.; Piquemal, J.-P.; Gresh, N. *J. Comput. Chem.* **2005**, *26*, 1131. (b) Roux, C.; Gresh, N.; Perera, L.; Piquemal, J.-P.; Salmon, L. *J. Comput. Chem.* **2007**, *28*, 938.

(30) Jenkins, L. M. M.; Hara, T.; Durell, S. R.; Hayashi, R.; Inman, J. K.; Piquemal, J.-P.; Gresh, N.; Appella, E. *J. Am. Chem. Soc.* **2007**, *129*, 11067.

(31) Sponer, J.; Sabat, M.; Burda, J.; Leszczynski, J.; Hobza, P. *J. Phys. Chem B* **1999**, *103*, 2528.

(32) Chalasinski, G.; Szczesniak, M. M. *Chem. Rev.* **1994**, *94*, 1723.

(33) Creuzet, S.; Langlet, J.; Gresh, N. *J. Chim.-Phys. Phys. Chim. Biol.* **1991**, *88*, 2399.

(34) Hobza, P.; Sponer, J. *Chem. Rev.* **1999**, *99*, 3247.

(35) Garmer, D. R.; Gresh, N. J. *Am. Chem. Soc.* **1994**, *116*, 3556.

(36) Rayon, V. M.; Valdes, H.; Diaz, N.; Suarez, D. K. *J. Chem. Theory Comput.* **2008**, *4*, 243.

(37) Bagus, P. S.; Hermann, K.; Bauschlicher, C. W. *J. Chem. Phys.* **1984**, *80*, 4378.

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

(39) Hobza, P.; Sponer, J. J. *Am. Chem. Soc.* **2002**, *124*, 11802.

(40) (a) Basch, H.; Stevens, W. J. *J. Mol. Struct. (THEOCHEM)* **1995**, *338*, 303. (b) Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 4177. (c) Cubero, E.; Lucque, F. J.; Orozco, M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 9576. (d) Dehez, F.; Angyan, J. G.; Gutierrez, I. S.; Luque, F. J.; Schulten, K.; Chipot, C. *J. Chem. Theory. Comput.* **2007**, *3*, 1914.

(41) Gresh, N.; Pullman, B. *Biochim. Biophys. Acta* **1980**, *625*, 356.

(42) (a) Ryde, U.; Olsen, L.; Nilsson, K. *J. Comput. Chem.* **2002**, *23*, 1058. (b) Ryde, U.; Nillson, K. *J. Am. Chem. Soc.* **2003**, *125*, 14232. (c) Rulisek, L.; Ryde, U. *J. Phys. Chem. B* **2006**, *110*, 11511.

(43) (a) Guillot, B.; Muzet, N.; Artacho, E.; Lecomte, C.; Jelsch, C. *J. Phys. Chem. B* **2003**, *107*, 9109. (b) Muzet, N.; Guillot, B.; Jelsch, C.; Howard, E.; Lecomte, C. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 8742. (c) Jelsch, C.; Guillot, B.; Lagoutte, B.; Lecomte, C. *J. Appl. Crystallogr.* **2005**, *38*, 38.

(44) Volkov, A.; Coppens, P. *J. Comput. Chem.* **2004**, *25*, 921.

CT800200J

JCTC Journal of Chemical Theory and Computation

# A Vulnerability in Popular Molecular Dynamics Packages Concerning Langevin and Andersen Dynamics

David S. Cerutti,*,† Robert Duke,‡,§ Peter L. Freddolino,|| Hao Fan,⊥ and Terry P. Lybrand†

*Center for Structural Biology, Department of Chemistry, Vanderbilt University, 5140 Medical Research Building III, 465 21st Avenue South, Nashville, Tennessee 37232-8725, Department of Chemistry, University of North Carolina, Campus Box 3290, Chapel Hill, North Carolina 27599-0001, Laboratory of Structural Biology, National Institute of Environmental Health Science, Research Triangle Park, 12 Davis Drive, Chapel Hill, North Carolina 27709-5900, Center for Biophysics and Computational Biology, University of Illinois, 156 Davenport Hall, 607 South Mathews Avenue, Urbana, Illinois 61801-3635, and Department of Biopharmaceutical Sciences, University of California, Byers Hall, 1700 Fourth Street, Suite 501, San Francisco, California 94158-2330*

Received June 10, 2008

**Abstract:** We report a serious problem associated with a number of current implementations of Andersen and Langevin dynamics algorithms. When long simulations are run in many segments, it is sometimes possible to have a repeating sequence of pseudorandom numbers enter the calcuation. We show that, if the sequence repeats rapidly, the resulting artifacts can quickly denature biomolecules and are then easily detectable. However, if the sequence repeats less frequently, the artifacts become subtle and easily overlooked. We derive a formula for the underlying cause of artifacts in the case of the Langevin thermostat, and find it vanishes slowly as the inverse square root of the number of time steps per simulation segment. Numerous examples of simulation artifacts are presented, including dissociation of a tetrameric protein after 110 ns of dynamics, reductions in atomic fluctuations for a small protein in implicit solvent, altered thermodynamic properties of a box of water molecules, and changes in the transition free energies between dihedral angle conformations. Finally, in the case of strong thermocoupling, we link the observed artifacts to previous work in nonlinear dynamics and show that it is possible to drive a 20-residue, implicitly solvated protein into periodic trajectories if the thermostat is not used properly. Our findings should help other investigators re-evaluate simulations that may have been corrupted and obtain more accurate results.

## Introduction

Molecular simulations of proteins and other complex biomolecules are performed routinely in atomic detail for tens of nanoseconds. A variety of thermodynamic ensembles are available for these simulations, but in virtually all cases, investigators wish to see the dynamics of a system at a particular temperature, corresponding to a Maxwell distribution of momenta for the particles of the molecular model. In simulations of complex biomolecules, the systems typically contain enough inhomogeneity that complete equilibration across all degrees of freedom is not possible over currently achievable simulation timescales, meaning that potential energy will tend to be released as structures relax. This, in addition to the slow but inevitable increase of energy

* To whom correspondence should be addressed. Phone: (615) 936-3569. Fax: (615) 936-2211. E-mail: david.cerutti@vanderbilt.edu.

† Vanderbilt University.

‡ University of North Carolina.

§ National Institute of Environmental Health Science.

|| University of Illinois at Urbana-Champaign.

⊥ University of California, San Francisco.

10.1021/ct8002173 CCC: $40.75 
Published on Web 09/10/2008

in the system because of the finite time steps taken to propagate the dynamics, leads to an upward drift in the system temperature as the simulation continues. Algorithms, such as SHAKE,[1] which apply constraints to a finite degree of precision, can also add to or even dissipate the system's energy, leading to more temperature drift.

To run simulations on the timescales needed to model chemical processes, a number of algorithms have been developed to maintain a specified system temperature. These include velocity rescaling approaches such as the Berendsen[2] and Nose-Hoover[3] thermostats and velocity modification approaches such as the Andersen[4] and Langevin thermostats.[5] In Andersen thermocoupling, particle velocities are periodically reassigned to pseudorandom values so that the resulting momenta follow a Maxwell distribution at the desired temperature. In the Langevin scheme, velocities of the particles in the simulations are modified with pseudorandom forces as if they were undergoing stochastic collisions with imaginary particles whose momenta follow a Maxwell distribution at the desired temperature.

The importance of generating long, decorrelated sequences of random numbers for accurate simulations has been discussed before,[6,7] and modern molecular dynamics codes use algorithms[8,9] that can generate sequences so long that they would be unlikely to repeat over the course of a simulation even if millions of particles were simulated for trillions of time steps (for example, if a virus capsid were simulated at atomic detail for several milliseconds).

However, modern molecular dynamics codes also offer a large number of options for managing simulations, and it is difficult to anticipate all the permutations of how those options might be used. Long simulations can generate tens of gigabytes of trajectory data and take weeks or months to complete. For this reason, checkpoint files are nearly always used to store the positions and velocities of atoms so that the simulation may be broken up into small segments that make it feasible to run on managed computing resources and easy to recover from a machine crash. However, in several popular molecular dynamics packages, the checkpoint files do not contain information on the state of the random number generator. In such cases, reuse of the same random number generator seed causes a finite sequence of random numbers to appear in every simulation segment. As will be shown, these repeating sequences of random numbers can drastically affect simulations using either Langevin or Andersen thermostats if the simulation segments are short; the effects can be subtle but significant if the segments are longer.

To help determine when this issue may produce significant problems in typical simulations, we have quantified the effects of repeating sequences of pseudorandom numbers in several test systems, including two explicitly solvated proteins, and surveyed existing codes to see which packages are vulnerable. We also show that simply incrementing the random number seed with each simulation segment effectively removes the artifacts.

## Theory

The effects of repeating sequences of pseudorandom numbers are straightforward to describe in the case of the Langevin thermostat, as we show in the following formalism. We expect that similar principles hold for the Andersen thermostat.

The Langevin thermostat maintains a desired temperature by application of a friction force with coefficient $\zeta$ and a random force $\mathbf{R}_i$ to all particles $i$ to simulate random collisions between particles in the simulation and imaginary particles in an external bath held at temperature $T$. In this framework, collisions with particle $i$ occur at a frequency $\gamma_i$

$$\gamma_i = \frac{\zeta}{m_i} \tag{1}$$

such that the central equation of motion is

$$\dot{\mathbf{p}}_i = \mathbf{f}_i - \gamma_i \mathbf{p}_i + \mathbf{R}_i \tag{2}$$

For real water, $\gamma$ has a value of roughly 50 $ps^{-1}$. In simulations, smaller values of 2−5 $ps^{-1}$ are typically used,[10] although some investigators have found that the full 50 $ps^{-1}$ gives better results.[11] The random force is related to $\gamma$ by eq 3

$$\langle \mathbf{R}(0)\mathbf{R}(t) \rangle = 2m_i k_b T \gamma_i \delta(t) \tag{3}$$

where the angular brackets denote an ensemble average, $k_b$ is Boltzmann's constant, and $\delta(t)$ is the Dirac delta function. By eq 3, the components of the instantaneous random force vector at time $t$ follow a Gaussian distribution with zero mean and variance $2m_i\gamma_i k_b T$, henceforth denoted $\sigma^2_{\mathbf{R}}$.

To understand the effect of repeating random number sequences on molecular dynamics simulations, we consider $\mathbf{\Psi}(N)$, the "residual" force on a particle causesed by Langevin collisions after $N$ steps of simulation. Each of the three components of $\mathbf{\Psi}$ can be expressed as

$$\mathbf{\Psi}_\alpha(N) = \langle \mathbf{R} \rangle = \frac{1}{N}\sum_{s=1}^{N} \mathbf{R}_s \tag{4}$$

where $\alpha$ represents $x$, $y$, or $z$. $\mathbf{\Psi}$ is an average rather than a sum because each of the random forces is only applied during one of the $N$ steps. By the Central Limit Theorem,[12] the distribution of $\mathbf{\Psi}$ is also Gaussian with zero mean and variance $\sigma^2_{\mathbf{R}}/N$. Therefore, the magnitude of each component of the residual force on an atom after $N$ Langevin dynamics steps of length $\Delta t$ can be expressed as

$$\mathbf{\Psi}_\alpha(N) = \sqrt{\frac{2m_i\gamma_i k_b T}{N\Delta t}} \tag{5}$$

If the same sequences of $N$ pseudorandom forces are used repeatedly in a Langevin dynamics simulation, each atom is exposed to a finite number of forces and therefore a nonvanishing residual force. Over many iterations, this is similar to applying a constant force to each particle, in a particular direction relative to the axes of the simulation cell, with a magnitude given by eq 5. The expected and observed magnitudes of residual forces for a Langevin thermostat with collision frequency 3 $ps^{-1}$ and a bath temperature of 298 K are plotted as a function of $N$ in Figure 1. (Observations of the residual forces were made with a modified version of the AMBER9 PMEMD software, available upon request.)

Although Figure 1 clearly shows significant residual forces acting on each atom even for lengthy simulation segments, the forces do not act in any concerted fashion (see Figure

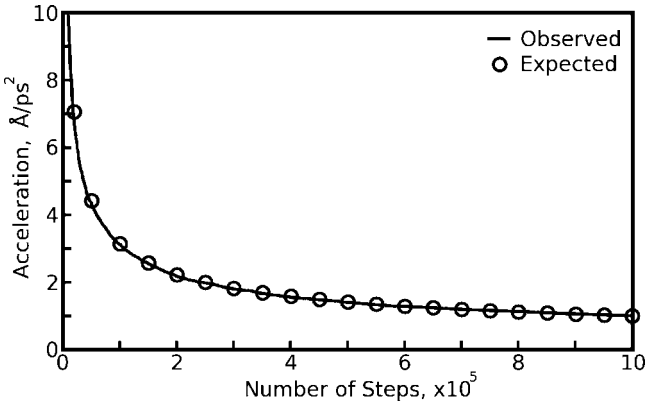



***Figure 1.*** Residual accelerations on atoms observed in Langevin dynamics. Langevin forces on individual atoms were summed over steps of a molecular dynamics run of the Trp-Cage miniprotein using collision frequency 3 ps$^{-1}$ and a bath temperature of 298 K. Averaging the forces over $N$ previous steps gives a value for the residual force on that atom, a quantity which tends to zero as $1/\sqrt{N}$. Residual forces on each atom were normalized by the atom's mass to give accelerations. The black line shows average residual acceleration for all atoms; circles show the values expected from eq 5.

S1 of the Supporting Information), and so, their overall effects must be determined by simulations. As we will show in the Results, these residual forces quickly give rise to severe artifacts when short simulation segments are used, but subtle artifacts can occur with greater segment lengths, such as those investigators might use in practice.

## Methods

Proteins for molecular dynamics simulations were obtained from the Protein Data Bank (PDB).[13] All proteins were protonated using the TLEAP module of AMBER9[14] and modeled using the AMBER ff99 force field,[15,16] with improvements suggested by Simmerling et al.[17] SPC/E water[18] was used for simulations in explicit solvent. The Generalized Born (GB) model of Onufriev et al.[19] combined with the LCPO pairwise surface area approximation[20] was used for simulations in implicit solvent. The PMEMD and SANDER modules of AMBER9 were used for simulations in explicit and implicit solvent, respectively.

Molecular dynamics simulations in explicit solvent were initiated by adjusting positions of added water molecules with 2000 steps of steepest-descent energy minimization, while restraining the positions of protein atoms, then performing similar energy minimization of the protein atoms with the solvent held fixed, and finally running energy minimization of the entire system with no restraints. Energy minimization of the protein was also done prior to implicit solvent simulations. Equilibration dynamics in all simulations were performed at a constant temperature of 298 K using a Langevin thermostat with a collision frequency of 3.0 ps$^{-1}$ (unless otherwise stated, this temperature and collision frequency were used in all simulations in this study). Position restraints were initially used to limit the motion of all heavy atoms; the restraints were gradually relaxed over a period of 500 ps. For simulations in explicit solvent, periodic boundary conditions were applied, and the simulation volume was held constant until the final stages of equilibration, when dynamics were continued in the constant-pressure ensemble. For implicit solvent calculations, no boundary conditions were used. The equilibration phase typically involved about ten restarts; different random seeds were used to initialize the pseudorandom number generator with each restart.

Force calculations for all stages of dynamics in explicit solvent were performed with a 9.0 Å cutoff on real-space interactions, particle-mesh Ewald electrostatics,[21] and Lennard-Jones tail corrections. Force calculations in implicit solvent were performed with no cutoff on real-space interactions and a 25 Å cutoff on calculations of the Born radii. The SHAKE algorithm[1] was used to constrain all bonds including hydrogen on protein atoms, and the SETTLE algorithm[22] was used to constrain the internal geometry of explicit water molecules. A time step of 1.5 fs was used for all production dynamics.

## Results

**Langevin Artifacts in Explicit Solvent.** We first became aware of the danger of repeating random number sequences when we noticed that the apostreptavidin tetramer (PDB accession code 1SWA) was relatively stable in explicit solvent when dynamics were propagated at 100 000 steps (150 ps) per segment but rapidly unfolded when dynamics were propagated at 1000 steps per segment. A Langevin thermostat with a collision frequency of 3 ps$^{-1}$ had been used to maintain the temperature at 298 K, and the same random seed had been provided to initialize the pseudorandom number generator (PRNG) in all cases. To quantify the protein destabilization effect, Figure 2 shows backbone root-mean-squared deviation (rmsd) results, taking the equilibrated conformation of the protein as a reference, for a series of 15 ns simulations of the tetramer using different segment lengths. When the PRNG is repeatedly initialized with the same seed, the segment length corresponds to the parameter $N$ as discussed in Theory. For comparison, a nonrepeating sequence of Langevin forces was generated by running the same simulation with 100 000 steps per segment and changing the PRNG seed with each restart. To demonstrate that the protein is destabilized by repeating sequences of Langevin forces and not some problem with restarting a simulation from a checkpoint file, we performed a 6 ns simulation with 1000 steps per segment, incrementing the PRNG seed with each restart. The results in Figure S3 of the Supporting Information show that this method also results in stable dynamics.

The streptavidin tetramer is a dimer of dimers.[23] The two dimers are each more stable than the tetramer as a whole, as demonstrated by the existence of dimeric streptavidin mutants[24] and the mechanism of tetramer stabilization by biotin binding.[25] For this reason, we tracked backbone rmsd not just for the tetramer but also for its dimer components. As shown in Figure 2, individual dimers maintained their original backbone conformations better than the tetramer as a whole under cycles of repeating sequences of Langevin forces. rmsd for the individual monomers is not shown, but it closely parallels the dimer rmsd.

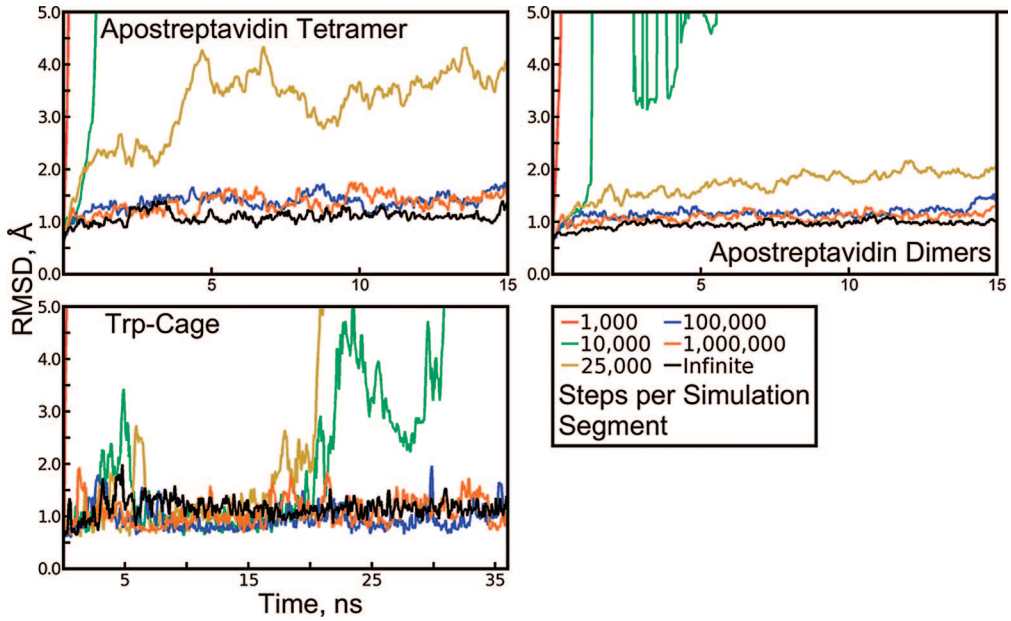



***Figure 2.*** Backbone rmsd of apostreptavidin and Trp-Cage, revealing artifacts in Langevin dynamics. Each protein was simulated in explicit solvent at 298 K using a Langevin thermostat with a collision frequency of 3 $ps^{-1}$ and simulation segments with lengths given in the figure legend. The same random seed was used to reinitialize the pseudorandom number generator (PRNG) at the beginning of every segment, except for the "infinite" case, in which segments of 100 000 steps were initiated with different PRNG seeds every time.

If very long sequences of repeating Langevin forces (100 000 and 1 000 000 steps per segment) are used, artifacts are difficult to detect in simulations of only 15 ns. Before we became aware of the problem with Langevin dynamics, a simulation of the apostreptavidin tetramer was carried out for 145 ns, using 100 000 and 1 000 000 step segments at different times but with the same PRNG seed in all cases. Backbone rmsds for monomers, dimers, and the tetramer in this system are shown in Figure 3. At a glance, the system appears to behave reasonably, except for the dissociation of the tetramer at 110 ns.

The apostreptavidin tetramer is known to be highly stable, even in concentrated urea,[26] so the dissociation seen in Figure 3 and in Figure S2 of the Supporting Information is not realistic. But because the tetramer is known to be stabilized by biotin binding[27] and because we had a 250 ns simulation showing the biotin-liganded tetramer to be stable in solution (data not shown), we initially believed that the dissociation of the unliganded tetramer was qualitatively correct. However, inspection of the rmsd for portions of the trajectory run with 1 000 000 versus 100 000 steps per segment suggests that over a very long simulation the tetramer can be destabilized by 100 000 step segments with identical PRNG seeds in much the same way that shorter segments destabilize it more quickly. Indeed, with sequences of 100 000 pseudorandom Langevin forces acting on each atom, the residual forces described in equ 5 would have been half as strong as those obtained with sequences of 25 000 Langevin forces, which created artifacts immediately.

To further investigate the extent of these artifacts, we conducted 36 ns simulations of the 20-residue Trp-Cage miniprotein in explicit solvent and subjected the system to repeating sequences of Langevin forces in the same manner as was done with the 500-residue apostreptavidin tetramer.

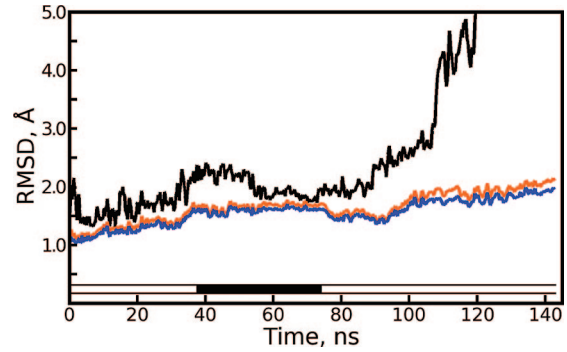



***Figure 3.*** Long-time scale Langevin dynamics of apostreptavidin tetramer under repeating sequences of Langevin forces. Root-mean-sqaured deviation (rmsd) of the tetramer (black line), average dimer rmsd (orange line), and average monomer rmsd (blue line) are plotted over 145 ns. The bar just above the *x*-axis is solid black when 1 000 000 step segments were used and white when 100 000 step segments were used. The period of simulation using the longer segments shows a slight reduction in the rmsd of the tetramer and stable RMSDs for dimers and monomers. In contrast, all of these rmsd values steadily increase, particularly that of the tetramer, when segments of 100 000 steps are used. Figure S2 of the Supporting Information illustrates the tetramer dissociation at 110 ns.

The results in Figure 2 show that Trp-Cage also unfolds under rapidly repeating Langevin forces, but remains stable if the Langevin thermostat is used correctly. Notably, whereas residual forces from a repeating sequence of 25 000 Langevin forces caused some instability in the apostreptavidin system, residual forces of the same magnitude denatured the Trp-Cage miniprotein. Moreover, under repeating sequences of 10 000−25 000 Langevin forces, Trp-Cage appeared to be stable for 17−20 ns before suddenly unfolding.

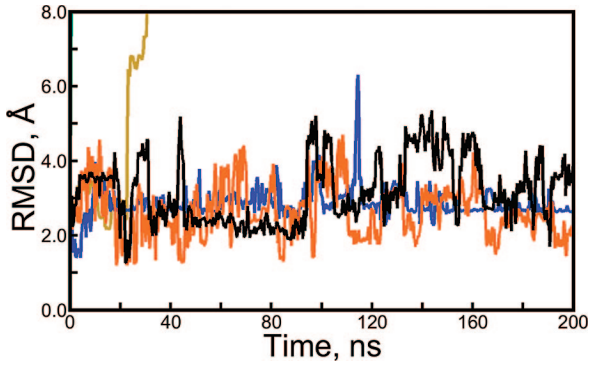



***Figure 4.*** Backbone rmsd of the Trp-Cage miniprotein revealing artifacts in Langevin dynamics. The Trp-Cage miniprotein was simulated in Generalized Born solvent using a Langevin thermostat. Simulations with different segment lengths are plotted in different colors following the legend in Figure 2. Simulations with 1000 and 10 000 steps per segment unfolded within 2 ns and, so, are not visible on the plot.

Still, the extent of artifacts in simulations using 100 000 or 1 000 000 steps per segment remains uncertain. Because the residual forces do not act in a concerted fashion, their effects may be more pronounced on local features of the protein structure. We therefore computed atomic root-mean-squared (rms) fluctuations for backbone atoms over the final 30 ns of each simulation as shown in Figure 5. Error bars were determined by computing rms fluctuations over four 7.5 ns subintervals and taking the standard deviation. In explicit solvent, the computed fluctuations do not differ greatly if the thermostat applies sequences of 100 000, 1 000 000, or an infinite number of Langevin forces. Furthermore, there is no apparent trend in the data; atomic fluctuations for nearly all backbone atoms increase slightly if repeating sequences of 1 000 000 Langevin forces instead of 100 000 are used, but they decrease again if an infinite sequence of Langevin forces is used. The amount of sampling in 36 ns of dynamics is rather small, however. In the next section, we sample protein conformations over longer time scales with a Langevin thermostat and an accelerated dynamics method.

**Langevin Artifacts in Implicit Solvent.** Although a Langevin thermostat can be used with explicitly solvated systems, it is more commonly used to simulate stochastic collisions with imaginary solvent particles in an implicitly solvated system. We therefore conducted simulations of the Trp-Cage miniprotein[28] in Generalized Born (GB) solvent. Because the system is so small (300 atoms versus 8000 for the explicitly solvated Trp-Cage versus 40 000 for the explicitly solvated apostreptavidin tetramer), we were able to obtain very long (200 ns) simulations and more convergent estimates of atomic fluctuations.

As shown in Figure 4, the Trp-Cage miniprotein explores conformations with larger backbone rmsd relative to the native state in GB implicit solvent as opposed to SPC/E explicit solvent. Again, with repeating sequences of 100 000 or more Langevin forces acting on each atom, Trp-Cage is stable, but with fewer sequences, it becomes denatured quickly.

Atomic fluctuations for backbone C atoms obtained from the final 180 ns of trajectories with 100 000 and 1 000 000 step segments using repeating random seeds are compared to those obtained from a trajectory generated with constantly changing random seeds in Figure 5. As before, error bars were created by splitting the data into four 45 ns segments and computing standard deviations. In implicit solvent, the atomic fluctuations generally increase as the length of the repeating sequence of random forces goes from 100 000 to infinity. Because more than six times as many conformations were used to calculate these fluctuations, the results are somewhat more certain than those from the explicit solvent simulations. Although the error bars look larger in the implicit solvent case, as a fraction of the corresponding fluctuations the error bars in implicit solvent are in fact roughly two times smaller. While short repeating sequences of Langevin forces acting on each atom denatured the protein, sequences of 100 000 forces appeared to reduce its mobility relative to much longer ones. This apparent contradiction may be explained by looking at the backbone rmsd obtained for shorter repeating sequences of Langevin forces, as shown in Figure S4 of the Supporting Information. In such cases, the rmsd may climb to very high values, but then hovers around particular values for extended periods of time, suggesting that the denatured conformations do not fluctuate very much.

The atomic fluctuations only appear to diminish in the absence of explicit solvent particles, however (see Figure

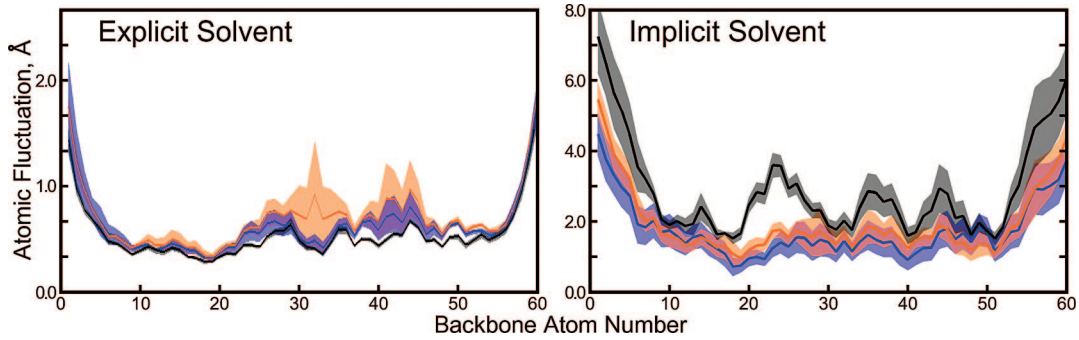



***Figure 5.*** Atomic fluctuations of Trp-Cage backbone atoms in two solvent environments. The numbering of atoms on the *x*-axis proceeds as (residue 1) N, CA, C, (residue 2) N, CA, C,..., (residue 20) N, CA, C. Fluctuations for simulations with sequences of $10^5$, $10^6$, and an infinite number of Langevin forces are shown as the blue, orange, and black lines, respectively. Error bars are given in the same colors as partially transparent regions surrounding each line. Simulations in explicit solvent were run for 36 ns, and simulations in implicit solvent were run for 200 ns. Note that the *y*-axis has a different scale in each panel.

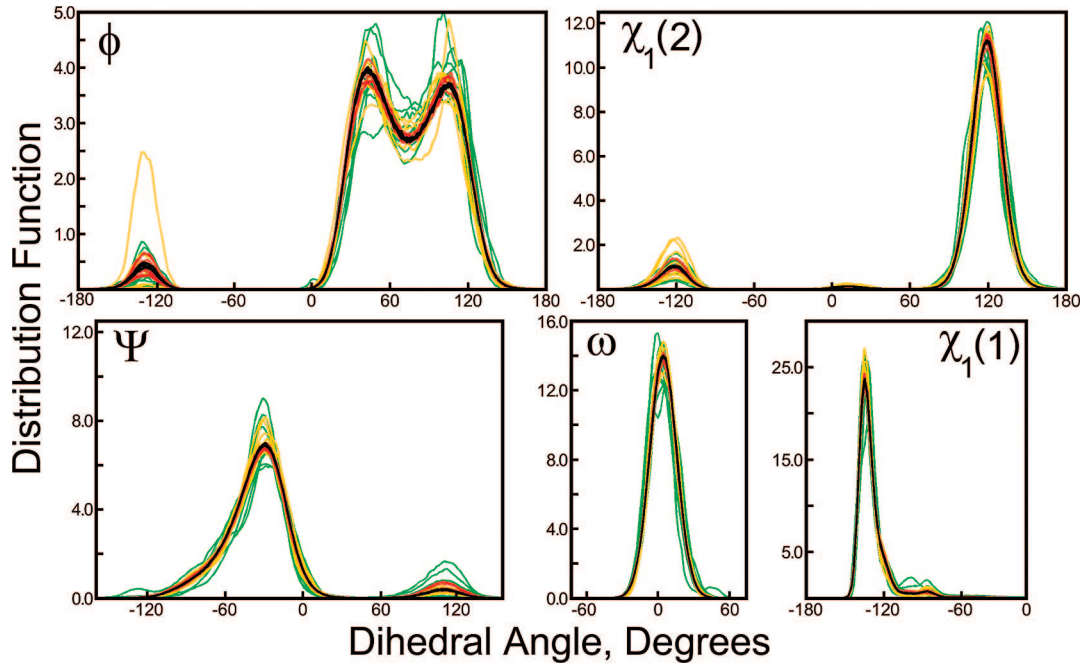



***Figure 6.*** Distributions of five dihedral angles in the seryl-serine system under finite sequences of Langevin forces. Eight independent trajectories of the seryl-serine system were computed with 25 000 (green lines), 100 000 (yellow lines), 1 000 000 (red lines), and an infinite sequence of Langevin forces (black lines) acting on each atom. The distributions above are normalized by the expected population of each dihedral angle value if the potential energy surface were completely flat. The distributions obtained for infinite sequences of Langevin forces are mutually convergent, demonstrating the thoroughness of the sampling from these 1000 ns simulations. However, finite sequences of Langevin forces tend to perturb the distributions. These perturbations are quantified in terms of transition free energies in Table 1.

S5, Supporting Information). When solvent is represented explicitly, denatured protein structures tend to fluctuate more even after the native conformation is lost. This dichotomy likely arises as the individual water molecules can migrate to different regions of the protein even if subjected to repeating sequences of Langevin forces. (Indeed, as will be discussed later in the Results, if the sequences are very short, the water molecules are all being propelled in particular directions and the polypeptides are literally showered with rapidly moving water molecules.) These solvent interactions impart instability on the polypeptide motion, increasing the atomic fluctuations, whereas in implicit solvent the polypeptide moves only according to the Langevin forces acting on its own atoms and thereby becomes trapped in a particular conformation.

The dissociation of the apostreptavidin tetramer over very long simulations in explicit solvent and reduced atomic fluctuations of the Trp-Cage miniprotein in implicit solvent give indications that simulations run with repeating sequences of 100 000 Langevin forces are not safe from artifacts. However, different PRNG seeds will create unique repeating sequences of Langevin forces that will affect the system in different ways, whereas our results thus far have shown the effects of only one sequence of a given length on each system tested. To precisely quantify the microscopic effects of residual forces as a function of the simulation segment length, we needed to be able to thoroughly sample the entire conformational space of a system and run many simulations with different sequences of Langevin forces.

For this purpose, we chose to study the seryl-serine peptide in implicit solvent. Because the serine side-chain is so small, residual forces acting on it will not be averaged over many atoms, and therefore, its $\chi_1$ angle should be very prone to reorientation due to these forces. Eight independent simulations of 1 $\mu$s were done using segments of 25 000, 100 000, and 1 000 000 steps with repeating random seeds, as well as segments of 100 000 steps with changing random seeds. Distributions of $\chi_1$ angles for each serine side-chain, as well as three backbone dihedral angles, are shown in Figure 6. While nonrepeating sequences of Langevin forces consistently generate the same distribution for each of the dihedral angles, unique repeating sequences of Langevin forces each impart their own bias on the system, causing the distribution of dihedral angles to converge differently in each case. As expected, the distortions grow larger as the simulation segment length decreases.

The thoroughness of equilibrium sampling in the seryl-serine system permitted direct calculation of transition free energies, $\Delta G$, by comparing the probability of finding each of the five dihedral angles at two values $\nu$ and $\eta$ in the unbiased ensemble (the trajectory computed with a nonrepeating sequence of Langevin forces). We also computed changes in the transition free energies, $\Delta\Delta G$, between the unbiased ensemble and each of the biased ensembles generated with finite sequences of Langevin forces. These quantities are defined mathematically as

$$\Delta G = RT \ln\left(\frac{P(\nu)}{P(\eta)}\right) \tag{6}$$

$$\Delta\Delta G = RT\left[\ln\left(\frac{P(\nu, \text{biased})}{P(\eta, \text{biased})}\right) - \ln\left(\frac{P(\nu, \text{unbiased})}{P(\eta, \text{unbiased})}\right)\right] \tag{7}$$

In the above equations, $T$ represents the temperature (298

***Table 1.*** Free Energies for Transitions between Two Values, $\nu$ and $\eta$, of Various Dihedral Angles in the Seryl-Serine System[a]

| dihedral angle | $\nu$ (deg) | $\eta$ (deg) | $\langle \Delta G \rangle$ | $\langle \lvert \Delta\Delta G \rvert \rangle$ ($2.5 \times 10^4$) | $\langle \lvert \Delta\Delta G \rvert \rangle$ ($10^5$) | $\langle \lvert \Delta\Delta G \rvert \rangle$ ($10^6$) |
|---|---|---|---|---|---|---|
| residue 1, $\chi_1$ | −123 | −76 | $-2.07 \pm 0.01$ | $0.37 \pm 0.25$ | $0.28 \pm 0.18$ | $0.10 \pm 0.07$ |
| residue 2, $\chi_2$ | 120 | −120 | $-1.42 \pm 0.02$ | $0.31 \pm 0.16$ | $0.32 \pm 0.17$ | $0.07 \pm 0.07$ |
| residue 1, $\psi$ | 107 | −29 | $1.72 \pm 0.04$ | $0.47 \pm 0.31$ | $0.20 \pm 0.11$ | $0.14 \pm 0.14$ |
| residue 2, $\phi$ | 43 | 105 | $-0.04 \pm 0.01$ | $0.11 \pm 0.09$ | $0.08 \pm 0.08$ | $0.03 \pm 0.02$ |
| residue 2, $\phi$ | 105 | 231 | $-1.27 \pm 0.05$ | $0.77 \pm 0.50$ | $0.70 \pm 0.50$ | $0.20 \pm 0.11$ |
| residue 2, $\phi$ | 43 | 231 | $-1.31 \pm 0.06$ | $0.74 \pm 0.48$ | $0.69 \pm 0.52$ | $0.21 \pm 0.12$ |

[a] The chosen values of $\nu$ and $\eta$ correspond to relative maxima indentified in the unbiased ensemble (see Figure 6) generated with an infinite sequence of Langevin forces. ΔG values, reported with standard deviations in kcal/mol, refer to the transition free energy for the unbiased ensemble (see eq 6); angular brackets ⟨⟩ refer to averages from eight independent trajectories. Similarly, |ΔΔG| values refer to absolute changes in the transition free energy if a finite sequence of Langevin forces (of length specified in parentheses) is used (see eq 7).

K), and $R$ represents the gas constant. Results from this analysis are given in Table 1. The table only reports average values of $\Delta\Delta G$, but individual cases showed changes in the transition free energies in excess of 1 kcal/mol for some of the biased ensembles obtained with 100 000 steps per segment. Contrary to our expectations, the largest $\Delta\Delta G$ values were obtained in the backbone $\phi$ angle of the second residue; residual forces on many atoms exert torques about this dihedral, yet the distortion resulting from an average of all these torques remains large. Although the distributions of each dihedral angle in the unbiased ensemble may not be totally accurate, the computed values of $\Delta G$ and $\Delta\Delta G$ provide precise measurements of the degree to which simulations using finite sequences of Langevin forces are biased, as well as the degree of bias present in short simulations (40 ps to 1.5 ns) using Langevin dynamics.

**Severity of Artifacts As a Function of the Langevin Collision Frequency.** As was predicted in Theory and shown in the preceding results, the severity of artifacts from the Langevin thermostat diminishes as the length of the repeating seqeunce of pseudorandom forces grows. However, by eq 5, the magnitude of residual forces and thus the severity of artifacts is also proportional to the square root of the collision frequency $\gamma_i$, and different values of this parameter have been used in the past.[10,11] We therefore repeated some of the simulations of Trp-Cage in implicit solvent with $\gamma_i$ set to 50 $ps^{-1}$ rather than 3 $ps^{-1}$. By eq 5, we would expect the higher collision frequency to increase the average residual force on each atom roughly by a factor of 4. With a collision frequency of 3 $ps^{-1}$, a segment length of 6000 steps would be needed to obtain residual forces of comparable magnitude (this was verified with the modified AMBER9 PMEMD code used to generate Figure 1).

The results in Figure 7 confirm that, even with relatively long 100 000 step segments, the 50 $ps^{-1}$ Langevin collision frequency can generate a striking artifact when combined with repeating sequences of pseudorandom forces. As indicated by the system's convergent backbone rmsd, the Trp-Cage miniprotein is driven to a very small set of structures under these conditions. Examination of the checkpoint files from each segment of the simulation shows that, within 12 ns, the coordinates and velocities are converged to $1.0 \times 10^{-7}$ Å and $1.0 \times 10^{-7}$ Å $ps^{-1}$, respectively, and the trajectory segments are identical thereafter. Although it was surprising to obtain periodic behavior over such long (150 ps) intervals in such a complex system, we also observed periodic behavior for 50 000 step segments and

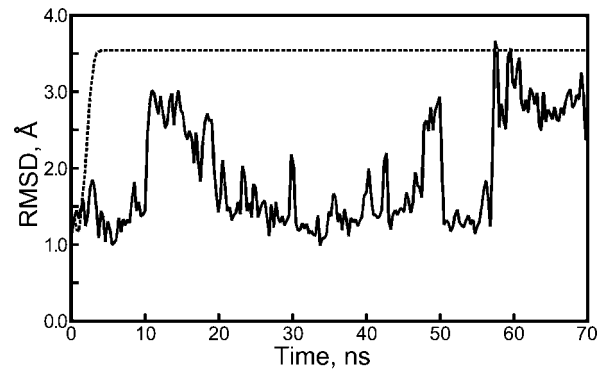



***Figure 7.*** Backbone rmsd of the Trp-Cage miniprotein during Langevin dynamics with strong thermocoupling. When each segment is initiated with the same random seed (dashed line), the repeating sequence of 100 000 Langevin forces drives the protein into a periodic trajectory (see Results, Severity of Artifacts As a Function of the Langevin Collision Frequency section). No such behavior is seen if an infinite sequence of Langevin forces is used instead (solid line).

200 000 step segments (data not shown). Periodicity was not observed in the trajectory if the random seed was changed with each restart (see Figure 7) or if a collision frequency of 3 $ps^{-1}$ was used (see Figure 4).

These observations led us to consider the possibility that the periodic behavior observed with strong thermocoupling was related to the protein unfolding seen in previous sections. If so, the fact that the strongly thermocoupled Trp-Cage system run in long segments did not unfold to the same extent as the weakly thermocoupled Trp-Cage system run in short segments (see Figure 4) needed further investigation. We emphasize that, as discussed in the preceding Langevin Artifacts in Implicit Solvent section, different repeating sequences of Langevin forces may drive the system into different conformations, and it is conceivable that occasionally these conformations would fall close to the native state. We therefore ran three additional simulations with 100 000 steps per segment, repeating random seeds, and a collision frequency of 50 $ps^{-1}$. All trajectories eventually became periodic, but the time to obtain this behavior varied for each different sequence of Langevin forces (see Figure S6 of the Supporting Information), and the length of the period was five simulation segments, rather than just one, in one of the cases. Although each periodic trajectory displayed a different level of backbone rmsd relative to the native state, all of the backbone RMSDs were much lower than the backbone

RMSDs eventually seen in similar runs with repeating sequences of 10 000−25 000 Langevin forces (see Figure 4).

In summary, a Langevin thermostat with a collision frequency of 50 $ps^{-1}$ drove the Trp-cage miniprotein into periodic trajectories 100 000−500 000 steps long. Under such strong thermocoupling, repeating sequences of Langevin forces did not denature the system to the extent seen before, but a periodic trajectory does represent an extreme restriction of the protein's conformational space.

**Artifacts in a Simulation of Pure Water.** With an explicitly solvated protein system, the motions of atoms in the protein are tightly coupled, but the motions of solvent particles are not. In the previous section, we tested the effects of repeating sequences of Langevin forces if the system contains only the tightly coupled degrees of freedom; conversely, we can look for artifacts in the thermodynamic properties of a system containing many small, unconnected particles.

Multiple 6 ns simulations of a box of 512 SPC/E water molecules were conducted at 1 atm pressure and 298 K using a Langevin thermostat (collision frequency 3 $ps^{-1}$). Constant random seeds were used to restart the simulations in segments ranging from 250−16 000 steps, and four independent simulations were conducted using unique random seeds for each segment length. As shown in Figure 8, the density, heat of vaporization, and heat capacity of SPC/E water all change noticeably for segments with fewer than 4000 steps. In such simulations, one cannot obtain a convergent value of the diffusion coefficient because every water molecule suffers a net displacement along a particular direction during each segment. However, compared to the artifacts observed in solvated proteins, the density, heat of vaporization, and heat capacity of water are not very sensitive to Langevin artifacts.

**Artifacts Created by Repeating Random Number Sequences with the Andersen Thermostat.** Although we did not provide a formal description of the way repeating sequences of velocity reassignments could create artifacts if the system temperature is maintained by an Andersen thermostat, we expected that this would have similar effects to applying repeating sequences of forces. An array of 15 ns simulations was carried out for the apostreptavidin tetramer with the same repeating random seeds and segment lengths as in the case of the Langevin thermostat. Results are shown in Figure 9. As before, the use of the Andersen thermostat with a repeating PRNG seed destabilized the tetramer, indicating that the Andersen thermostat can create artifacts in much the same manner as the Langevin thermostat.

Although the severity of the artifacts appear to be smaller in terms of backbone rmsd than the artifacts created by Langevin dynamics with similar segment lengths, we stress that the strength of thermocoupling in each thermostat is determined differently and that this can also influence the severity of artifacts (see Severity of Artifacts As a Function of the Langevin Collision Frequency section). We did not try to match the degree of thermocoupling in the Andersen dynamics simulations with that used in our other explicit solvent simulations.

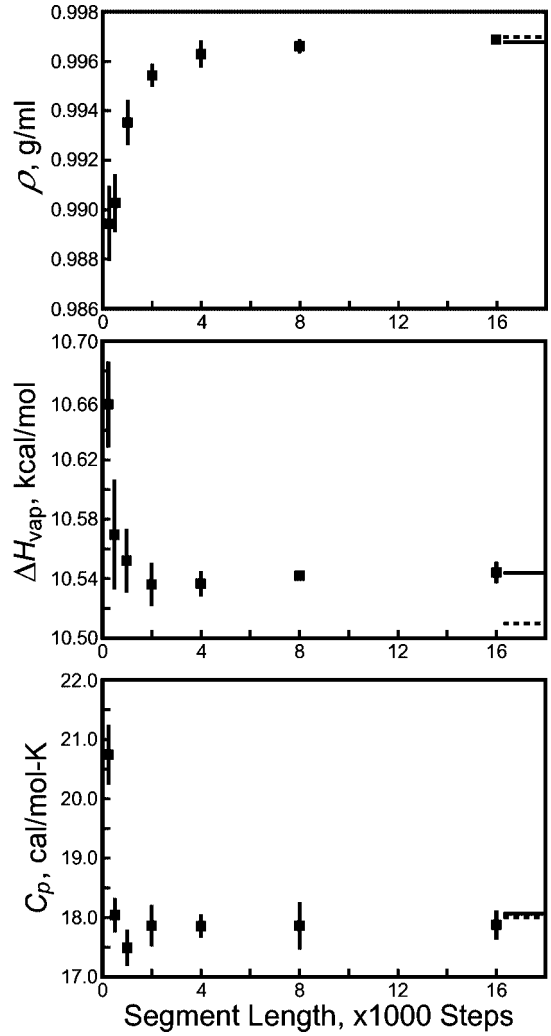



***Figure 8.*** Thermodynamic properties of a box of 512 SPC/E water molecules revealing Langevin artifacts. Formulas for the density ($\rho$), heat of vaporization ($\Delta H_{vap}$), and heat capacity ($C_p$) can be found in work by Jorgensen and Jenson[43] (note that the polarization energy correction[18] is invoked in computing $\Delta H_{vap}$). Solid black lines extending from the right border indicate the values of each quantity if a nonrepeating sequence of Langevin forces is used; dashed lines indicate experimental results for water at 298 K. Error bars are obtained from four indpendent simulations.

## Discussion

**Common Features of Artifacts Resulting from Repeating Random Number Sequences.** In the Results, we identified a number of abnormal behaviors that can be observed in systems run with repeating sequences of Langevin forces. Most of the backbone root-mean-squared deviation (rmsd) artifacts can be explained as consequences of residual forces, which decay slowly as a function of the length of the sequence of Langevin forces as shown in Figure 1. Together, these residual forces do not act in any concerted fashion, but individually they do act in a particular direction relative to the coordinate axes of the simulation box. Each atom of the protein is therefore forced in a unique random

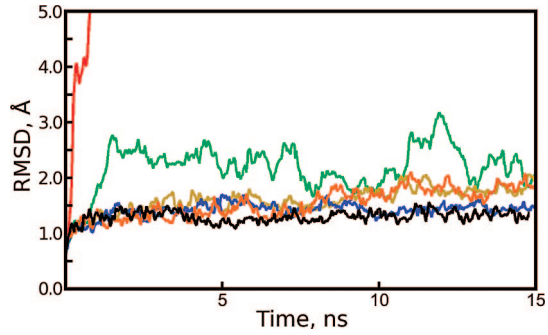



***Figure 9.*** Backbone rmsd of the apostreptavidin tetramer revealing artifacts in Andersen dynamics. The apostreptavidin tetramer was simulated in explicit solvent with repeating sequences of Andersen velocity reassignments. The legend in Figure 2 indicates the length of segments in each simulation; velocity reassignment occurred every 1000 steps (e.g., the red line presents backbone rmsd of the tetramer when all atoms are reassigned to the same set of velocities every 1000 steps).

direction, and the protein becomes distorted until the forces on each atom are counterbalanced by gradients of the system's potential energy function. Weak residual forces, such as those encountered with 100 000 steps and a collision frequency of 3 $ps^{-1}$, appear to be enough to break apart globular domains along their weak interfaces (see Figure S2, Supporting Information), but stronger residual forces can denature the domains themselves (see Figures 2 and 4), regardless of the type of solvent used. Similar artifacts obtained with the Andersen thermostat (see Figure 9) are likely the products of "residual momenta."

Initially, it would seem that the relative positions of larger groups of atoms would be less prone to artifacts than smaller groups of atoms because the residual forces acting on individual atoms would be averaged such that the net force pulling two groups of atoms apart would be small. However, the larger $\Delta\Delta G$ values observed for the backbone $\phi$ angle in Figure 6 and the separation of the apostreptavidin tetramer seen in Figure 3 do not support this reasoning. Instead, because each atom of a rigid molecular structure has a different moment arm about some center of rotation, the residual forces on just a few atoms could be amplified, creating the large $\Delta\Delta G$ values between populations of certain dihedral angles and the hinge-bending motion of the apostreptavidin tetramer dissociation (see Figure S2 of the Supporting Information).

In the Theory section, we stated that repeatedly applying a finite sequence of pseudorandom forces to an atom was similar to applying a constant net force on that atom. However, a more precise description is needed to explain the periodicity of trajectories observed in Results, Severity of Artifacts As a Function of the Langevin Collision Frequency section, and the differences in atomic fluctuations observed in Results, Langevin Artifacts in Implicit Solvent section. Separate trajectories initiated from distinct conformations of a system have been observed to synchronize if identical sequences of pseudorandom noise are used to propagate Langevin dynamics.[29] This synchronization occurs after the trajectories remain uncorrelated for some amount of time, the length of which depends on the strength of the pseudorandom noise. In the examples given throughout the Results, the checkpoint files written at the end of each segment of a simulation provide distinct conformations of the system, and the collision frequency tunes the strength of the noise. In the Results, Severity of Artifacts As a Function of the Langevin Collision Frequency section, the 100 000-step segments of the trajectory become synchronized as identical sequences of strong pseudorandom noise are repeatedly applied. This offers an explanation of how synchronization of successive trajectory segments could occur in as little as 12 ns with $\gamma$ set to 50 $ps^{-1}$ but not in 200 ns if $\gamma$ is set to 3 $ps^{-1}$.

An earlier work by Fahy and Hamann[30] performed similar calculations on small systems driven with a rudimentary Andersen-like thermostat. In this work, they noted the existence of a critical length of time between velocity reassignments, $\tau_c$, such that reassigning velocities more frequently resulted in synchronization of the trajectories and reassigning them less frequently resulted in indefinite chaotic behavior. Noting that $\tau_c$ corresponds to the strength of thermocoupling in the Anderson thermostat, we can hypothesize that there exists some critical strength of thermocoupling in the Langevin thermostat above which synchronization of trajectories is guaranteed and below which chaotic behavior will be observed. This is consistent with our results, and knowledge of the value of $\tau_c$ or equivalent $\gamma_c$ could help investigators make better choices about how to maintain the temperature of a simulation. However, more studies would be necessary to estimate these critical thresholds for different system sizes and topologies.

On the basis of the above observations, we may extend our description of the artifacts created by repeating sequences of Langevin forces or Andersen velocity reassignments in molecular dynamics simulations and state it loosely as follows: *Thermostats operating with repeating finite sequences of random noise will cause incoherent perturbations in a system's potential energy surface, the strength of the perturbations being inversely proportional to the square root of the length of the noise sequence and directly proportional to the square root of the strength of the noise itself. The incoherent distortions tend to reduce the conformational space available to the system; in the limit of strong noise, the system may be driven into periodic trajectories according to the unique sequence of noise applied.*

Unfortunately, the artifacts caused by repeating sequences of Langevin forces or Andersen velocity reassignments seem to be very extensive because of the way the residual forces scale with the sequence length. Many published simulations could potentially have been affected; the results in this study show that, over very long simulations, some observables such as atomic fluctuations in implicit solvent display artifacts if a finite sequence of even 1 000 000 Langevin forces is used to control the temperature. Artifacts in backbone rmsd measurements may be detectable if a repeating sequence of 100 000 Langevin forces is used. With sequences of fewer than 100 000 Langevin forces, the artifacts may take tens of nanoseconds to appear, but they are often dramatic. We would like to offer a general statement such as "simulations

performed with sequences of 1 000 000 or more Langevin forces and a weak thermocoupling of 3 $ps^{-1}$ or less are safe from artifacts," but certain analyses other than those presented in this study may be more sensitive to thermostat artifacts.

**Survey of Current Molecular Dynamics Packages with Respect to Random Number Generation.** The potential for artificially distorting a biomolecule by incorrect use of the Langevin or Andersen thermostats represents a serious problem for molecular simulations. This prompted us to make a brief survey of existing molecular dynamics packages to see which implementations could allow users to unwittingly perturb their systems with repeating sequences of pseudorandom numbers. The most robust protection against the artifacts identified in the Results is to pass the state of the random number generator through the molecular dynamics checkpoint files and, by default, to override user-specified random seeds when restarting a molecular dynamics calculation. In this manner, the pseudorandom number generator (PRNG) would produce a single sequence for the entire simulation.

As stated in the results, we discovered this problem while running Langevin dynamics with the AMBER9 software package.[14] By default, both of its simulation modules use a random seed of 71277, and users may specify other values. The state of the PRNG is not passed via the checkpoint file, however, so Langevin and Andersen dynamics simulations are prone to artifacts unless the user specifically requests that the random seed be set using the clock time, changes the random seed with a script running outside of the AMBER software, or performs simulations in very long segments. Similarly, the GROMACS (version 3.*)[31−34] software runs with a default random seed of 1993 and does not pass the state of the PRNG through its checkpoint files, but users may request that the seed be set using the clock time. Tests with the GROMACS software presented in the Supporting Information confirm that artifacts can be generated in the same manner as was shown for the AMBER code throughout the Results. Robust protection against random number artifacts will be implemented in future versions of both AMBER and GROMACS.

In the DL_POLY package (version 3),[35] the random seed is set at compile time, although if segments of a Langevin or stochastic dynamics simulation are run in parallel on a varying number of processors, different series of pseudorandom numbers will be generated.

The NAMD code[36] is highly resistant to Langevin artifacts because of the manner in which it generates random numbers. By default, the PRNG seed is set by the clock time, although users may set it to a specific value. The state of the PRNG is not passed through the checkpoint file, but some unique aspects of the NAMD code offer added protection against random number artifacts (see the Supporting Information). These aspects of the code make it very difficult to obtain such artifacts with NAMD.

The CHARMM[37] and DESMOND[38] packages both implement the robust solution by passing the state of the PRNG through simulation checkpoint files and, so, can be considered safe from the artifacts identified in this study.

**Future Directions: Thermostats for Molecular Dynamics Calculations.** We have shown that, if used correctly, the Langevin thermostat produces stable dynamics for explicitly solvated proteins over tens of nanoseconds. Other studies[39] have reported stable dynamics for tens to hundreds of nanoseconds using the Berendsen "weak-coupling" approach to temperature regulation. However, the results in Figure 1 suggest that a sufficiently large number of random forces (or, by extension, velocity reassignments) must be applied to each atom to ensure that a thermostat based on random numbers has not applied significant net forces or momenta to the individual atoms of the system. Furthermore, data in Results, Severity of Artifacts As a Function of the Langevin Collision Frequency section, corroborate the findings of Ciesla and co-workers,[29] suggesting that separate trajectories of large molecular systems can become synchronized if both simulations are run with the same sequence of Langevin forces, even if that sequence is infinite. Investigators should therefore carefully consider the manner in which their thermostat functions, beyond simple qualifications such as stable dynamics.

All molecular dynamics thermostats attempt to simulate coupling of the system to an external bath at the desired temperature, but none of the methods are entirely physically meaningful. In reality, "heating" and "cooling" refer to the equilibration of the momenta of particles in two systems brought into contact with one another. In common biomolecular simulations with explicit solvent, the solvent is typically an accessory, while the analysis focuses on the biomolecule itself. It may therefore be desirable to modify existing thermocoupling schemes to regulate only the temperature of the solvent, or perhaps only the temperature of solvent particles further than some minimum distance from the biomolecule. This method, similar in spirit to stochastic dynamics,[40,41] would regulate the temperature of the biomolecule indirectly, hopefully causing very little perturbation to its dynamics. Other modifications of simple thermostats[42] should also be considered.

The goal of this study was to expose a serious problem associated with the use of Langevin and Andersen thermostats in molecular simulations and to present in detail possible artifacts that might arise. However, the results from simulations with strong thermocoupling raise questions about the way thermostats affect the dynamical properties of biomolecular models, and whether it would be helpful to modify current thermocoupling schemes as increasing computational resources make it possible to study kinetic properties of events such as protein folding or ligand binding through simulations.

**Acknowledgment.** This research was supported by National Institutes of Health Grant GM080214. D.S.C. thanks the GROMACS development team, Dr. Bernard R. Brooks, Dr. Justin Gullingsrud, and Dr. Ilian Todorov for explanations about the operation of GROMACS, CHARMM, DESMOND, and DL POLY, respectively. Dr. Robert Konecny and Dr. Barrett Abel of the Center for Theoretical Biophysics at the University of California, San Diego, provided part of the computing resources for this study via National Science Foundation Grant PHY-0216576.

**Supporting Information Available:** Illustration of residual Langevin forces, illustration of the dissociation of the apostreptavidin tetramer, demonstration that rapidly restarting simulations with a changing pseudorandom number generator (PRNG) seed permits stable dynamics, clarification of the fact that Langevin artifacts tend to lead to higher protein backbone rmsd relative to the native state but lower atomic fluctuations thereafter, further demonstration of periodic trajectories obtained with PRNG artifacts and strong thermocoupling, and illustrations of Langevin dynamics artifacts in simulations performed with the GROMACS and NAMD codes. This material is available free of charge via the Internet at http://pubs.acs.org.

## References


(1) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C.; Hirasawa, K. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **1997**, *23*, 327–341.

(2) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(3) Hoover, W. G. Canonical dynamics: Equilibrium phase−space distributions. *Phys. Rev. A.* **1985**, *31*, 1695–1697.

(4) Andersen, H. C. Molecular dynamics at constant pressure and/or temperature. *J. Chem. Phys.* **1980**, *72*, 2384–2393.

(5) Izaguirre, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D. Langevin stabilization of molecular dynamics. *J. Chem. Phys.* **2001**, *114*, 2090–2098.

(6) Ferrenberg, A. M.; Landau, D. P.; Wong, J. Y. Monte Carlo simulations: Hidden errors from "good" random number generators. *Phys. Rev. Lett.* **1992**, *69*, 3382–3384.

(7) Vattulainen, I. Framework for testing random numbers in parallel calculations. *Phys. Rev. E.* **1999**, *59*, 7200–7204.

(8) Holian, B. L.; Percus, O. E.; Warnock, T. T.; Whitlock, P. A. Pseudorandom number generator for massively parallel molecular-dynamics simulations. *Phys. Rev. E.* **1994**, *50*, 1607–1615.

(9) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. Random Numbers. In *Numerical Recipes: The Art of Scientific Computing*, 3rd ed.; Cambridge University Press: Cambridge, U.K., 1989; pp 380−385.

(10) Wu, X.; Brooks, B. R. Self-guided Langevin dynamics simulation method. *Chem. Phys. Lett.* **2003**, *381*, 512–518.

(11) Fan, H.; Mark, A. E.; Zhu, J.; Honig, B. Comparative study of generalized Born models: Protein dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6760–6764.

(12) Feller, W. The Fundamental Limit Theorems in Probability. *Bull. Amer. Math. Soc.* **1945**, *51*, 800–832.

(13) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(14) Case, D. A.; Cheatham, T. E.; Darden, T. A.; Gohlke, H.; Luor, R.; Merz, M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. The AMBER biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(15) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(16) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second-generation force field for the simulations of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(17) Simmerling, C.; Strockbine, B.; Roitberg, A. E. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **2002**, *124*, 11258–11259.

(18) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(19) Onufriev, A.; Bashford, D.; Case, D. A. Modification of the generalized Born model suitable for macromolecules. *J. Phys. Chem. B.* **2000**, *104*, 3712–3720.

(20) Weiser, J.; Shenkin, P. S.; Still, W. C. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J. Comput. Chem.* **1999**, *20*, 217–230.

(21) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. H. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(22) Miyamoto, S.; Kollman, P. A. SETTLE: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952–962.

(23) Reznik, G. O.; Vajda, S.; Smith, C.; Cantor, C. R.; Sano, T. Streptavidins with intersubunit crosslinks have enhanced stability. *Nat. Biotechnol.* **1996**, *14*, 1007–1011.

(24) Pazy, Y.; Eisenberg-Domovich, Y.; Laitinen, O. H.; Kulomaa, M. S.; Bayer, E. A.; Wilchek, M.; Livnah, O. Dimer−tetramer transition between solution and crystalline states of streptavidin and avidin mutants. *J. Bacteriol.* **2003**, *185*, 4050–4056.

(25) Katz, B. A. Binding of biotin to streptavidin stabilizes intersubunit salt bridges between Asp61 and His87 at low pH. *J. Mol. Biol.* **1997**, *274*, 776–800.

(26) Kurzban, G. P.; Bayer, E. A.; Wilcheck, M.; Horowitz, P. M. The quaternary structure of streptavidin in urea. *J. Biol. Chem.* **1991**, *266*, 14470–14477.

(27) Sano, T.; Cantor, C. R. Cooperative biotin binding by streptavidin. *J. Biol. Chem.* **1990**, *265*, 3369–3373.

(28) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. Designing a 20-residue protein. *Nat. Struct. Biol.* **2002**, *9*, 425–430.

(29) Ciesla, M.; Dias, S. P.; Longa, L.; Oliveira, F. A. Synchronization induced by Langevin dynamics. *Phys. Rev. E.* **2001**, *63*, 065202.

(30) Fahy, S.; Hamann, D. R. Transition from chaotic to nonchaotic behavior in randomly driven systems. *Phys. Rev. Lett.* **1992**, *69*, 761–764.

(31) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing molecular dynamics implementation. *Comput. Phys. Commun.* **1995**, *91*, 43–45.

(32) Daura, X.; Oliva, B.; Querol, E.; Aviles, F. X.; Tapia, O. On the sensitivity of MD trajectories to changes in water-protein interaction parameters: The potato carboxypeptidase inhibitor in water as a test case for the GROMOS force field. *Proteins* **1996**, *25*, 89–103.

(33) Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: A package for molecular simulations and trajectory analysis. *J. Mol. Model.* **2001**, *7*, 306–317.

(34) van der Spoel, D.; van Buuren, A. R.; Tieleman, P.; Berendsen, H. J. C. Molecular dynamics simulations of peptides from BPTI: A closer look at amide-aromatic interactions. *J. Biomol. NMR.* **1996**, *8*, 229–238.

(35) Smith, W.; Yong, C. W.; Rodger, P. M. DL POLY: application to molecular simulation. *Mol. Simulat.* **2002**, *28*, 385–471.

(36) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kal, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(37) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

(38) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; and Shaw, D. E. Scalable algorithms for molecular dynamics simulations on commodity clusters In *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*; San Jose, CA, December 3−6, 2006; Buhyan, L. Ed.; Association for Computing Machinery, Inc.: New York, 2006.

(39) Lei, H.; Duan, Y. Two-stage folding of HP-35 from ab-initio simulations. *J. Mol. Biol.* **2007**, *370*, 196–206.

(40) Berkowitz, M.; McCammon, J. A. Molecular dynamics with stochastic boundary conditions. *Chem. Phys. Lett.* **1982**, *90*, 215–217.

(41) Brünger, A.; Brooks, C. L., III; Karplus, M. Stochastic boundary conditions for molecular dynamics simulations of ST2 water. *Chem. Phys. Lett.* **1984**, *105*, 495–500.

(42) Koopman, E. A.; Lowe, C. P. Advantages of a Lowe−Andersen thermostat in molecular dynamics simulations. *J. Chem. Phys.* **2006**, *124*, 204103.

(43) Jorgensen, W. L.; Jenson, C. Temperature dependence of TIP3P, SPC, and TIP4P water from NPT Monte Carlo simulations: Seeking temperatures of maximum density. *J. Comput. Chem.* **1998**, *19*, 1179–1186.

CT8002173

JCTC Journal of Chemical Theory and Computation

# Donor−Acceptor Dissociation Energies of Group 13−15 Donor−Acceptor Complexes Containing Fluorinated Substituents: Approximate Lewis Acidities of $(F_3C)_3M$ vs $(F_5C_6)_3M$ and the Effects of Phosphine Steric Bulk

Austin L. Gille and Thomas M. Gilbert*

*Department of Chemistry and Biochemistry, Northern Illinois University, DeKalb, Illinois 60115*

Received May 24, 2008

**Abstract:** To study donor−acceptor complexes containing fluoroalkyl and -aryl substituents on the acceptors, ONIOM methods for optimizing large complexes and determining single point energies were tested. A two-layer ONIOM optimization procedure utilizing the MPW1K model followed by single point calculations using the composite three-layer ONIOM G2R3 method proved acceptable. The optimization model predicts M-X bond distances well when compared to experiment and shows that the distances increase discontinuously with the bulk of the phosphine. Unexpectedly, $(R_F)_3B\text{-}XR_3$ and $(R_F)_3Al\text{-}XR_3$ bond dissociation energies ($\Delta E_{DA}$) are comparable for several R substituents. For $R_F = CF_3$, both are predicted to exhibit M-X $\Delta E_{DA}$ values in the range 55−80 kcal mol$^{-1}$, exceptionally strong for dative bond energies. For $R_F = C_6F_5$, the $\Delta E_{DA}$ values are predicted to lie in the range 30−45 kcal mol$^{-1}$. $(F_5C_6)_3BP(t\text{-}Bu)_3$, which does not contain a B−P bond, is predicted to display $\Delta E_{DA} = 19$ kcal mol$^{-1}$. The $\Delta E_{DA}$ energies do not change smoothly as the steric bulk of the phosphine increases. However, intrinsic $\Delta E_{DA}$ energies $\Delta E_{int}$ show a regular increase as the donor ability of the phosphine increases, confirming that the reorganization energy of the individual moieties contributes sizably to the overall $\Delta E_{DA}$. The data indicate that $PPh_3$ is approximately equivalent to $PMe_3$ as a donor in terms of $\Delta E_{int}$.

## Introduction

Main group Lewis acid−base complexes remain of interest as archetypal tests of bonding theories but have recently gained new emphasis owing to Stephan's recent report[1] of the heterolytic cleavage of $H_2$ by the "frustrated Lewis pairs"[2] (FLPs) $(F_5C_6)_3BPR_3$ (R = *t*-Bu, mesityl) and $Ph_3BP(t\text{-}Bu)_3$. The steric bulk of the substituents provides the key to this reactivity, apparently by prohibiting formation of strong B−P bonds. The fluorines in the former complexes also seem to play a role, as computational studies suggest that $(F_5C_6)_3BP(t\text{-}Bu)_3$ is held together to some extent by intramolecular F...H interactions across the BP core.[3] Such effects have been observed experimentally and studied computationally for a range of $(F_5C_6)_3M\text{-}XR_3$ systems (M = B, Al; X = N, P).[4−6] It is plausible that interactions between the arene $\pi$ clouds and the *t*-Bu hydrogens act similarly in $Ph_3BP(t\text{-}Bu)_3$.

Despite these extensive studies, in general the factors that determine M-X bond dissociation energies ($\Delta E_{DA}$) in these and related fluorinated systems have not been systematically examined. Indeed, $\Delta E_{DA}$ values and trends have been examined only sparsely.[3,4] We are particularly interested in determining $\Delta E_{DA}$ values and trends for complexes where $(F_3C)_3M$ acts as the acceptor, as experiments suggest $(F_3C)_3B\text{-}NR_3$ complexes exhibit exceptionally strong B−N bonds.[7] For example, $(CF_3)_3B\text{-}NH(CH_3)_2$ forms through treatment of $[(CF_3)_3B\text{-}N(CH_3)_2]^-$ with excess concentrated HCl;[8] $(CF_3)_3B\text{-}NH(CH_2CH_3)_2$ converts to $(CF_3)_3B\text{-}NH_3$ upon treatment with excess $KOH/Br_2/H_2O$.[9] Structural and computational studies[10] show that the substituents on B and N adopt conformations intermediate between staggered and

* Corresponding author e-mail: tgilbert@niu.edu.

10.1021/ct8001859 CCC: $40.75 
Published on Web 09/23/2008

***Table 1.*** M-X Bond Distances (Å) and $\Delta E_{DA}$(raw)[17] Values (kcal mol$^{-1}$) Using Different Models, Basis Sets, and ONIOM Layer Sizes (OLS)

| | model | basis set | OLS | distance | $\Delta E_{DA}$(raw) |
|---|---|---|---|---|---|
| $(F_5C_6)_3BPMe_3$ | MP2 | 6−311++G(d,p) | 1 | 2.046 | 34 |
| | MP2 | 6−31+G(d) | 1 | 2.052 | 31 |
| | MPW1K | 6−311++G(d,p) | 1 | 2.022 | 38 |
| | MPW1K | 6−31+G(d) | 1 | 2.026 | 38 |
| | MPW1K | 6311++G(d,p) | 2 | 2.066 | 28 |
| | MPW1K | 631+G(d) | 2 | 2.068 | 26 |
| | MPW1K | 6311++G(d,p) | 3 | 2.069 | 28 |
| | MPW1K | 6311+G(d) | 3 | 2.069 | 28 |
| | MPW1K | 631+G(d,p) | 3 | 2.072 | 26 |
| | MPW1K | 6−31+G(d) | 3 | 2.072 | 26 |
| $(F_3C)_3AlPPh_3$ | MP2 | 6−311++G(d,p) | 1 | 2.401 | 59 |
| | MP2 | 6−31+G(d) | 1 | 2.403 | 59 |
| | MPW1K | 6−311++G(d,p) | 1 | 2.380 | 62 |
| | MPW1K | 631+G(d) | 1 | 2.382 | 63 |
| | MPW1K | 6311++G(d,p) | 2 | 2.444 | 51 |
| | MPW1K | 6−311+G(d) | 2 | 2.444 | 51 |
| | MPW1K | 6−31+G(d,p) | 2 | 2.443 | 51 |
| | MPW1K | 6−31+G(d) | 2 | 2.443 | 51 |
| | MPW1K | 6−311+G(d) | 3 | 2.461 | 46 |
| | MPW1K | 6−31+G(d) | 3 | 2.462 | 45 |

eclipsed, suggesting that intramolecular F...H interactions are less important than in the pentafluorophenyl analogues.

We published model tests on bond dissociation energies of alkyl- and fluoroalkyl borane amine complexes,[10] showing that the MP2 model gave results closest to experiment and that DFT models other than the MPW1K approach generally performed poorly. In accord with the experiments noted above, we predicted $\Delta E_{DA}$ = 68 kcal mol$^{-1}$ for $(CF_3)_3B$-$NMe_3$, an exceptionally large value for a dative bond by any measure and almost four times the value for $Me_3B$-$NMe_3$. This, combined with the steric effects noted above, motivated us to examine the structures and dissociation energies for fluorinated borane acceptor−donor complexes where the donor bulk was increased systematically. From a computational chemistry standpoint, we felt that the MP2 model is too resource-intensive to apply to molecules as large as $(F_5C_6)_3BP(t\text{-}Bu)_3$, so we wanted to examine ONIOM-based approaches that would allow study of such complexes. We report here a variety of tests leading to ONIOM G2R3-based $\Delta E_{DA}$ values for the series $(R_F)_3M$-$XR'_3$ ($R_F$ = $F_3C$, $F_5C_6$; M = B, Al; X = N, P; R = Me, Et, *i*-Pr, *t*-Bu, Ph). Our testing efforts suggest that the values predicted are likely to be acceptably accurate. The data provide a means of quantifying the effect of steric bulk and of the Lewis acidity of the tris(fluoroalkyl)borane moieties on these donor−acceptor systems and point to candidates for experimental study.

## Computational Methods and Tests

All calculations were performed with the Gaussian 98 (G98) suite of programs.[11] As the molecules studied are far too large to study with our computational resources, we followed the approach suggested by Vreven and Morokuma for study of $\Delta E_{DA}$ of $Ph_3C$-$CPh_3$.[12] This involves the usual methodology of a composite method, such as the $G_n$ models:[13] optimize to an acceptably accurate structure and then use this for vibrational/temperature corrections and for single point energy calculations using perturbation theory models of increasing sophistication and increasingly large basis sets. Vreven and Morokuma's contribution showed that ONIOM methods could be used to provide both an acceptable structure and the single point energies. Their three-layer ONIOM G2R approach for the single point energies, based on two-layer ONIOM B3LYP optimized structures, gave $\Delta E_{CC}$ within 1 kcal mol$^{-1}$ of the experimental value for $Ph_3C$-$CH_3$ and suggested the quite reasonable value of $\Delta E_{CC}$ = 16 kcal mol$^{-1}$ for $Ph_3C$-$CPh_3$.

As this method has not been broadly applied yet, and as our systems involve breaking dative bonds rather than homolytic ones, we felt it necessary to test it against datively bound systems. Testing required determining the effects of three variables. The model and basis set size affect only the optimization part of the method; the ONIOM layer space affects both the optimization and the single point calculations. Thus both structural parameters and $\Delta E_{DA}$ values were examined as test indicators. The complexes $(F_5C_6)_3B$-$PMe_3$ and $(F_3C)_3Al$-$PPh_3$ were selected for testing. These are representative of the spectrum of complexes we wished to examine because they (a) have fluorinated substituents on the acceptor moieties, (b) have the two group 13 atoms, and (c) have phenyl rings either on the donor or acceptor, which provide the most challenging part of using the ONIOM approach (see below).

The structures of the two complexes were optimized without constraints using a two level ONIOM approach, either the (MPW1K/various basis sets: MPW1K/3−21G) combination or the (MP2/various basis sets:HF/3−21G) combination (Table 1). The MP2[14] perturbation theory model was used as coded in G98, while the MPW1K model was generated using IOp keywords.[15] We selected the DFT MPW1K model, rather than the B3LYP model used by Vreven and Morokuma,[12] since we showed previously that it outperforms B3LYP significantly for both optimizations and energies of donor−acceptor complexes, particularly those where the donor and acceptor are trialkyl substituted.[10] Moreover, since the MPW1K model was designed to model transition state energies, it is plausible that it models weakly

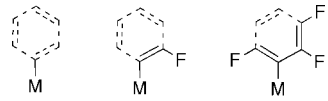



***Figure 1.*** ONIOM layer sizes examined for phenyl/pentafluorophenyl rings. Fluorine atoms included in the high layer for pentafluorophenyl rings are shown for each layer size.

bound transition state structures well. This could prove important in modeling the structures of weakly bound donor−acceptor complexes. In many cases, the optimization procedure did not readily minimize the four parameters G98 uses to determine structural convergence. This problem has been noted previously[16] and is apparently more common when DFT methods are employed in the ONIOM approach.[12] We halted the optimization in such cases when 20 consecutive steps failed to change the predicted energy by more than 1 kcal $mol^{-1}$. Usually this corresponded to a point where the forces had converged but the displacements had not.

Selection of a proper layer size is a key aspect of using ONIOM. For optimizations of $(F_5C_6)_3B$-$PMe_3$ and $(F_3C)_3Al$-$PPh_3$, the Group 13/15 core atoms and the carbon atoms bound to them were always placed in the high layer; the fluorine atoms of trifluoromethyl substituents were placed there as well. Hydrogens of methyl groups were always placed in the low layer. Following Vreven and Morokuma,[12a] we examined three options for the layering of atoms of phenyl/pentafluorophenyl rings, shown in Figure 1. One can view these as the "hydrogen" level (OLS 1), the "vinyl" level (OLS 2), and the "2-butadienyl" level (OLS 3), respectively. Specifically, in OLS 1, only the core group 13/15 atoms are placed in the high level; all other atoms are added to the low level. In OLS 2, the core atoms are joined in the high layer by the phenyl ipso carbons, the phenyl ortho carbons that are oriented over the donor−acceptor axis, and the fluorines or hydrogens attached to these carbons. These selections were made so that the fluorines/hydrogens most likely to participate in intramolecular interactions (because they point toward the other half of the complex) were treated in the most extensive way available. The OLS 3 layering system works similarly save that both phenyl ortho carbons and one meta carbon, with their associated substituent atoms, are placed in the high layer. OLS 1 is the least complete but sufficiently small so that one can employ the MP2/6−311++G(d,p) model chemistry on the high level (Table 1); OLS 3 is the most complete but quite resource-intensive for $C_6F_5$ rings, even with the DFT model and the relatively small 6−31+G(d) basis set.

On the basis of internal consistency, the data for $(F_5C_6)_3B$-$PMe_3$ and $(F_3C)_3Al$-$PPh_3$ suggest that the OLS 1 approach performs poorly for structure and $\Delta E_{DA}$ prediction (Table 1).[17] The B−P and Al−P distances are 0.4−0.8 Å shorter, and the donor and acceptor are overbound by 8−10 kcal $mol^{-1}$, as compared with values predicted using more expansive layer sizes. Moreover, this approach leaves the fluorine atoms in the low layer. We felt it likely that studying intramolecular F...H interactions correctly would require having the fluorine atoms that are oriented over the donor−acceptor axis in the high layer (which holds for OLS 2 and 3). Thus, we excluded the OLS 1 model for further calculations. This necessitated excluding the ONIOM MP2:HF combination as well, as this pairing is too resource-intensive to employ at the OLS 2 level. We discarded the OLS 3 model for the opposite reason: using it provides no significant improvement either in the bond distances or $\Delta E_{DA}$ values over the OLS 2 approach.

Similarly, the data in Table 1 show that using large basis sets such as 6−311++G(d,p) or 6−311+G(d) in the high level may provide slightly more accurate structures than the smaller 6−31+G(d) basis set but does not predict significantly different $\Delta E_{DA}$ values. We therefore performed all subsequent optimizations using the ONIOM (OLS 2) (MPW1K/6−31+G(d):MPW1K/3−21G) approach. In addition to the layering choices already described, we opted to model alkyl substituents larger than methyl as methyl groups. For example, for *t*-Bu, the tertiary carbon atom only was placed in the high layer, while the primary carbons and hydrogens were placed in the low layer.

Having selected an ONIOM optimization method, we adopted Vreven and Morokuma's[12] three-layer ONIOM G2R approach (hereafter called OG2R3) for final single point energy calculations. The OG2R3 composite approximates a CCSD(T)/6−311+G(2df,2p) calculation using fifteen separate energy determinations (several of which are duplicates) to determine a final molecular energy; however, all are of sizes that are compatible with typical computational resources and require only a few hours maximum per calculation. The overall OG2R3 energy is calculated as $\Delta E_{DA}$ (OG2R3) = $\Delta E_{CCSD(T)} + \Delta E_{MP2large} - \Delta E_{MP2small}$, where $\Delta E_{CCSD(T)}$ is the energy from a three-layer ONIOM calculation symbolized as ONIOM(CCSD(T)/6−31G(d):MP2/6−31G(d):B3LYP/3−21G), $\Delta E_{MP2large}$ is the energy from a three-layer ONIOM calculation symbolized as ONIOM(MP2/6−311+G(2df,2p):MP2/6−31G(d):B3LYP/3−21G), and $\Delta E_{MP2small}$ is the energy from a three-layer ONIOM calculation symbolized as ONIOM(MP2/6−31G(d):MP2/6−31G(d):B3LYP/3−21G). For each component energy determination, the high layer was restricted, as per Vreven and Morokuma,[12a] to the two core Group 13/15 atoms. Since our resources allowed it, we chose an OLS 3 medium layer rather than Vreven and Morokuma's suggestion of OLS 2; the low layer encompassed the entire complex.

To determine the utility of the OG2R3 approach, we needed standards to compare to. As a first step, we determined the OG2R3 energies of the complexes $H_xMe_{3-x}B$-$NH_xMe_{3-x}$ (first seven entries in Table 2), as experimental dissociation enthalpies are available for them.[10,18] It should be noted that optimizations were performed without resorting to ONIOM procedures; for the OG2R3 energies, the boron and nitrogen atoms were set in the high layer, carbon or hydrogen atoms bound to these were added to form the medium layer, and all atoms were placed in the low layer. While these simple amine-boranes do not fit our broad requirements of containing fluorinated substituents or aryl groups, they provide a sense of the agreement between experiment and theory for donor−acceptor complexes. The overall rms difference between experimental and computed energies is 2.8 kcal $mol^{-1}$. This value is skewed somewhat

***Table 2.*** M-X Bond Distances (Å) and $\Delta E_{DA}$ Values (kcal mol$^{-1}$) for $H_xMe_{3-x}B$-$NH_xMe_{3-x}$ and $(F_3C)_3M$-$XMe_3$ Complexes Using Various Approaches[a]

| | model | basis set | distance | $\Delta E_{DA}$ | $\Delta E_{DA}$(OG2R3) |
|---|---|---|---|---|---|
| $H_3BNH_2Me$ | MPW1K | 6−311++G(d,p) | | 33 | 31 (35.0) |
| $H_3BNHMe_2$ | MPW1K | 6−311++G(d,p) | | 34 | 33 (36.4) |
| $H_3BNMe_3$ | MPW1K | 6−311++G(d,p) | 1.630 (1.656) | 34 | 33 (34.8) |
| $Me_3BNH_3$ | MPW1K | 6−311++G(d,p) | | 12 | 16 (13.8) |
| $Me_3BNH_2Me$ | MPW1K | 6−311++G(d,p) | | 14 | 19 (17.6) |
| $Me_3BNHMe_2$ | MPW1K | 6−311++G(d,p) | 1.691 (1.656) | 13 | 19 (19.3) |
| $Me_3BNMe_3$ | MPW1K | 6−311++G(d,p) | 1.727(1.70) | 10 | 16 (17.6) |
| $(F_3C)_3BNMe_3$ | MP2 | 6−311++G(d,p) | 1.638 | 67 | 65 |
| | MPW1K | 6−311++G(d,p) | 1.634 | 53 | 65 |
| | ONIOM | 6−31+G(d) | 1.637 | 52 | 64 |
| $(F_3C)_3BPMe_3$ | MP2 | 6−311++G(d,p) | 1.967 | 72 | 75 |
| | MPW1K | 6−311++G(d,p) | 1.974 | 64 | 75 |
| | ONIOM | 6−31+G(d) | 1.976 | 62 | 74 |
| $(F_3C)_3AlNMe_3$ | MP2 | 6−311++G(d,p) | 2.011 | 57 | 55 |
| | MPW1K | 6−311++G(d,p) | 2.003 | 49 | 55 |
| | ONIOM | 6−31+G(d) | 2.007 | 48 | 55 |
| $(F_3C)_3AlPMe_3$ | MP2 | 6−311++G(d,p) | 2.423 | 51 | 54 |
| | MPW1K | 6−311++G(d,p) | 2.424 | 48 | 54 |
| | ONIOM | 6−31+G(d) | 2.423 | 48 | 54 |

[a] $\Delta E_{DA}$ values in the left most column are those from the model chemistry listed, and $\Delta E_{DA}$(OG2R3) entries are single point ONIOM G2R3 values using the optimized structures from the listed model chemistries. Experimental values are in parentheses.

by the large differences seen for the $H_3B$-$NR_3$ complexes; the rms error for the four $Me_3B$-$NR_3$ complexes is only 1.6 kcal mol$^{-1}$. The data set is too small to allow firm conclusions to be drawn, but it appears that the OG2R3 approach gives $\Delta E_{DA}$ values that should differ no more than 5 kcal mol$^{-1}$ from experiment on average.

While no experimental $\Delta E_{DA}$ values for $(R_F)_3M$-$XMe_3$ complexes exist, we showed previously that the MP2/6−311++G(d,p) method gives results most in agreement with experiment for $H_xMe_{3-x}B$-$NH_xMe_{3-x}$ complexes.[10] As a second step, we selected the four $(F_3C)_3M$-$XMe_3$ molecules as standards, since they are small enough to allow use of this model. As these do not contain phenyl rings, setting the ONIOM layers was straightforward. For the optimizations, the high layer contained the Group 13/15 core atoms and the carbons, while the low layer encompassed the entire complex. For the OG2R3 energy calculations, the high layer included only the core Group 13/15 atoms, the medium layer added the carbons to these, and the low layer held the entire complex. Data appear in Table 2. The OLS 2 ONIOM MPW1K distances for the complexes agree well with the standard MP2 and MPW1K distances, indicating that this model performs adequately for structure prediction. The ONIOM MPW1K approach, like the full MPW1K model, systematically underbinds the complexes, slightly for the Al systems and significantly for the B systems. However, the OG2R3 composite performs superbly, in that it predicts essentially identical energies regardless of the model employed and that it agrees very well with the MP2/6−311++G(d,p) standard values at a fraction of the resource cost.

Combining these results, we thus chose to perform all further production calculations using ONIOM MPW1K/6−31+G(d):MPW1K/3−21G optimizations with the high layer containing OLS 2 atoms for phenyl/pentafluorophenyl substituents, donor/acceptor-bound carbons for methyl/trifluoromethyl substituents, and trifluoromethyl group fluorine atoms. All other atoms were placed in the low layer. Optimized structures were used as bases for OG2R3 single point energy calculations with the two core Group 13/15 atoms in the high layer, a medium layer employing OLS 3 for aryl rings, carbon atoms for alkyl groups and fluorine atoms for $CF_3$ substituents, and a low layer covering the entire complex.

To obtain zero point energies (ZPEs), the structures found using ONIOM optimizations were used as starting points for optimizations using the Hartree−Fock/3−21G approach. Structures were proved to be minima by analytical frequency analysis at this level. ZPEs obtained were scaled by 0.9207 when used to correct the raw energy values,[19a] giving the final values listed in Table 4. In a few instances ($(F_5C_6)_3BP(i\text{-}Pr)_3$, for example), we found that the ONIOM-predicted structure differed significantly from that predicted by HF/3−21G, in that the M-X bond was significantly longer with the smaller basis set. This almost certainly reflects the absence of diffuse and polarization functions in the 3−21G basis set. We therefore redetermined the ZPEs for these and their components using frequency calculations employing the two-layer ONIOM approach. ZPEs obtained this way were not scaled when used to correct the raw energies. $\Delta E_{DA}$ values calculated this way differed insignificantly from those calculated using HF/3−21G ZPEs.

In the Discussion section below, we will refer to intrinsic bond dissociation energy, $\Delta E_{int}$. This term arises from broader methods of bond energy decomposition; it is sometimes called the snap bond energy or the instantaneous interaction energy.[20] The characterization and use of $\Delta E_{int}$ has been discussed in several places, so we describe it only briefly.

One way $\Delta E_{DA}$ may be decomposed is as follows

$$\Delta E_{DA} = \Delta E_{int} - \Delta E_{prep} = \Delta E_{int} - (\Delta E_{prep}(D) + \Delta E_{prep}(A))$$

where the reorganization energy $\Delta E_{prep}$ is the energy associated with deforming/relaxing the fragments of interest to their geometries in the molecule/ion (equally, the difference between the energy of the free moiety and that of the energy

***Table 3.*** Selected Structural Parameters for Donor−Acceptor Complexes (Distances in Å, Angles in deg) at the ONIOM MPW1K/6-31+G(d):MPW1K/3-21G Level

| | M-X | C-M-X-C | no. of F...H < 2.6 | no of F...H/max no. of F...H |
|---|---|---|---|---|
| $(F_3C)_3BNMe_3$ | 1.637 | 41.3 | 9 | |
| $(F_3C)_3BPMe_3$ | 1.976 | 40.2 | 0 | 0 |
| $(F_3C)_3BPEt_3$ | 1.987 | 51.4 | 6 | 0.5 |
| $(F_3C)_3BP(i\text{-}Pr)_3$ | 2.048 | 59.5 | 9 | 0.38 |
| $(F_3C)_3BP(t\text{-}Bu)_3$ | 2.163 | 46.2 | 9 | 0.38 |
| $(F_3C)_3BPPh_3$ | 2.016 | 29.6 | 3 | 1 |
| $(F_3C)_3AlNMe_3$ | 2.007 | 45.2 | 2 | |
| $(F_3C)_3AlPMe_3$ | 2.423 | 44.7 | 0 | 0 |
| $(F_3C)_3AlPEt_3$ | 2.421 | 51.8 | 0 | 0 |
| $(F_3C)_3AlP(i\text{-}Pr)_3$ | 2.437 | 50.9 | 6 | 0.25 |
| $(F_3C)_3AlP(t\text{-}Bu)_3$ | 2.442 | 38.9 | 8 | 0.33 |
| $(F_3C)_3AlPPh_3$ | 2.443 | 40.2 | 3 | 1 |
| $(F_5C_6)_3BNMe_3$ | 1.797 | 16.4 | | |
| $(F_5C_6)_3BPMe_3$ | 2.068 | 14.9 | | |
| $(F_5C_6)_3BPEt_3$ | 2.077 | 16.4 | | |
| $(F_5C_6)_3BP(i\text{-}Pr)_3$ | 2.194 | 0.4 | | |
| $(F_5C_6)_3BP(t\text{-}Bu)_3$ | 3.838 | 5.4 | | |
| $(F_5C_6)_3BPPh_3$ | 2.159 | 0.3 | | |
| $(F_5C_6)_3AlNMe_3$ | 2.060 | 22.0 | | |
| $(F_5C_6)_3AlPMe_3$ | 2.454 | 20.6 | | |
| $(F_5C_6)_3AlPEt_3$ | 2.438 | 21.4 | | |
| $(F_5C_6)_3AlP(i\text{-}Pr)_3$ | 2.519 | 26.4 | | |
| $(F_5C_6)_3AlP(t\text{-}Bu)_3$ | 2.621 | 15.7 | | |
| $(F_5C_6)_3AlPPh_3$ | 2.503 | 5.4 | | |

***Table 4.*** $\Delta E_{DA}$, $\Delta E_{DA}$(raw),[17] $\Delta E_{prep}$, and $\Delta E_{int}$ Energies (kcal mol$^{-1}$, OG2R3 Model) for Donor−Acceptor Complexes

| | $\Delta E_{DA}$ | $\Delta E_{DA}$(raw) | $\Delta E_{prep}$(A) | $\Delta E_{prep}$(D) | $\Delta E_{int}$[a] |
|---|---|---|---|---|---|
| $(F_3C)_3BNMe_3$ | 64 | 71 | 25 | 3 | 99 |
| $(F_3C)_3BPMe_3$ | 74 | 77 | 20 | 6 | 103 |
| $(F_3C)_3BPEt_3$ | 70 | 73 | 21 | 13 | 107 |
| $(F_3C)_3BP(i\text{-}Pr)_3$ | 77 | 81 | 29 | 8 | 118 |
| $(F_3C)_3BP(t\text{-}Bu)_3$ | 69 | 74 | 40 | 6 | 120 |
| $(F_3C)_3BPPh_3$ | 68 | 71 | 22 | 5 | 98 |
| $(F_5C_6)_3BNMe_3$ | 21 | 25 | 29 | 5 | 59 |
| $(F_5C_6)_3BPMe_3$ | 41 | 44 | 20 | 4 | 68 |
| $(F_5C_6)_3BPEt_3$ | 36 | 39 | 22 | 11 | 72 |
| $(F_5C_6)_3BP(i\text{-}Pr)_3$ | 32 | 33 | 29 | 6 | 68 |
| $(F_5C_6)_3BP(t\text{-}Bu)_3$ | 19 | 20 | 1 | 0 | 21 |
| $(F_5C_6)_3BPPh_3$ | 39 | 41 | 20 | 3 | 64 |
| $(F_3C)_3AlNMe_3$ | 55 | 59 | 7 | 1 | 67 |
| $(F_3C)_3AlPMe_3$ | 54 | 56 | 6 | 5 | 67 |
| $(F_3C)_3AlPEt_3$ | 51 | 53 | 7 | 12 | 72 |
| $(F_3C)_3AlP(i\text{-}Pr)_3$ | 68 | 71 | 10 | 6 | 87 |
| $(F_3C)_3AlP(t\text{-}Bu)_3$ | 79 | 82 | 15 | 5 | 102 |
| $(F_3C)_3AlPPh_3$ | 52 | 54 | 8 | 4 | 66 |
| $(F_5C_6)_3AlNMe_3$ | 39 | 42 | 10 | 2 | 55 |
| $(F_5C_6)_3AlPMe_3$ | 43 | 44 | 7 | 4 | 55 |
| $(F_5C_6)_3AlPEt_3$ | 39 | 41 | 8 | 11 | 60 |
| $(F_5C_6)_3AlP(i\text{-}Pr)_3$ | 42 | 44 | 15 | 9 | 68 |
| $(F_5C_6)_3AlP(t\text{-}Bu)_3$ | 42 | 46 | 20 | 3 | 69 |
| $(F_5C_6)_3AlPPh_3$ | 44 | 46 | 8 | 3 | 57 |
| $Ph_3BP(t\text{-}Bu)_3$ | 5 | 5 | 0 | 0 | 5 |
| $Ph_3BPPh_3$ | 22 | 22 | 20 | 2 | 44 |

[a] See the Computational Methods section.

of the moiety fixed in the orientation it has when bound to another moiety), and $\Delta E_{int}$ is therefore the intrinsic energy of the bond; that is, the energy required to dissociate the moieties from each other before they relax to their preferred separated structures. $\Delta E_{prep}$ here is treated as a positive value. It may be further subdivided into $\Delta E_{prep}$(D), the energy associated with deforming/relaxing the donor moiety, and $\Delta E_{prep}$(A), the analogous energy for the acceptor moiety.

This equation can be rewritten to define $\Delta E_{int}$:

$$\Delta E_{int} = \Delta E_{DA} + (\Delta E_{prep}(D) + \Delta E_{prep}(A))$$

This equation is that used to calculate $\Delta E_{int}$ values in Table 4. One should be aware that, as $\Delta E_{prep}$(D) and $\Delta E_{prep}$(A) correspond to energies of species not necessarily at their structural minima, they cannot be properly corrected for ZPE, as the mathematics of ZPE estimation require use of a minimum structure. Therefore, when determining $\Delta E_{int}$, one must use $\Delta E_{DA}$ uncorrected for ZPE. In Table 4 and the associated discussion, these uncorrected $\Delta E_{DA}$ values will be referred to as raw $\Delta E_{DA}$ values and symbolized as $\Delta E_{DA}$(raw).[17]

Though we will not examine this here, it may be of interest to know that $\Delta E_{int}$ may be considered composed of three terms: $\Delta E_{elstat}$, the electrostatic interaction energy between the fragments; $\Delta E_{orbital}$, the energy associated with relaxation of the orbitals as self-consistency is reached; and $\Delta E_{Pauli}$, the repulsive interaction energy between the fragments resulting from interactions between occupied orbitals. $\Delta E_{elstat}$ and $\Delta E_{orbital}$ broadly describe electrostatic and covalent attractive aspects of bonding, respectively, while $\Delta E_{Pauli}$ describes repulsive aspects.

As the OG2R3 composite employs relatively small basis sets for many of the energy determinations, it is useful to examine the basis set superposition energy (BSSE) corrections for systems like those here. However, few examples of BSSE-corrected ONIOM calculations have appeared.[21] This stems largely from two issues: one, the Gaussian program does not allow keyword-based calculation of BSSE for ONIOM runs (the Counterpoise[22] and ONIOM keywords cannot be used in the same run); and two, each BSSE run requires five separate energy determinations, thus being resource-intensive. For our OG2R3 energies, determining the BSSE for a particular molecule requires determining indi-

vidual BSSEs using six different model chemistry/layer combinations: CCSD(T)/6−31G(d) (the BSSE of which is hereafter denoted A), MP2/6−311+G(2df,2p) (B), and MP2/6−31G(d) (C) BSSEs for the small layer; MP2/6−31G(d) (D) and B3LYP/3−21G (E) BSSEs for the medium layer; and the B3LYP/3−21G (F) BSSE for the low layer. Thus a total of 30 individual energy calculations must be performed per molecule to estimate the BSSE by the counterpoise method. This is too demanding to be practical; thus we selected two molecules, $(F_3C)_3BPPh_3$ and $(F_5C_6)_3AlPMe_3$, and determined the BSSE of each on the assumption that the values would be representative. The relevant equation combining the individual BSSEs to give a total BSSE is as follows: BSSE (total) = (A + B − C) + (D − C) − (E − F).

The values are 18.2 and 10.6 kcal $mol^{-1}$, respectively, so on average one expects BSSEs for molecules in the set to be 14−15 kcal $mol^{-1}$.

These BSSEs appear large compared to those typical and call into question the utility of the OG2R3 approach. We examined the BSSEs of each model chemistry/layer combination, finding that those for the small layers (the A + B − C part) combined contributed only 2−3 kcal $mol^{-1}$ to the total, while the B3LYP model chemistry part (E−F) contributed less than 1 kcal $mol^{-1}$. The largest contribution to overall BSSE comes from the MP2-based (D−C) part, wherein the BSSE of the MP2/6−31G(d)/medium layer component is much larger than the BSSE of the MP2/6−31G(d)/small layer component. This suggests that using a smaller layer size for the former will lower the (D−C) correction and so the overall BSSE. We plan to explore this in future work. For now, we note that the BSSEs represent upper limits to the correction, and so the real correction for a particular molecule may be lower. In particular, the average BSSE probably cannot be applied to the weakly bound $(F_5C_6)_3BP(t\text{-}Bu)_3$ and $Ph_3BP(t\text{-}Bu)_3$ complexes, as these cannot readily share basis functions owing to the long distances between donor and acceptor atoms.

## Results and Discussion

**Structures.** The predicted structures (Table 3) of these donor−acceptor complexes display no remarkable features save those noted below. As mentioned, $(F_5C_6)_3M\text{-}XR_3$ complexes have previously been analyzed computationally and experimentally in terms of the impact of intramolecular forces on structure, particularly the observation of F...H interactions and eclipsed (or nearly so) conformations. We refer the reader to that work[3−6] and simply note here several observations. First, our optimization method appears to give excellent agreement with experiment. For example, the B−P distance in $(F_5C_6)_3B\text{-}PMe_3$ is 2.061(4) Å;[23] we predict 2.068 Å. The B−P distance in $(F_5C_6)_3B\text{-}PPh_3$ is 2.180(6) Å;[5] we predict 2.158 Å, an improvement over the DFT-D-PBE/TZVP level prediction of 2.22 Å.[6] With this in mind, we note the method predicts a B−P distance of 3.838 Å for $(F_5C_6)_3BP(t\text{-}Bu)_3$, some 0.4 Å shorter than that predicted by Pàpai et al.[3] Second, the M-P bond distances in the $(F_5C_6)_3B\text{-}PR_3$ series are nearly identical for R = Me and Et, increase about 0.1 Å for R = *i*-Pr and Ph, and then increase again for R = *t*-Bu (substantially for M = B). This suggests similar steric bulk for *i*-Pr and Ph substituents, somewhat at odds with their suggested cone angles.[24] However, the view is supported by the similar M-P distances for the $(F_5C_6)_3M\text{-}P(i\text{-}Pr)_3$ and $(F_5C_6)_3M\text{-}PPh_3$ complexes, which are quite distinct from, for example, the M-P distances in $(F_5C_6)_3M\text{-}PEt_3$ or $(F_5C_6)_3MP(t\text{-}Bu)_3$. Third, the method finds six intramolecular F...H interactions of ≤2.6 Å (the sum of the van der Waals radii) for $(F_5C_6)_3B\text{-}PR_3$ (R = Me, Et, *i*-Pr) but only three for R = *t*-Bu and Ph. In the case of R = *t*-Bu, this reflects the long B−P distance; in the case of R = Ph, it reflects the planarity of the phenyl substituents. Analogous results are seen for the aluminum complexes. This illustrates the point that the core bond distance, the substituent geometry, and the number of hydrogens available all determine the number of intramolecular contacts, and thus one must analyze the data carefully to ascertain the presence and importance of such contacts. Fourth, in support of this view, the torsional angles rapidly *decrease* with the steric bulk of the phosphorus-bound substituents; i.e. as the substituents get larger, the complexes adopt more eclipsed geometries, in stark contrast to expectation. This effect is most pronounced for the borane-phosphines (some ambiguity exists on this point for the borane systems owing to the lack of a B−P bond in $(F_5C_6)_3BP(t\text{-}Bu)_3$) but appears in the alane-phosphines as well. It is notable that $PPh_3$ complexes show the most eclipsed conformations; this may reflect the presence of $\pi$-$\pi$ interactions arising from intramolecular ring stacking in addition to the F...H interactions.[6]

No detailed analyses of intramolecular interactions in $(F_3C)_3M\text{-}XR_3$ complexes have appeared. We commented on the structures of $(F_3C)_3B\text{-}NR_3$ (R = H, Me) complexes previously, comparing computational to experimental results.[10] At that time, the observation of conformational deviations from staggered structures was of minor concern, being reflected mostly in the energies required to rotate groups around the B−N bonds. It is appropriate here to expand on this, including different core atoms and larger substituents. Before doing so, we note that the $(F_3C)_3M$ complexes display M-X bond distances markedly shorter than those for $(F_5C_6)_3M$ complexes (Table 2). This is particularly true for the $(R_F)_3M\text{-}NMe_3$ systems, where the differences are 0.16 Å for B and 0.053 Å for Al. The extent to which the distances increase in $(F_3C)_3M\text{-}PR_3$ complexes as the phosphine bulk increases from R = Me to R = *i*-Pr is smaller as well. Both observations suggest that the $CF_3$ substituent is more electron-withdrawing and sterically smaller than the $C_6F_5$ group.[25]

The data in Table 3 point out the difficulties in assessing the importance of intramolecular F...H interactions for $(F_3C)_3M\text{-}XR_3$ complexes but in general suggest that they are of little importance. First, the $F_3C\text{-}B\text{-}P\text{-}C$ torsion angles are much closer to the ideal staggered value of 60° than are their $F_5C_6\text{-}B\text{-}P\text{-}C$ counterparts. Also, the torsion angles do not vary regularly with the size or number of H atoms available for intermolecular interactions. One sees that for M = B and Al, the angle increases with the size of the R substituent to a maximum for R = *i*-Pr (i.e., the conformation becomes more staggered as size increases) and then decreases for R

= *t*-Bu. This holds for M = Al despite the fact that the Al−P bond distance increases relatively little across the series. Moreover, while the number of F...H distances ≤2.6 Å increases with the size and number of hydrogen atoms, the ratio of these to the maximum number of possible such interactions remains fairly constant.[26]

Considering all these observations, it is clear that the structural data do not allow an unambiguous evaluation of the presence or importance of intramolecular interactions here, but they are clearly less important than in the pentafluorophenyl systems. The close contacts observed may simply reflect the steric needs of the substituents compared to the sizes of the core atoms. One notes in this regard that $(F_3C)_3B$-$NMe_3$ and $(F_3C)_3B$-$PEt_3$ exhibit respectively nine and six F...H distances ≤2.6 Å, while the Al analogues exhibit none. The torsional angles clearly express competition between the steric needs of the substituents vs attempts to create close F...H contacts. As will be seen below, the energetic data provide a clearer means by which to assess the importance of intramolecular interactions and support their minimal contribution to the $\Delta E_{DA}$ values.

**Donor−Acceptor Dissociation Energies.** The $\Delta E_{DA}$ (OG2R3) values for the $(R_F)_3M$-$XR_3$ complexes studied appear in the first column of Table 4. As $(F_5C_6)_3BP(t\text{-Bu})_3$ uniquely does not contain a B−P bond, we except it from the discussion save when we include it explicitly. That said, several observations can be made, and several trends discerned.

The $(F_3C)_3BXR_3$ complexes display slightly larger $\Delta E_{DA}$ values than do their Al analogues, but the $(F_5C_6)_3BXR_3$ complexes display slightly smaller $\Delta E_{DA}$ values than their Al analogues. This presumably reflects crowding of the more compact boron systems with the larger fluoroaryl substituents. This finds support in the observation that $\Delta E_{DA}$ values for the $(F_5C_6)_3AlPR_3$ systems scarcely change as the phosphine increases in size, while those for the boron analogues decrease regularly, culminating in the lack of a B−P interaction in $(F_5C_6)_3BP(t\text{-Bu})_3$ (although this still shows a sizable dissociation energy of 19 kcal $mol^{-1}$).

As defined by relative $\Delta E_{DA}$ values, the $(F_3C)_3B$ moiety is a much stronger Lewis acid than is the $(F_5C_6)_3B$ moiety. Timoshkin and Frenking[27] have recently calculated $\Delta E_{DA}$ values for (aryl)- and (fluoroaryl)boranes, alanes, and gallanes, finding that all three Group 13 $M(C_6F_5)_3$ species are essentially as Lewis acidic as the corresponding $MCl_3$ species. They describe $Al(C_6F_5)_3$ as "one of the strongest Lewis acids". The data in Table 4 clearly show that $B(CF_3)_3$ is approximately twice as strong as this as assessed by $\Delta E_{DA}$, while $Al(CF_3)_3$, if prepared, would be about 1.3 times as strong. These views are supported by the unusual stability of $(F_3C)_3B$ complexes noted above (and by its nonexistence as a free borane). The synthesis of $Al(CF_3)_3$ and further studies of the reactivities of both $M(CF_3)_3$ would be of considerable interest; that said, the bonds between these and donors may prove too strong to break easily, limiting their chemistry.

As mentioned above, the structural data provide little support for the presence of intramolecular F...H interactions in $(F_3C)_3MXR_3$ complexes. The energetic data show similar lacks. The $\Delta E_{DA}$ values for the systems showing the most extensive close F...H contacts, the $(F_3C)_3BPR_3$ series, change little as the number of phosphine H atoms increases. One would anticipate that, if such interactions were important, $\Delta E_{DA}$ for $(F_3C)_3BP(t\text{-Bu})_3$, with 9 F...H contacts ≤2.6 Å, would be sizably larger than that for $(F_3C)_3BPMe_3$, with no such short contacts. This does not hold. It is true that $\Delta E_{DA}$ values for the $(F_3C)_3AlPR_3$ series increase with size, so possibly intramolecular contacts play a larger role for these. We cannot assess this quantitatively given the data in hand.

In contrast to the general view that bonds between elements higher in the Periodic Table exhibit larger dissociation energies, the M-P bonds are generally stronger than the M-N bonds. Surprisingly, this holds particularly for B-N/B-P systems, where 2p-2p orbital overlap in the former is usually thought to provide stronger bonding than 2p-3p overlap in the latter. Jacobsen et al. analyzed[4] the related $(F_5C_6)_3B$-NCMe and $(F_5C_6)_3B$-$PH_3$ complexes, finding that the smaller covalent and electrostatic bonding interactions in the borane-phosphine were countered significantly by decreased electronic repulsions resulting from the longer B-P bond length. It appears similar effects occur here.

We noted above that, in terms of the M-P bond length, $PPh_3$ appeared similar to $P(i\text{-Pr})_3$. In terms of $\Delta E_{DA}$ values, $PPh_3$ is clearly most similar to $PMe_3$. This holds particularly when considering intrinsic $\Delta E_{int}$ energies, indicating that the electronic properties of the two are similar. Since $PPh_3$ is generally observed to be a poorer donor than $PMe_3$, this suggests that the interplay between bonding attractions and electronic repulsions mentioned above applies here as well. In every case, the $(R_F)_3M$-$PPh_3$ bond is longer than the $(R_F)_3M$-$PMe_3$ bond, so the repulsions decrease in the former apparently to the same extent that the attractions increase in the latter.

Curiously, the dissociation energies for $(R_F)_3M$-$PR_3$ complexes change erratically with the size of the phosphine. This holds despite the fact that the M-P bonds increase in length fairly regularly as the phosphine bulk increases along the series. This suggests a fairly flat potential energy surface for bond breaking in these systems. It also suggests that the reorganization energy $\Delta E_{prep}$ of the fragments contributes detectably to the overall $\Delta E_{DA}$.

The data in Table 4 bear this out. One sees that, as expected, generally $\Delta E_{prep}$ is substantially larger for the acceptors than for the donors, as the acceptors change geometry from pseudotetrahedral to pseudotrigonal planar, while the donors change only from pseudotetrahedral to pseudopyramidal. $\Delta E_{prep}$ is larger for boranes than for alanes, because the more compact boranes encounter more repulsions as the substituents adjust from being 120° apart to being at more acute angles. Interestingly, the $\Delta E_{prep}$ values for the acceptors are approximately the same for $(F_3C)_3M$ and $(F_5C_6)_3M$ complexes for a particular M, indicating that the parameter is fairly insensitive to the size or electron-withdrawing properties of the fluorinated substituents. The total $\Delta E_{prep} = (\Delta E_{prep}(A) + \Delta E_{prep}(D))$ tends to increase from $(R_F)_3MPMe_3$ to $(R_F)_3MP(t\text{-Bu})_3$; this value for $PPh_3$ complexes tends to mimic that for $PMe_3$ complexes (see above). Of course, $(F_5C_6)_3BP(t\text{-Bu})_3$, which represents a "bound"

system where the fragments are already in their relaxed forms, shows $\Delta E_{prep} \approx 0$.

The intrinsic dissociation energies $\Delta E_{int}$ increase regularly with the size of the phosphine, indicating that they represent only the effect of increasing phosphine basicity. Remarkably, the model predicts $(F_3C)_3BP(t\text{-}Bu)_3$ to display $\Delta E_{int} = 120$ kcal mol$^{-1}$, a value in excess of most covalent bonds spilt homolytically. $\Delta E_{int}$ values for $PPh_3$ complexes are closest to those for $PMe_3$ complexes, reiterating the view that these two donors behave similarly. It is evident that all these $(R_F)_3MXR_3$ complexes are strongly bound, making Stephan's discovery of the distinctiveness of weakly bound $(F_5C_6)_3BP(t\text{-}Bu)_3$ all the more impressive. That this complex is so reactive despite its being bound by 20 kcal mol$^{-1}$ indicates that this value represents an approximate "reactivity limit"; complexes with dissociation energies much larger than this are likely to be unreactive.

We have included in Table 4 our calculated dissociation energies for $Ph_3BP(t\text{-}Bu)_3$ and $Ph_3BPPh_3$. The former, as noted in the Introduction, also heterolytically breaks the $H_2$ bond; we included the latter for comparison as a phosphine-borane with large substituents. The complexes differ in that we find the *t*-Bu complex to be weakly bound by dispersion effects and not to contain a B−P bond, while the latter contains a B−P bond but displays a dissociation energy similar to that of $(F_5C_6)_3BP(t\text{-}Bu)_3$. The data show that $Ph_3BP(t\text{-}Bu)_3$ is bound by only 5 kcal mol$^{-1}$. This implies that F...H interactions in the fluorinated homologue $(F_5C_6)_3BP(t\text{-}Bu)_3$ contribute ca. 15 kcal mol$^{-1}$ to the binding energy. The weak dissociation energy implies that if separation of the fragments dictates the $H_2$ splitting rate, then the parent system should react faster. In fact, it reacts more slowly,[1] implying that another factor, such as the Lewis acidity of the acceptor, is critically important in determining the rate. This being so, it seems plausible that the $H_2$ splitting involves the "reaction cell" mechanism suggested by Papai et al.[3a] See the Computational Methods section.

## Conclusions

The OG2R3 model predicts dissociation energies of large donor−acceptor complexes with acceptable resource usage. As a reviewer noted, it is impossible to tell whether the energies are trustworthy, as no comparable experimental data exist. That said, they agree reasonably with known experimental energies and are consistent with the reactivity patterns exhibited by systems for which dissociation energies are unknown. We hope that such large complexes can be examined this way will spur effort to examine systems that previously were simplified.

From an experimental standpoint, the data suggest that one should be able to isolate all the $(R_F)_3MXR_3$ complexes examined. Difficulties in this regard lie in the systems finding other reaction pathways,[28] but assiduous effort should overcome this. While $(F_5C_6)_3BP(t\text{-}Bu)_3$ has so far defied isolation, the interaction energy is large enough that one could conceivably crystallize it, but the fact that it is bound through dispersive effects means that no spectroscopic method is likely to detect it. Only $Ph_3BP(t\text{-}Bu)_3$ appears to have too little binding energy to allow its isolation, and so far the reactivity of this complex does not merit extensive study.

Our finding that $PPh_3$ mimics $PMe_3$ in terms of the M-P bond length and dissociation energy is intriguing, given that the proton affinities of the two are so different. We plan to probe the reasons for this observation, beginning with the hypothesis that, for a sufficiently sized acceptor, the steric and electronic effects associated with binding cancel.

**Acknowledgment.** The NIU Computational Chemistry Laboratory was created using funds from the U.S. Department of Education Grant P116Z020095 and is supported in part by the taxpayers of the State of Illinois.

**Supporting Information Available:** Optimized Cartesian coordinates of all molecules examined, with absolute energies. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Welch, G. C.; Stephan, D. W. *J. Am. Chem. Soc.* **2007**, *129*, 1880–1881.

(2) Stephan, D. W. *Org. Biomol. Chem.* **2008**, *6*, 1535–1539.

(3) (a) Rokob, T. A.; Hamza, A.; Stirling, A.; Soós, T.; Pápai, I. *Angew Chem. Int. Ed.* **2008**, *47*, 2435–2438.

(4) Jacobsen, H.; Berke, H.; Dšring, S.; Kehr, G.; Erker, G.; Fröhlich, R.; Meyer, O. *Organometallics* **1999**, *18*, 1724–1735.

(5) Mountford, A. J.; Lancaster, S. J.; Coles, S. J.; Horton, P. N.; Hughes, D. L.; Hursthouse, M. B.; Light, M. E. *Inorg. Chem.* **2005**, *44*, 5921–5933.

(6) Spies, P.; Fröhlich, R.; Kehr, G.; Erker, G.; Grimme, S. *Chem. Eur. J.* **2008**, *14*, 333–343.

(7) Pawelke, G.; Bürger, H. *Appl. Organomet. Chem.* **1996**, *10*, 47–174.

(8) Brauer, D. J.; Bürger, H.; Dörrenbach, F.; Krumm, B.; Pawelke, G.; Weuter, W. *J. Organomet. Chem.* **1990**, *385*, 161–172.

(9) Pawelke, G. *J. Fluorine Chem.* **1995**, *73*, 51–55.

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

(11) 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.; Rega, N.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98, Revision A.11.4*; Gaussian, Inc.: Pittsburgh, PA, 2002.

(12) (a) Vreven, T.; Morokuma, K. *J. Phys. Chem. A* **2002**, *106*, 6167–6170. (b) Vreven, T.; Morokuma, K. *J. Chem. Phys.* **1999**, *111*, 8799–8803.

(13) Curtiss, L. A.; Raghavachari, K. *Theor. Chem. Acc.* **2002**, *108*, 61–70.

(14) Möller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.

(15) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2001**, *105*, 2936–2941.

(16) Vreven, T.; Morokuma, K.; Farkas, O.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, *24*, 760–769.

(17) Energies in Table 1 are not corrected for ZPE, as they were used solely for test comparison and not as predicted values. Throughout this work, energies not corrected for ZPE are designated as raw energies and symbolized as $\Delta E_{DA}$(raw).

(18) These enthalpies, while technically being different than $\Delta E_{DA}$ values determined computationally, have nonetheless been modeled well by various computational methods. See: Haaland, A. *Angew. Chem., Int. Ed. Engl.* **1989**, *28*, 992–1007.

(19) (a) Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502–16513. (b) Truhlar, D. G. Database of Frequency Scaling Factors for Electronic Structure Methods. http://comp.chem.umn.edu/database/freq_scale.htm (accessed July 7, 2008).

(20) (a) Diefenbach, A.; Bickelhaupt, F. M.; Frenking, G. *J. Am. Chem. Soc.* **2000**, *122*, 6449–6458. (b) Szilagyi, R.; Frenking, G. *Organometallics* **1997**, *16*, 4807–4815. (c) Ziegler, T. *Can. J. Chem.* **1995**, *73*, 743–761. (d) Ziegler, T. *Chem. Rev* **1991**, *91*, 651–667. (e) Ziegler, T. In *Metal-Ligand Interactions: from Atoms to Clusters to Surfaces*; Salahub, D. R.,Russo, N., Eds.; Kluwer: The Netherlands, 1992: pp 367−396.

(21) For examples, see: (a) Kuno, M.; Hongkrengkai, R.; Hannongbua, S. *Chem. Phys. Lett.* **2006**, *424*, 172–177. (b) Tschumper, G. S.; Morokuma, K. *J. Mol. Struct. (THEOCHEM)* **2002**, *592*, 137–147.

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

(23) Chase, P. A.; Parvez, M.; Piers, W. E. *Acta Crystallogr.* **2006**, *E62*, o5181−o5183.

(24) Huheey, J. E.; Keiter, E. A.; Keiter, R. L. *Inorganic Chemistry*, 4th ed.;Harper Collins: New York, 1993; p 690.

(25) (a) Suresh, C. H. *Inorg. Chem.* **2006**, *45*, 4982–4986. (b) Boere, R. T.; Zhang, Y. *J. Organomet. Chem.* **2005**, *690*, 2651–2657.

(26) The maximum number of F. .H interactions was determined by using models to examine conformations. The values are 12 for R = Me, 18 for R = Et, and 24 for R = *i*-Pr and *t*-Bu.

(27) Timoshkin, A. Y.; Frenking, G. *Organometallics* **2008**, *27*, 371–380.

(28) (a) Welch, G. C.; Holtrichter-Roessmann, T.; Stephan, D. W. *Inorg. Chem.* **2008**, *47*, 1904–1906. (b) Finze, M.; Bernhardt, E.; Willner, H.; Lehmann, C. W. *Inorg. Chem.* **2006**, *45*, 669–678.

CT8001859

JCTC Journal of Chemical Theory and Computation

# First Principles Study of NO and NNO Chemisorption on Silicon Carbide Nanotubes and Other Nanotubes

Guohua Gao† and Hong Seok Kang*

*Department of Nano and Advanced Materials, College of Engineering, Jeonju University, Hyoja-dong, Wansan-ku, Chonju, Chonbuk 560-759, Republic of Korea*

Received June 8, 2008

**Abstract:** Using methods based on first principles, we find that NO and NNO molecules can be chemisorbed on silicon carbide nanotubes (SiCNTs) with an appreciable binding energy and that this is not the case for either carbon nanotubes (CNTs) or boron nitride nanotubes (BNNTs). A detailed analysis of the energetics, geometry, and electronic structure of various isomers of the complexes was performed. The adsorption energy (∼−0.7 eV) is larger for the SiCNT-NO complex. The complex exhibits magnetism, and a ferromagnetic coupling of spins is observed when more than one NO molecule is adsorbed. This observation suggests that magnetic properties can be used to sense the amount of NO molecules adsorbed. The SiCNT-NNO complex is a nonmagnetic system in which five-membered rings form at the binding site.

## 1. Introduction

Among the many materials for semiconductors, bulk SiC has been found to be potentially useful for high power, high frequency, and high temperature electronic devices.[1] Therefore, motivated by the recent discovery of carbon nanotubes (CNTs), studies have sought to synthesize tubular forms of SiC. SiC nanotubes (SiCNTs) have been successfully synthesized from the reaction of SiO and multiwalled CNTs.[2] In SiCNTs, carbon and silicon atoms exist in a 1:1 ratio, and a theoretical calculation shows that the tubes consist of alternating C and Si atoms forming $sp^2$ Si−C bonds.[3] Single-walled SiCNTs are known to be semiconductors irrespective of their chiral indices, because of the ionicity of Si−C bonds. The tubes are expected to be amenable to a variety of external functionalizations due to the presence of the Si atoms, which prefer $sp^3$ hybridization. In fact, the exterior surface of SiCNTs is much more reactive than the exterior of either CNTs or boron nitride nanotubes (BNNTs), making SiCNTs much more interesting for chemical functionalization than the other type of nanotubes. For example, transition metal atoms can be chemically adsorbed on SiCNTs with binding energies greater than 1.17 eV.[4] SiCNTs have also proven useful for hydrogen storage, since the binding energy of the tubes with hydrogen molecules is 20% larger than that of CNTs.[5]

Nitrogen oxides (NOx) are compounds notorious for their harmful impact on the environment. They are mostly generated as byproducts of high-temperature combustion of fossil fuels.[6] Although there have been many efforts to use catalysts to reduce the amount of nitrogen oxides in the air,[7] an efficient method of sensing and removing the pollutants is still needed. Very recently, $SnO_2-In_2O_3$ nanocomposites have been shown to be useful as semiconductor NOx sensors, since they can selectively and reproducibly detect the gases at the level of parts-per-million.[8] Heme-nitric oxides, which constitute a newly discovered family of heme proteins, have also been found to detect the molecules selectively and sensitively.[9] Nanocrystalline titanium dioxide is known to be useful for photocatalytic degradation of the molecules.[10]

These results prompted us to take a theoretical approach to ask whether nanotubes could be used for this purpose, particularly in light of the high chemical reactivity of SiCNTs. To our knowledge, there has been only one theoretical work related to that problem: Rafati et al. found that an NO molecule can be physisorbed on the surface of CNTs endothermically.[11] Nevertheless, recent quantum chemical calculations have elucidated the mechanism of reaction of tungsten with NOx.[12] There has been an

* Corresponding author e-mail: hsk@jj.ac.kr.

† Alternate address: Pohl Institute of Solid State Physics, Tongji University, Shanghai 200092, P. R. China.

10.1021/ct800273c CCC: $40.75 
Published on Web 09/12/2008

investigation of the structure and energetics of NOx adsorption on clusters[13] and studies of NOx binding on various surfaces.[14]

The present work is devoted to a first-principles investigation of using SiCNTs as a sensor and adsorbent of NO and NNO molecules. The study also compares the adsorption performance of SiCNTs, CNTs, and BNNTs.

## 2. Theoretical Methods

Total energy calculations were performed using the Vienna *ab initio* simulation package (VASP).[15,16] Electron-ion interactions were described by the projected augmented wave (PAW) method.[17] Exchange-correlation effects were treated within the generalized gradient approximation (GGA) of Perdew, Burke, and Ernzerhof.[18] The cutoff energy was set high enough (400 eV) to ensure accurate results, and the conjugate gradient method was employed to optimize the geometry until the Hellmann−Feynman force exerted on an atom was less than 0.03 eV/Å.

In order to investigate differences in adsorption properties of armchair and zigzag tubes, we examined supercells, which consisted of three and four primitive cells of (8,0) and (5,5) nanotubes, respectively. Although we focused on SiCNTs, we also analyzed CNTs and BNNTs for the purposes of comparison. The total number of atoms, the diameter, and the lattice parameter were (80, 8.62 Å, 12.39 Å) and (96, 8.06 Å, 16.02 Å) for (5,5) and (8,0) SiCNTs, respectively. The corresponding parameters for (5,5) CNT, (8,0) CNT, (5,5) BNNT, and (8,0) BNNT were (80, 6.87 Å, 9.90 Å), (96, 6.46 Å, 12.78 Å), (80, 7.00 Å, 9.90 Å), and (96, 6.58 Å, 12.90 Å), respectively. Here, lattice parameters are optimized values in such a way that the axial stress is zero. Two *k*-points were used for *k*-point sampling in the irreducible region of the first Brillouin zone along the tube axis (X), which ensures the accuracy of the calculation within 1 meV even for metallic systems. In these analyses, we used large supercells to guarantee that interatomic distances between neighboring cells along the Y and Z directions were greater than 10.3 Å.

## 3. Results

First, we investigated NO adsorption on the surface of nanotubes. We describe our calculations on three different configurations of the NO molecule on a SiCNT of a specific chiral index (*n, m*). In Table 1, these configurations are denoted (E, Z, $Z^R$) and (Z, A, $Z^R$) for (5,5) and (8,0) SiCNTs, respectively. As shown in Figure 1 the letters Z, E, and A indicate whether the NO molecule is located directly above the zigzag, equatorial, or axial Si−C bond, respectively. A zigzag bond is defined as a bond that is neither parallel nor perpendicular to the tube axis. An equatorial bond is perpendicular to the tube axis, while an axial bond is parallel to the axis. We note that there are no axial bonds in armchair tubes and no equatorial bonds in zigzag tubes. In the E, Z, and A configurations, the oxygen atom, which is more electronegative than the nitrogen atom in the NO molecule, is located directly above a silicon atom ($Si_1$), which is more electropositive than carbon atoms. Likewise, the nitrogen atom is located directly above a carbon atom ($C_1$) that is bonded to $Si_1$. Therefore, there can be N−$C_1$ and O−$Si_1$ bonds. On the other hand, in configuration $Z^R$, the nitrogen atom is located above the silicon atom of a zigzag bond, so there can be O−$C_1$ and N−$Si_1$ bonds instead.

***Table 1.*** Energetic, Magnetic, and Geometric Parameters for Various Configurations of SiCNT-NO Complexes in Which the SiCNTs Have (5,5) and (8,0) Chiral Indices

| | chiral index | | | | | |
|---|---|---|---|---|---|---|
| | (5,5) | | | (8,0) | | |
| configuration | E | Z | $Z^R$ | Z | A | $Z^R$ |
| $E_b(1)$[a] (eV) | −0.67 | −0.63 | 0.82 | −0.62 | −0.61 | 0.69 |
| $E_b(2)$[b] (eV) | | −1.54 | | −1.41 | | |
| $\mu(1)$[c] ($\mu_B$) | 1.00 | 0.95 | 1.00 | 0.92 | 1.00 | 1.00 |
| $\mu(2)$[d] ($\mu_B$) | | 1.98 | | 2.00 | | |
| $l$ (N−$C_1$)[e] (Å) | 1.50 | 1.49 | 1.54 | 1.48 | 1.52 | 1.51 |
| $l$ (O−$Si_1$)[f] (Å) | 1.75 | 1.80 | 1.83 | 1.79 | 1.76 | 1.81 |
| $l$ (N−$Si_2$)[g] (Å) | 2.78 | 1.93 | 3.00 | 1.98 | 2.73 | 2.69 |
| $l$ ($C_1$−$Si_1$)[h] (Å) | 1.98 (1.79) | 1.90 (1.79) | | 1.92 (1.79) | 1.91 (1.79) | |
| $l$ (O−N)[i] (Å) | 1.40 (1.17) | 1.40 (1.17) | 1.39 (1.17) | 1.40 (1.17) | 1.40 (1.17) | 1.39 (1.17) |
| $q$(NO)[j] | −0.48 (0) | −0.51 (0) | | −0.52 (0) | −0.46 (0) | |

[a] Binding energy of one NO molecule on the surface of the SiCNT. [b] Binding energy of two NO molecules on the surface of the SiCNT. [c] Magnetic moment of the SiCNT-NO complex with one NO molecule adsorbed. [d] Magnetic moment of the SiCNT-NO complex with two NO molecules adsorbed. [e] N−$C_1$ distance, where $C_1$ is the carbon atom of the SiCNT above which the nitrogen atom of the NO molecule is located. For configuration $Z^R$, the number corresponds to the O−$C_1$ distance instead. [f] O−$Si_1$ distance, where $Si_1$ is the silicon atom of the SiCNT above which the oxygen atom of the NO molecule is located. For configuration $Z^R$, the number corresponds to the N−$Si_1$ distance instead. [g] N−$Si_2$ distance, where $Si_2$ is a silicon atom of the SiCNT in the bonds $Si_1$−$C_1$−$Si_2$ of the SiCNT. For configuration $Z^R$, the number corresponds to the O−$Si_2$ distance. [h] $C_1$−$Si_1$ bond length. The numbers in parentheses denote the corresponding data in pristine tubes. [i] N−O bond length in the NO molecule. [j] Total Mulliken charge of the NO molecule. The numbers in parentheses denote the corresponding data for an isolated NO molecule.

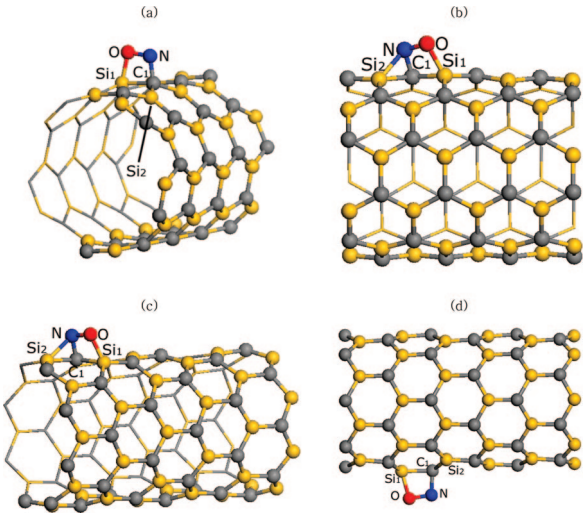



***Figure 1.*** Optimized structures for stable configurations of SiCNT-NO complexes: configurations E (a) and Z (b) of the (5,5) complex and configurations Z (c) and A (d) of the (8,0) complex. Atomic labels are also defined.

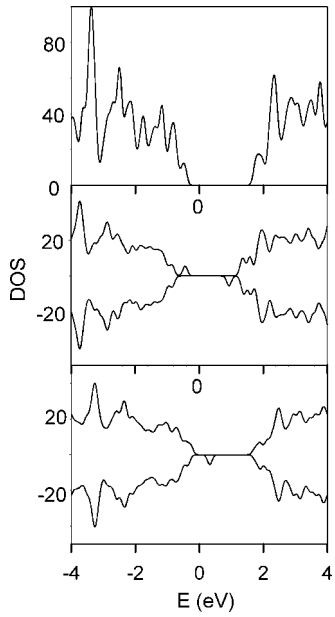



***Figure 2.*** Comparison of the electronic density of states (DOS) for (5,5) SiCNT-NO complexes: pristine tube (top), configuration E (middle), and configuration Z (bottom). The Fermi level is set to zero. For the complex, DOSs for spin-up and spin-down states are drawn separately.

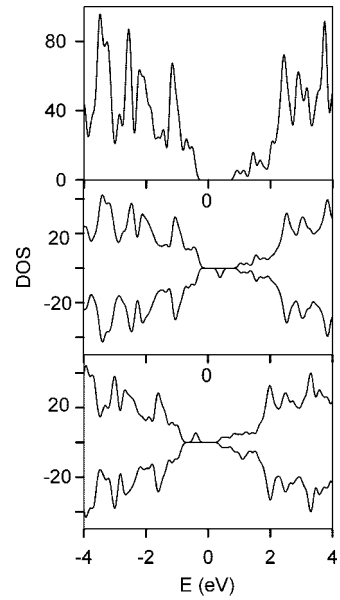



***Figure 3.*** Comparison of the electronic density of states (DOS) for (8,0) SiCNT-NO complexes: pristine tube (top), configuration Z (middle), and configuration A (bottom). The Fermi level is set to zero. For the complex, DOSs for spin-up and spin-down states are drawn separately.

Table 1 shows that there is appreciable binding between an NO molecule and SiCNTs in configurations E, Z, and A. In fact, the magnitude of the binding energy or adsorption energy ($E_b$) of an NO molecule, which is defined as the energy change associated with the process SiCNT + NO(doublet) → NO-SiCNT, is between −0.61 and −0.67 eV for (5,5) and (8,0) SiCNTs. We do not observe any significant difference in binding energy between the tubes of two different chiral indices. Table 1 and Figures 2 and 3 show that the NO-SiCNTs are magnetic semiconductors with magnetic moments close to 1.0 $\mu_B$.[19] A separate analysis shows that the spin polarization is largely concentrated on the NO molecule. As two examples, Figure 4 shows the spin density distributions of configuration E of (5,5) and configuration Z of (8,0) SiCNT-NO complexes depicted in Figure 1(a),(c).

When two NO atoms are adsorbed, the binding of the second NO molecule is stronger than the first one, indicating that the NO binding is cooperative. For example, Table 1 shows that the adsorption energy of the first and the second NO molecules for isomer Z of (5,5) SiCNTs is −0.63 eV and −0.91 eV, respectively. Spins of two NO molecules couple ferromagnetically, resulting in magnetic moments of 2.0 $\mu_B$ for both of (5,5) and (8,0) tubes. In order to estimate the coupling strength between local magnetic moments of two NO molecules, we have calculated the energy difference ($\Delta E_{F\text{-}AF}$) between ferromagnetic and antiferromagnetic states from the relation: $\Delta E_{F\text{-}AF} = E_F - E_{AF}$. Its values (= −0.16 eV and −0.35 eV, respectively) show that there are strong couplings between local magnetic moments for both of (5,5) and (8,0) complexes, suggesting the possibility of long-range magnetic ordering at room temperature.

As Figure 5 shows, the second NO molecule is adsorbed on the nearest carbon atom of the tubes in the most stable

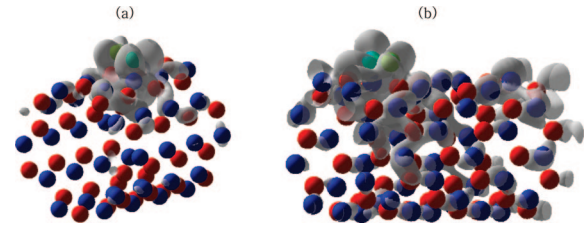



***Figure 4.*** Spin density plot of configuration E (a) of (5,5) and configuration Z (b) of (8,0) SiCNT-NO complexes, in each of which the tube has the same orientation as in Figure 1(a) and (c).

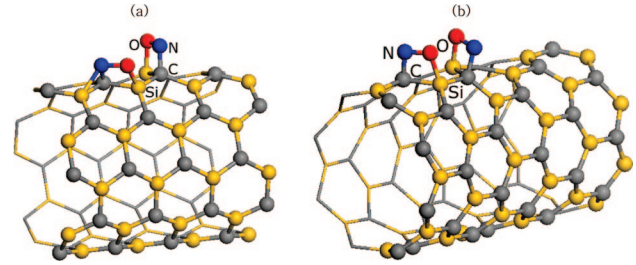
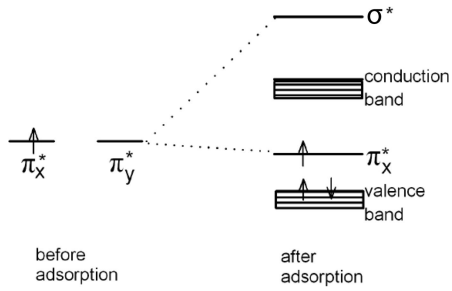



***Figure 5.*** Optimized structures of (5,5) (a) and (8,0) (b) SiCNT-2NO complexes.

configuration. Namely, two NO molecules are found to be adsorbed on $C_1$ and $C_2$ in the bonds $C_1-Si-C_2$ of the tube. In the figure, two NO molecules adopt E and Z configurations, respectively, for the (5,5) complex. For the (8,0) complex, both of them adopt Z configurations. The configurations are more stable than another configuration by 0.26 and 0.14 eV for (5,5) and (8,0) tubes, respectively, in which two NO molecules are located far away on the opposite sides of the circumference of the tube. Therefore, the change in the magnetic moment of SiCNTs can be used to sense the amount of NO gas adsorbed. As shown in Table 1, $Z^R$

***Scheme 1.*** Schematic Energy Level Diagram of Two $\pi^*$ Molecular Orbitals of the NO Molecule before and after Adsorption on Isomer E of (5,5) SiCNT




configurations are not expected to be observable in either (5,5) or (8,0) tube, since the binding energy is largely positive. Therefore, we concentrate our work on configurations E, Z, and A.

As was mentioned in the previous paragraph, Figure 1 shows the molecular structure of SiCNTs with one NO molecule adsorbed on the surface for configurations E and Z of the (5,5) complex and Z and A of the (8,0) complex. The N−$C_1$ bond length is 1.49−1.52 Å, almost achieving the length of a single covalent bond.[20] There is an O−$Si_1$ bond that is somewhat weaker than a single bond, as indicated by its bond length (1.75−1.80 Å), which is longer than that (1.66 Å) of an Si−O single bond.[19] As a result, the $C_1$−$Si_1$ and N−O bonds are appreciably weakened, which is seen in their bond lengths (1.90−1.98 Å and 1.39−1.40 Å, respectively), which are longer than those in pristine tubes (1.79 Å) and an isolated NO molecule (1.17 Å). In fact, comparison of the length of the N−O bond in the adsorbed molecule with the lengths of single and double N−O bonds indicates that the bond order changes from 2.5 to 1. There are also weak N−$Si_2$ bonds in the Z configuration of both (5,5) and (8,0) tubes, which is not the case in other configurations.

In order to better understand the electronic properties of the complex, we used the E configuration of the (5,5) complex as an example. (We recall that the X-axis is parallel to the tube axis. For armchair tubes, we also assume that the equatorial direction of the NO molecule is parallel to the Z-axis. Therefore, the $\pi$ orbitals of the tube are directed along the Y-axis.) Let us assume the electronic configuration ($2\pi_x^{*1}$, $2\pi_y^{*0}$) in an isolated NO molecule. Our analysis of *l,m*-projected local density of states (LDOS) shows that the $2\pi_x$*(NO) orbital does not interact significantly with tube states, that it remains half-filled, and that it contributes to the spin polarization of the complex. $2\pi_y$*(NO), which is degenerate with $2\pi_x$*(NO) in an isolated NO molecule, is now located above the conduction band edge. [See Scheme 1.] This is because it becomes a $\sigma$* orbital for the O−$Si_1$ and N−$C_1$ bonds. In short, the spin polarization of the complex can be attributed to that in the $\pi$*(NO) orbital, which is perpendicular to the N−$C_1$ and O−$Si_1$ bonds.

Figures S1 and S2 show the change of the band structure after the adsorption of one and two NO molecules. As was mentioned in the previous paragraph, a new state is introduced in the band gap of the pristine SiCNT when one NO molecule is adsorbed, which is $2\pi_x$*(NO) state weakly

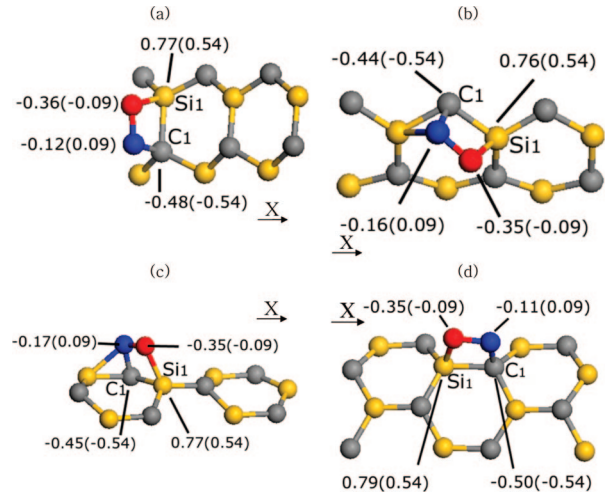



***Figure 6.*** Mulliken charges of atoms around the adsorption site for SiCNT-NO complexes depicted in Figure 1: configurations E (a) and Z (b) of the (5,5) tube and configurations Z (c) and A (d) of the (8,0) tube. For the better understanding, the former two figures, (a) and (b), are rotated along the tube axis (= X), while others have the same orientations as in Figure 1.

interacting with a $\pi$(tube) state. This state is located below the Fermi level for the spin-up state. On the other hand, it is located above the Fermi level for the spin-down state, causing the spin-polarization of the complex. When two NO molecules are adsorbed, two $2\pi_x$*(NO) states are introduced from two NO molecules. As a result, the band gap of the system decreases from those of pristine SiCNTs after the adsorption. However, Figures S1 and S2 indicate that there is no definite trend in the change of the gap as a function of NO molecules adsorbed.

A separate PBEPBE/6−31G(d) calculation using the GAUSSIAN03 program shows that both the N and O atoms have negative values of the Mulliken charge. For example, the charges are −0.12 and −0.36, respectively, in the E configuration of the (5,5) SiCNT. Upon complex formation, there is a net charge transfer of 0.46−0.52*e* from the tube to the molecule. Figure 6 shows Mulliken charges of atoms around the adsorption site for configurations of (5,5) and (8,0) SiCNT-NO complexes depicted in Figure 1. It shows that approximately one-half of the transferred electrons come from $Si_1$.

Figure S3 shows four other initial configurations of the (5,5) SiCNT-NO complex considered in this work. In configuration H1, which corresponds to S3(a), we built an initial O−$Si_1$ bond perpendicular to the tube surface. Similarly, we introduced an initial N−$C_1$ bond perpendicular to the tube surface in configuration H2. In configurations H3 and H4, which correspond to Figure S3(c),(d), the NO molecule is located diagonal to a hexagon of the tube. In all of them, initial bond lengths of O−Si and N−C bonds were chosen to be similar to those in stable configurations shown in Table 1. Two of them, which correspond to configurations H3 and H4, result in the same final structure in which the NO molecule is chemisorbed in such a way that a Si−N bond is formed. [See Figure S4(a).] However, its binding energy (= −0.34 eV) is smaller than those of configurations E and Z shown in Table 1. This is manifested in the Si−N

***Table 2.*** Energetic and Geometric Parameters for Various Configurations of CNT-NO and BNNT-NO Complexes in Which the Nanotubes Have (5,5) and (8,0) Chiral Indices

| | CNT | | | | BNNT | | | |
|---|---|---|---|---|---|---|---|---|
| | (5,5)[e] | | (8,0)[e] | | (5,5)[e] | | (8,0)[e] | |
| configuration | E | Z | Z | A | E | Z | Z | A |
| $E_b$[a] (eV) | 2.07 | 2.08 | 2.03 | 1.71 | 1.59 | 1.77 | 1.62 | 1.57 |
| $l$ (O−$C_1$)[b] (Å) | 1.49 | 1.55 | 1.53 | 1.52 | 1.52 | 1.58 | 1.56 | 1.56 |
| $l$ (N−$C_2$)[c] (Å) | 1.49 | 1.50 | 1.49 | 1.49 | 1.51 | 1.58 | 1.53 | 1.57 |
| $l$ (O−N)[d] (Å) | 1.39 | 1.37 | 1.39 | 1.38 | 1.35 | 1.34 | 1.35 | 1.35 |
| | (1.17) | (1.17) | (1.17) | (1.17) | (1.17) | (1.17) | (1.17) | (1.17) |

[a] Binding energy of one NO molecule on the surface of the CNT and BNNT. [b] O−$C_1$ distance, where $C_1$ is the carbon atom of the CNT above which the oxygen atom of the NO molecule is located. In BNNT-NO complexes, $C_1$ denotes a boron atom of the BNNT instead. [c] N−$C_2$ distance, where $C_2$ is the carbon atom of the CNT bonded to $C_1$ above which the nitrogen atom of the NO molecule is located. In BNNT-NO complexes, $C_2$ denotes a nitrogen atom of the BNNT instead. [d] N−O bond length in the NO molecule. The numbers in parentheses denote the corresponding length in an isolated molecule. [e] Chiral index.

bond length (=2.02 Å) which is appreciably larger than Si−O bond lengths (∼1.80 Å) in configurations E and Z. Relaxations of other initial structures result in physisorption ($|E_b| < 0.1$ eV) in which no chemical bond is formed between the tube and the NO molecule.

Figure S5 also shows four other initial configurations (=G1−G4) of the (8,0) SiCNT-NO complex, each of which is characterized by the geometrical feature which is the same as the corresponding one for the (5,5) complex already described. Only configurations G3 and G4 result in the chemisorption ($E_b = -0.37$ eV) shown in Figure S4(b), in which the local geometry around the adsorption site is similar to that of configuration H3 of the (5,5) complex.

Now we investigate the adsorption of an NO molecule on CNTs and BNNTs. In contrast to SiCNTs, these nanotubes exhibit highly endothermic binding to NO, which is even more pronounced for CNTs (Table 2). These binding energy data predict that the chemisorption of NO molecules is practically impossible on these tubes. A separate calculation shows that there are barriers for the desorption of the NO molecule in all the configurations of the complex, indicating that the complex indeed corresponds to a metastable state. This observation suggests that barriers for the adsorption are much higher than the value (=0.026 eV) of $kT$ at room temperature. In fact, our separate calculation shows that the barriers for the adsorption are greater than 1.86 and 1.67 eV for configuration E of (5,5) and configuration A of (8,0) of CNT-NO complexes. Similarly, corresponding barriers for BNNT-NO complexes are greater than 1.40 and 1.32 eV, respectively.

Table 2 shows that upon adsorption of the NO molecule on CNTs, the O−$C_1$ and N−$C_2$ bonds that form are weaker than single bonds. (Here, $C_1$ and $C_2$ correspond to the carbon atoms of CNTs bonded to each other and on which the NO molecule sits.) The O−N bond of the NO molecule is also weakened to a single bond. A similar analysis holds for the complex involving BNNTs. We want to recall that our calculation is focused on the chemisorption of the molecule, not on its physisorption as was recently investigated by Rafati et al.[11] This difference is manifested in the O−$C_1$ distance (1.49−1.55 Å) of the CNT-NO complex obtained from our calculation, which is much shorter than that (∼3.15 Å) from their calculation.

Next, we examine the adsorption of an NNO molecule on the surface of SiCNTs. Table 3 shows the binding energy of the molecule in various configurations. In the E, Z, and A configurations, there is appreciable NNO-SiCNT binding, although this is weaker than for the NO molecule. Figure 7 shows that these configurations are characterized by the formation of five-membered rings around the equatorial, zigzag, and axial adsorption sites, depending on the configuration. (Note that atoms in the molecule $N_1$−$N_2$−O are labeled $N_1$, $N_2$, and O.) This ring formation is achieved by (1) the formation of O−$Si_1$ and N−$C_1$ single bonds, (2) bending of the $N_1$−$N_2$−O bond in such a way that the linear configuration of the isolated molecule is destroyed, and (3)

***Table 3.*** Energetic and Geometric Parameters for Various Configurations of SiCNT-NNO Complexes in Which the SiCNTs Have (5,5) and (8,0) Chiral Indices

| | chiral index | | | | | |
|---|---|---|---|---|---|---|
| | (5,5) | | | (8,0) | | |
| configuration | E | Z | C | Z | A | C |
| $E_b(1)$[a] (eV) | −0.44 | −0.40 | 0.66 | −0.43 | −0.56 | 0.56 |
| $E_{gap}$[b] | 1.99 | 1.84 | | 1.33 | 1.29 | |
| | (2.18) | (2.18) | | (1.35) | (1.35) | |
| $l$ ($N_1$−$C_1$)[c] (Å) | 1.52 | 1.52 | | 1.52 | 1.51 | |
| $l$ (O−$Si_1$)[d] (Å) | 1.75 | 1.74 | | 1.75 | 1.74 | |
| $l$ ($C_1$−$Si_1$)[e] (Å) | 1.96 | 1.93 | | 1.95 | 1.93 | |
| | (1.79) | (1.79) | | (1.79) | (1.79) | |
| $l$ ($N_1$−$N_2$)[f] (Å) | 1.25 | 1.25 | | 1.25 | 1.25 | |
| | (1.15) | (1.15) | | (1.15) | (1.15) | |
| $l$ ($N_2$−O)[g] (Å) | 1.39 | 1.40 | | 1.39 | 1.41 | |
| | (1.20) | (1.20) | | (1.20) | (1.20) | |
| $\theta$($N_1$−$N_2$−O)[h] | 117.4 | 117.0 | | 117.3 | 117.1 | |
| | (180) | (180) | | (180) | (180) | |
| $q$ (NNO)[i] | −0.48 | −0.48 | | −0.49 | −0.47 | |
| | (0) | (0) | | (0) | (0) | |

[a] Binding energy of one NNO molecule on the surface of the SiCNT. [b] The band gap of the SiCNT-NNO complex. [c] $N_1$−$C_1$ distance, where $C_1$ is the carbon atom of the SiCNT above which $N_1$ is located. Two nitrogen atoms of the NNO molecule are defined by the bonds $N_1$−$N_2$−O. [d] O−$Si_1$ distance, where O is the oxygen atom of the NNO molecule. $Si_1$ is the silicon atom of the SiCNT bonded to $C_1$ above which O is located. [e] $C_1$−$Si_1$ bond length. The numbers in parentheses denote the corresponding data in pristine tubes. [f] $N_1$−$N_2$ bond length of the NNO molecule. The numbers in parentheses denote the corresponding data for an isolated molecule. See footnote *c* for the definitions of $N_1$ and $N_2$. [g] $N_2$−O bond length of the NNO molecule. The numbers in parentheses denote the corresponding data in an isolated molecule. See footnote *c* for the definition of $N_2$. [h] Bond angle of the NNO molecule. [i] Total Mulliken charge of the NNO molecule. The numbers in parentheses denote the corresponding data in an isolated molecule.

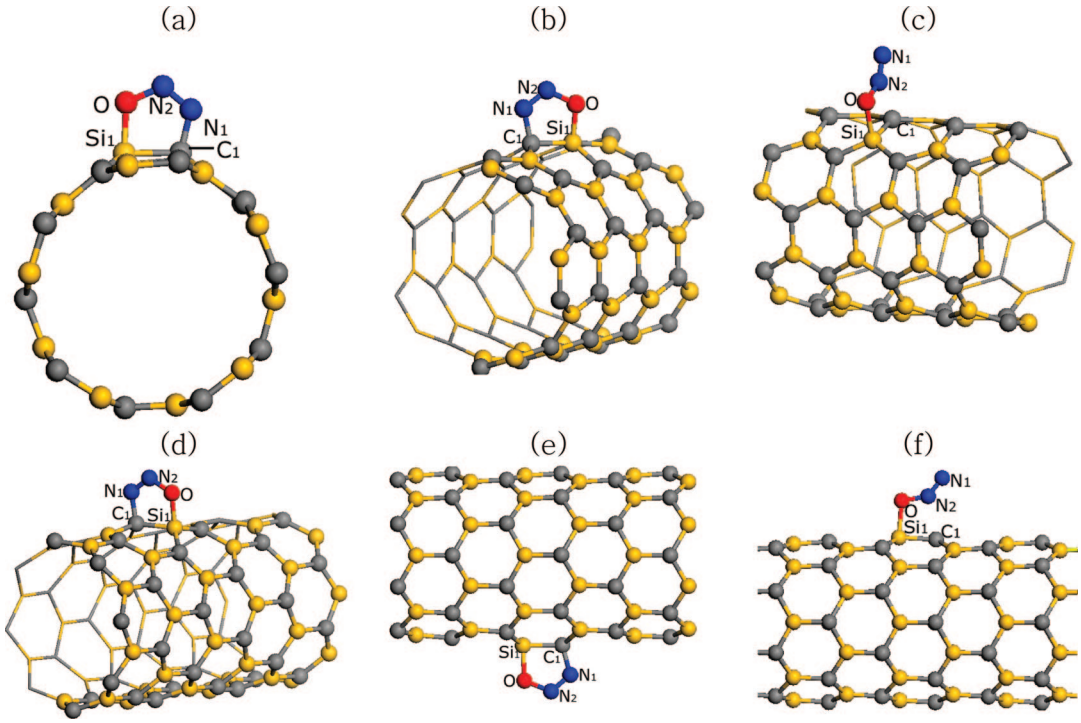



***Figure 7.*** Optimized structures for various configurations of SiCNT-NNO complexes: configurations E (a), Z (b), and C (c) of the (5,5) complex and configurations Z (d), A (e), and C (f) of the (8,0) complex. Atomic labels are also defined.

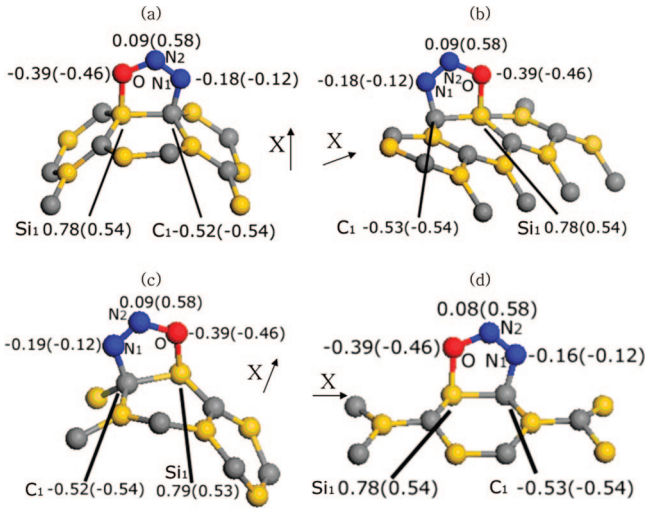



***Figure 8.*** Mulliken charges of atoms around the adsorption site for SiCNT-NNO complexes depicted in Figure 7: configurations E (a) and Z (b) of the (5,5) tube and configurations Z (c) and A (d) of the (8,0) tube. All of them have the same orientations as in Figure 7.

orientation of the NNO molecule such that its molecular plane is perpendicular to the tube surface.

Figure 7(c) shows another configuration (C) in which the five-membered ring is not formed. Here, the NNO molecule adopts a different orientation with respect to the tube. Table 3 shows that this configuration is much less stable than others for both (5,5) and (8,0) SiCNTs. Indeed, the positive values of binding energy indicate that the configuration corresponds to a metastable state. Thus we do not consider it in further analyses.

Table 3 shows that lengths of the O−$Si_1$ and $N_1$−$C_1$ bonds are similar to the corresponding values in a stable SiCNT-

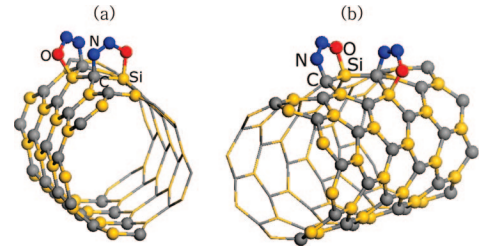



***Figure 9.*** Optimized structures of (5,5) (a) and (8,0) (b) SiCNT-2NNO complexes.

NO complex. As already noted, the geometry of the NNO molecule shows a significant degree of deformation, as evidenced by large deviations in the $N_1$−$N_2$−O bond angle from 180°. In fact, the bond angle (∼117°) shown in the table indicates that $N_2$ is nearly $sp^2$-hybridized, which is achieved by its protrusion from the circumference of the tube. The $N_1$−$N_2$ and $N_2$−O bonds are elongated by different amounts in such a way that the molecule has N1═N2−O bonds. (We recall that the bond orders of $N_1$−$N_2$ and $N_2$−O bonds are 2.5 and 2 in an isolated NNO molecule.) The $C_1$−$Si_1$ bond is also elongated by an appreciable amount (∼0.17 Å) compared to that in pristine tubes. As shown in Table 3, Mulliken charge analysis shows a charge transfer (∼0.48*e*) from the tubes to the molecule, which is comparable to what is seen with the SiCNT-NO complex. Figure 8 shows a more detailed analysis of the Mulliken charges for atoms around the adsorption site. Comparing the charges on the atoms around the adsorption site with the corresponding values in the isolated molecule and tube shows that most of the transferred charge is concentrated on the central nitrogen atom, $N_2$, making it almost neutral. A careful analysis of the changes in charge on $Si_1$, $N_2$, and O upon complex formation shows that most of the charge transfer

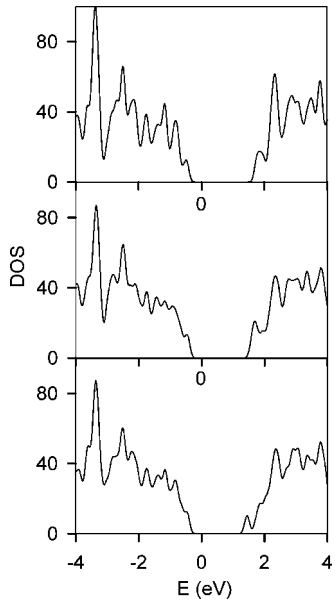



***Figure 10.*** Comparison of the electronic density of states (DOS) for (5,5) SiCNT-NNO complexes: pristine tube (top), configuration E (middle), and configuration Z (bottom).

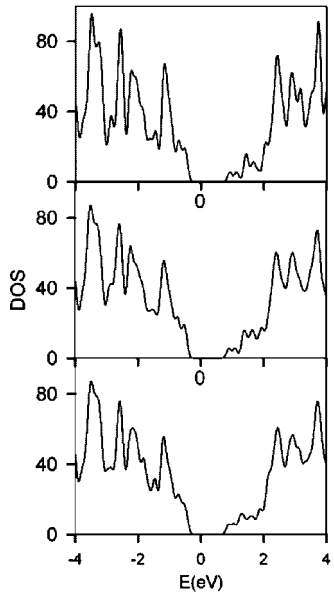



***Figure 11.*** Comparison of the electronic density of states (DOS) for (8,0) SiCNT-NNO complexes: pristine tube (top), configuration Z (middle), and configuration A (bottom).

from the tube to the molecule occurs from $Si_1$ to $N_2$ through O, allowing $N_2$ to have an unpaired electron in one of its $sp^2$ hybrid orbitals. Another transfer route running from $C_1$ to $N_2$ through $N_1$ makes a much smaller contribution to the charge transfer.

Figure S6 shows four additional configurations (K1−K4) of the (5,5) SiCNT-NNO complex investigated, each of which has a local geometry similar to the corresponding one for the SiCNT-NO complex shown in Figure S3. Except configuration K1, all of them result in the physisorption with the binding energy less than −0.02 eV upon the relaxation of geometry. Optimization of configuration K1 results in configuration C shown in Figure 7(c). Figure S7 also shows additional configurations (L1−L4) of the (8,0) SiCNT-NNO complex. Similar to the case of their correspondents for the (5,5) complex, all of them exhibit physisorption with the binding energy less than −0.01 eV except configuration L1. Optimization of configuration L1 also results in configuration C shown in Figure 7(f).

We have also investigated the physicochemical properties of SiCNTs when two NNO molecules are adsorbed on the surface. In order to do this, we take configurations E of the (5,5) SiCNT-NNO complex and A of the (8,0) complex, which are the most stable ones when one NO molecule is adsorbed. We also find that two NNO molecules tend to bind to adjacent sites rather than being far apart, as evidenced by the difference in the binding energy of 0.12 eV and 0.10 eV for (5,5) tube (8,0) tube, respectively. Figure 9 shows that the (5,5) SiCNT-2NNO complex adopts E and Z configurations, while the (8,0) complex adopts Z and A configurations of the two molecules. In addition, the adsorption is also cooperative. For the (5,5) tube, the binding energy of the first and the second molecule is −0.44 eV and −0.68 eV, respectively. For the (8,0) tube, the corresponding value is −0.56 eV and −0.72 eV, respectively.

Figures 10 and 11show that the electronic density of states (DOS) for stable configurations of SiCNTs does not differ much from that of the pristine tube near the Fermi level. One may note that the band gap decreases slightly (<0.2 eV) upon adsorption. In addition, there is little shift in the Fermi level after adsorption, indicating that the rigid charge transfer model does not apply to this system. A separate analysis shows that the top of the valence band is mostly composed of $\pi$ states of the tube in which electron densities are concentrated on carbon atoms. Doubly degenerate HOMOs of the isolated NNO molecule, which split into $\sigma$ and $\pi$ states upon adsorption on the tube, do not affect the electronic structure of the complex at the top of the valence band, since they are located more than 1.8 eV below the Fermi level. Likewise, doubly degenerate LUMOs of the NNO molecule also interact with conduction bands at 0.3 eV above their bottom. Figures S8 and S9 show band structures of the (5,5) and (8,0) complexes as the number of adsorbed molecules are varied.

Table 4shows that the adsorption of the NNO molecule on the surface of CNTs or BNNTs is also unfavorable, as indicated by significantly positive values of the binding energy. Once adsorbed, the geometry of the adsorbed complex is similar to that of the corresponding SiCNT-NNO complex, in that five-membered rings are formed at the adsorption site.

## 4. Conclusions

Using a theoretical method based on first principles, we have investigated the chemisorptions of NO and NNO molecules on SiCNTs, CNTs, and BNNTs. To the authors' knowledge, this is the first theoretical work examining the possibility of chemisorbing the notorious nitrogen oxide pollutants on these nanotubes in detail. Our calculations show that they can be adsorbed on silicon carbide nanotubes (SiCNTs) with an appreciable binding energy. A detailed investigation of energetics, electronic structure, and magnetic properties of the adsorbed complexes was made, which gives useful results for NOx sensing. On the other hand, chemisorptions on carbon nanotubes (CNTs) or boron nitride nanotubes (BNNTs) are severely

***Table 4.*** Energetic and Geometric Parameters for Various Configurations of CNT-NNO and BNNT-NNO Complexes in Which the Nanotubes Have (5,5) and (8,0) Chiral Indices

| | CNT | | | | BNNT | | | |
|---|---|---|---|---|---|---|---|---|
| | (5,5)[g] | | (8,0)[g] | | (5,5)[g] | | (8,0)[g] | |
| configuration | E | Z | Z | A | E | Z | Z | A |
| $E_b$[a] (eV) | 1.70 | 1.50 | 1.58 | 1.07 | 1.50 | no binding | no binding | 1.29 |
| $l$ (O−$C_1$)[b] (Å) | 1.54 (1.75) | 1.51 (1.74) | 1.53 (1.75) | 1.49 (1.74) | 1.59 (1.75) | | | 1.56 (1.74) |
| $l$ ($N_1$−$C_2$)[c] (Å) | 1.57 (1.52) | 1.55 (1.52) | 1.57 (1.52) | 1.52 (1.51) | 1.64 (1.52) | | | 1.63 (1.51) |
| $l$ (O−$N_2$)[d] (Å) | 1.37 (1.20) | 1.40 (1.20) | 1.38 (1.20) | 1.40 (1.20) | 1.32 (1.20) | | | 1.33 (1.20) |
| $l$ ($N_1$−$N_2$)[e] (Å) | 1.23 (1.15) | 1.23 (1.15) | 1.23 (1.15) | 1.23 (1.15) | 1.23 (1.15) | | | 1.23 (1.15) |
| $\theta$($N_1$−$N_2$−O)[f] | 116.3 (180) | 114.6 (180) | 115.3 (180) | 114.4 (180) | 119.9 (180) | | | 119.4 (180) |

[a] Binding energy of one NNO molecule on the surface of the CNT or BNNT. [b] O−$C_1$ distance, where $C_1$ is the carbon atom of the CNT above which O is located. In the BNNT, $C_1$ represents a boron atom of the tube. [c] $N_1$−$C_2$ distance, where $C_2$ is the carbon atom of the CNT above which $N_1$ is located. In the BNNT, $C_2$ represents a nitrogen atom of the tube. Two nitrogen atoms of the NNO molecule are defined by the bonds $N_1$−$N_2$−O. [d] $N_2$−O bond length of the NNO molecule. The numbers in parentheses denote the corresponding data in an isolated molecule. See footnote *c* for the definition of $N_2$. [e] $N_1$−$N_2$ bond length of the NNO molecule. The numbers in parentheses denote the corresponding data in an isolated molecule. See footnote *c* for the definitions of $N_1$ and $N_2$. [f] Bond angle of the NNO molecule. [g] Chiral index.

endothermic, resulting in metastable complexes. In some cases, not even these metastable states are predicted to exist. In conjunction with Rafati et al. result on the physisorption of the NO molecule on CNTs, these results suggest that only SiCNTs would be practically useful for the removal of NO and NNO molecules among three nanotubes. Change of magnetic properties will be particularly useful for sensing the amount of NO molecules adsorbed on SiCNTs.

**Acknowledgment.** This work was supported by a Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-313-C00334), in which the main calculations were performed using the supercomputing resource of the Korea Institute of Science and Technology Information (KISTI).

**Supporting Information Available:** Figures S1−S9. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Janzen, E.; Kordina, O.; Henry, A.; Chen, W. M.; Son, N. T.; Monemar, B.; Sorman, E.; Berganman, P.; Harris, C. I.; Yakimova, R.; Tuominen, M.; Konstantinov, A. O.; Hallin, C.; Hemmingsson, C. *Phys. Scr.* **1994**, *T54*, 283.

(2) Sun, X.-H.; Li, C.-P.; Wong, W.-K.; Wong, N.-B.; Lee, C.-S.; Lee, S.-T.; Teo, B.-K. *J. Am. Chem. Soc.* **2002**, *124*, 14464.

(3) (a) Menon, M.; Richter, E.; Mavrandonakis, A.; Froudakis, G.; Andriotis, A. N. *Phys. Rev. B* **2004**, *69*, 115322. (b) Miyamoto, Y.; Yu, B. D. *Appl. Phys. Lett.* **2002**, *80*, 586. (c) Mavrandonakis, A.; Froudakis, G. E.; Schnell, M.; Muhlhauser, M. *Nano Lett.* **2003**, *3*, 1481.

(4) Zhao, J.-x.; Ding, Y.-b. *J. Phys. Chem. C* **2008**, *112*, 2558.

(5) Mopurmpakis, G.; Froudakis, E.; Lithoxoos, G. P.; Samios, J. *Nano Lett.* **2006**, *6*, 1581.

(6) Zel'dovich, Ya. B.; Sadovnikov, P. Ya.; Frank-Kamenetskii, D. A. *Oxidation of Nitrogen in Combustion*; Acad. of Sci. USSR: Moscow, 1947.

(7) Shelef, M. *Chem. Rev.* **1995**, *95*, 209.

(8) McCue, J. T.; Ying, J. Y. *Chem. Mater.* **2007**, *19*, 1009.

(9) Boon, E. M.; Marletta, M. A. *J. Am. Chem. Soc.* **2006**, *128*, 10022.

(10) Huang, Y.; Ho, W.; Lee, S.; Zhang, L.; Li, G.; Yu, J. C. *Langmuir* **2008**, *24*, 3510.

(11) Rafati, A. A.; Hashemianzadeh, S. M.; Nojini, Z. B. *J. Phys. Chem. C* **2008**, *112*, 3597.

(12) Chen, H.-T.; Musaev, D. G.; Irle, S.; Lin, M. C. *J. Phys. Chem. A* **2007**, *111*, 982.

(13) Gronbeck, H.; Hellman, A.; Gavrin, A. *J. Phys. Chem. A* **2007**, *111*, 6062.

(14) (a) Broqvist, P.; Panas, H.; Fridell, E.; Persson, H. J. *Phys. Chem. B* **2002**, *106*, 137. (b) Schneider, W. F.; Hass, K. C.; Miletic, M.; Gland, J. L. *J. Phys. Chem. B* **2002**, *106*, 7405. (c) Miletic, M.; Gland, J. L.; Hass, K. C.; Schneider, W. F. *J. Phys. Chem. B* **2003**, *107*, 157. (d) Karlsen, E. J.; Nygren, M. A.; Petterson, G. M. *J. Phys. Chem. B* **2003**, *107*, 7795. (e) Broqvist, P.; Gronbeck, H.; Fridell, E.; Panas, I. *J. Phys. Chem. B* **2004**, *108*, 3523. (f) Xu, S. C.; Irle, S.; Musaev, D. G.; Lin, M. C. J. *Phys. Chem. B* **2006**, *110*, 21135.

(15) Kresse, G.; Hafner, *Phys. Rev. B* **1993**, *47*, 558.

(16) Kresse, G.; Joubert, D. *Phys. Rev. B* **1999**, *59*, 1758.

(17) Kresse, G.; Furthmuller, J. *Phys. Rev. B* **1996**, *54*, 11169.

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

(19) Throughout this work, the accuracy of calculations on the electronic LDOS was carefully confirmed by separate band structure calculation.

(20) Huheey, J. E.; Keiter, E. A.; Keiter, R. L. *Inorganic Chemistry*, 4th ed.; HarperCollins College Publishing: New York, 1993; Chapter A-28.

CT800273C

JCTC Journal of Chemical Theory and Computation

# A Systematic Comparison of Pairwise and Many-Body Silica Potentials

Sterling Paramore, Liwen Cheng, and Bruce J. Berne*

*Department of Chemistry, Columbia University, 3000 Broadway, Mail Code 3103, New York City, New York 10027*

Received June 24, 2008

**Abstract:** The role of many-body effects in modeling silica was investigated using self-consistent force matching. Both pairwise and polarizable classical force fields were developed systematically from ab initio density functional theory force calculations, allowing for a direct comparison of the role of polarization in silica. It was observed that the pairwise potential performed remarkably well at reproducing the basic silica tetrahedral structure. However, the Si−O−Si angle that links the silica tetrahedra showed small but distinct differences with the polarizable potential, a result of the inability of the pairwise potential to properly account for variations in the polarization of the oxygens. Furthermore, the transferability of the polarizable potential was investigated and suggests that additional forces may be necessary to more completely describe silica annealing.

## I. Introduction

The development of accurate classical potentials for silica has been the subject of intense research for over 30 years.[1] While enormous advances in computing technology have recently enabled researchers to perform ab initio molecular dynamics simulations of silica,[2−4] the time and length scales accessible to such calculations are still fairly limiting, necessitating the use of fast classical methods. Some of the earliest and still widely used classical models of silica are fixed charge models, where the Si and O atoms are treated as point charges that experience Coulombic interactions in addition to short-range pairwise forces.[5,6] Parameters for these models were obtained either through fitting them to ab initio data for small silica clusters, adjusting the parameters to reproduce known experimental results, or some combination of these methods.

Many-body effects have long been believed to be of critical importance in governing the structure and dynamics of amorphous silica.[7−12] Of particular interest are the role of many-body effects in silica subject to heterogeneous environments, such as those encountered at surfaces and liquid−silica interfaces.[8,10,13−15] Three-body angle terms have been added to the potential and parametrized to reproduce the angle distributions deduced from experiments.[7,12,13] The three-body potentials are found to be in closer agreement with experimental angle distributions than the pairwise force fields. However, the pairwise potentials did not include any information about experimental angles in the parametrization. It is thus not clear whether the pairwise potentials are inherently unable to reproduce the correct angle distributions, or if they fail just because that information was not included in the parametrization. When comparing the behavior of different force fields using different parametrization procedures, it is not always obvious whether a given type of interaction is actually necessary for an accurate description of the system. In order to directly compare force fields and determine whether certain interactions are necessary, the force field construction needs to be systematically derived from a common data set.

Force matching was originally developed heuristically as an algorithm for generating a classical force field for aluminum from a set of ab initio calculations.[16] It has since been used to develop force fields for numerous other systems including water,[17] liquid hydrogen fluoride,[18] room temperature ionic liquids,[19] and bulk amorphous silica.[11] Force matching as an algorithm has provided remarkably robust and accurate classical force fields, but also has a solid foundation in statistical mechanics.[20] The algorithm involves constructing a representation of a classical force field (which may include pairwise, 3-body, or polarizable terms) and tuning the parameters of the potential to reproduce forces obtained from ab initio calculations. In force matching, any

* Corresponding author. Email: bb8@columbia.edu.

10.1021/ct800244q CCC: $40.75 
Published on Web 09/24/2008

given representation of the classical force field is parametrized using the same set of ab initio data, enabling a systematic comparison of different force fields.

Force matching has traditionally been performed by defining a function with a small number of parameters (e.g., the Lennard-Jones size and energy parameters) and then fitting those parameters to the ab initio forces.[11,16,19] One of the primary difficulties associated with this method is that the forces are nonlinearly dependent on the parameters, which dramatically increases the difficulty of the fitting procedure. In addition, the chosen functional form may or may not be a good description of the actual forces involved. To remedy this problem, the force field needs to be both flexible and linearly dependent on its parameters. This has recently been accomplished by defining the force as a spline,[17,18] or as a discrete tabularized function.[20]

One of the primary objectives of this work is to use force matching to determine the importance of including many-body effects in silica. Tangney and Scandolo[11] (TS) constructed a many-body silica model using nonlinear force matching where the oxygens are treated as polarizable atoms. The TS potential has been shown to be very successful at reproducing many of the structural and dynamical properties of several crystalline phases of silica.[21] However, it is plausible that the success of the TS potential is due to the force matching parametrization procedure and not due to the form of the force field. In other words, it is not clear whether this success is a consequence of the explicit many-body nature of the force field, or if the parametrization procedure could have performed equally well using a different representation of the force field. This article investigates whether a pairwise potential, parametrized using the same force matching procedure, can reproduce the structural properties of the many-body silica force field. In addition, we have also examined the effects of using less restrictive functions to describe the pairwise interactions involved in the TS force field.

## II. Methods

**Force Matching.** The main principle behind force matching involves defining a residual, which is the ensemble average of the difference between the ab initio force (i.e., the Hellman−Feynman force) on atomic site $i$ for a given configuration, $\mathbf{F}_i^{\mathrm{AI}}$, and the force given by the classical force field, $\mathbf{F}_i^{\mathrm{FF}}$, summed over all atomic sites

$$\chi^2 = \frac{1}{3N}\sum_i \langle |\mathbf{F}_i^{\mathrm{FF}} - \mathbf{F}_i^{\mathrm{AI}}|^2 \rangle \tag{1}$$

Force matching involves finding the classical force field, $\mathbf{F}_i^{\mathrm{FF}}$, that minimizes the residual, or the force field that is the best fit to the ab initio forces. The residual is zero when the classical force field perfectly reproduces the ab initio forces for all configurations in the ensemble. Of course, such a force field would be exceedingly complex and would likely remove any computational advantages to using a classical force field.

There have been several approaches to defining an appropriate potential and minimizing the residual in eq 1. Typically, the force depends nonlinearly on the parameters of the force field, in which case nonlinear minimization methods, such as simulated annealing, must be used.[11,16,19] However, these methods cannot guarantee that the residual is minimized and can require a relatively large amount of computational effort. As will be discussed below, it is possible to represent a general short-range pairwise force field as one that depends linearly on its parameters. Linear dependence allows one to use efficient computational methods to solve for the parameters and guarantees minimization of the residual (as long as the problem is not ill-conditioned).[22] Somewhat surprisingly, force-matched force fields that contain only central pairwise terms have proven to be remarkably good at reproducing structural properties of the systems simulated with ab initio dynamics.[17,18] The accuracy of pairwise potentials generated using force matching often exceeds expectations in that behavior thought to be critically dependent on many-body effects can sometimes be reproduced using a force-matched pairwise potential.[23]

For a classical force field that is short-range, pairwise, and central, the force on atom $i$ of type $\alpha$ can be written

$$\mathbf{F}_{i_\alpha}^{\mathrm{FF}} = \sum_{\beta=1}^{N_{\mathrm{T}}} \sum_{j_\beta \neq i_\alpha}^{N_\beta} f_{\alpha\beta}(r_{ij})\hat{\mathbf{r}}_{ij} \tag{2}$$

where $N_{\mathrm{T}}$ is the number of types of atoms in the system, $N_\beta$ is the number of atoms of type $\beta$, and $f_{\alpha\beta}(r_{ij})$ determines the magnitude of the force between two atoms. In this article, greek letters will be used to indicate the type of the atomic species and lowercase letters will indicate a particular atom. To reduce notational clutter, type subscripts on individual atoms (e.g., $i_\alpha$) will be omitted when the atom index appears as the argument of the sum, since the type can be inferred from the summation index. While there are numerous and varied forms that the $f_{\alpha\beta}(r_{ij})$ term can take (e.g., Lennard-Jones, Born−Mayer, splines, etc.), we follow the work of Noid et al.[20] and discretize the pairwise force according to

$$f_{\alpha\beta}(r_{i_\alpha j_\beta}) = \sum_{d=1}^{N_{\mathrm{d}}} f_{\alpha\beta}^d \delta_d(r_{i_\alpha j_\beta} - r_d) \tag{3}$$

where $f_{\alpha\beta}^d$ are the parameters of the force field and $\delta_{\mathrm{d}}(r)$ is a discrete delta function defined as

$$\delta_{\mathrm{d}}(r) = \begin{cases} 1 & -\Delta r/2 \leq r < \Delta r/2 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Here, $\Delta r$ determines the resolution of the discretization and $N_{\mathrm{d}}$ is the number of discrete gridpoints used to describe the force, giving a total of $N_{\mathrm{d}} \times N_{\mathrm{T}} \times (N_{\mathrm{T}} - 1)/2$ parameters. Discretizing the force in eq 3 makes the total pairwise force *linearly* dependent on its parameters. The important consequence of this relationship is that minimization of the residual of eq 1 can be written as a linear equation

$$\mathbf{Ax} = \mathbf{b} \tag{5}$$

where the elements of the matrix $\mathbf{A}$ are

$$A_{lm} = \sum_{\alpha=1}^{N_T} \sum_{i_\alpha=1}^{N_\alpha} \left\langle \frac{\partial \mathbf{F}_{i_\alpha}^{FF}}{\partial x_l} \cdot \frac{\partial \mathbf{F}_{i_\alpha}^{FF}}{\partial x_m} \right\rangle \quad (6)$$

and the elements of the vector **b** are

$$b_l = \sum_{\alpha=1}^{N_T} \sum_{i_\alpha=1}^{N_\alpha} \left\langle \mathbf{F}_{i_\alpha}^{AI} \cdot \frac{\partial \mathbf{F}_{i_\alpha}^{FF}}{\partial x_l} \right\rangle \quad (7)$$

and $x_l$ is parameter $l$ of the force field (i.e., $x_l = f_{\alpha\beta}^d$ for a given unique set of $\alpha$, $\beta$, and $d$).

For the sake of discussing in more detail the nature of the force matching equations, consider a single-component system that is described by a pairwise force field (eqs 23). In this case, the matrix **A** and vector **b** of the force matching equations can be written

$$A_{lm} = \left\langle \sum_{j\neq i} \sum_{k\neq i} \delta_D(r_{ij} - r_l)\delta_D(r_{ik} - r_m)\hat{\mathbf{r}}_{ij} \cdot \hat{\mathbf{r}}_{ik} \right\rangle \quad (8)$$

$$b_l = \left\langle \sum_{j\neq i} \delta_D(r_{ij} - r_l)\mathbf{F}_i^{AI} \cdot \hat{\mathbf{r}}_{ij} \right\rangle \quad (9)$$

As aptly discussed by Noid et al.,[20] **A** contains two- and three-body correlation information and **b** is related to the potential of mean force between the atoms. The force matching algorithm takes as input this correlation information and, through solving eq 5, derives the underlying force field which gives that correlation information. In many ways, force matching can be thought of as the inverse of molecular dynamics. In molecular dynamics, one starts with a given force field and then calculates trajectories in order to obtain structural information (e.g., the radial distribution function). In force matching, one starts with the structural information needed to construct **A** and **b** and then solves eq 5 to obtain the force field parameters.

**A Many-Body Force Field.** While it has been shown that pairwise force-matched force fields can perform remarkably well in many situations, it is expected that more complex force fields may be necessary in some systems. The Tangney and Scandolo[11] (TS) potential explicitly incorporates many-body effects by treating the oxygen atoms as polarizable ions. It was parametrized using a nonlinear force matching procedure, where the forces were obtained from DFT calculations of configurations sampled from 3000 K liquid silica simulations. While the TS force field has been shown to give good agreement with experimental data for various crystalline phases of silica,[21] it has not been used to examine amorphous structures. Our attempts to use the TS force field construct 300 K amorphous silica structures via simulated annealing produced a large number of anomalous two-membered silica rings.[24] Two-membered rings are formed by edge-sharing silica tetrahedra and are often observed as defect sites on silica surfaces.[8,25−28] However, there is no experimental evidence and no other reported simulations that support their existence in the bulk at room temperature. These two-membered ring artifacts observed are due to the presence of an attractive Si−Si interaction at short distances that is sampled during the annealing process. While the artifacts could be removed by adding hard restraining potentials to counteract the attractions, they needed to be placed so far out that the entire first peak of the Si−Si radial distribution function only sampled the hard restraining potential and not the original potential.

One of the goals of this work is to investigate whether the artifacts in the potential could be a consequence of the force matching procedures used by Tangney and Scandolo. It can be particularly difficult to determine whether the global minimum of the residual is found in nonlinear force matching. This is not a problem with the linear force matching method described above. In addition, by force matching onto a strict functional form, it is possible that some regions of the force field may fit the ab initio forces better than others. To determine if the formation of the two-membered ring artifacts could be a consequence of either the nonlinear minimization method failing to find the global residual minimum or a poor fit of the ab initio forces in certain regions of the force field, we have reparameterized the TS model using linear form-free tabularized potentials.

The energy of the TS model is given as a sum of short-range pairwise, Coulombic, dipole polarization, and short-range charge−dipole screening terms

$$U^{TS} = U^{PW} + U^C + U^\mu + U^S \quad (10)$$

In our reparameterization of the TS potential, the pairwise term will be such that its force is given in the discrete tabularized form of eqs 2 and 3. The Coulombic term accounts for the classical electrostatic interaction between all charges and dipoles[29]

$$U^C = \frac{1}{2}\sum_i \sum_{j\neq i} \left[ \frac{Q_i Q_j}{r_{ij}} - 2Q_j \frac{\mathbf{r}_{ij} \cdot \boldsymbol{\mu}_i}{r_{ij}^3} + \boldsymbol{\mu}_i \cdot \mathbf{T}_{ij} \cdot \boldsymbol{\mu}_j \right] \quad (11)$$

where $Q_i$ is the charge of atom $i$, $\boldsymbol{\mu}_i$ is the dipole associated with atom $i$, and $\mathbf{T}_{ij}$ is the dipole propagator, a second rank tensor

$$\mathbf{T}_{ij} = \left[ \mathbf{I} - 3\frac{\mathbf{r}_{ij}\mathbf{r}_{ij}^T}{\mathbf{r}_{ij}^2} \right] \frac{1}{r_{ij}^3} \quad (12)$$

The energy required to polarize an ion is $U^\mu = \sum_i \boldsymbol{\mu}_i^2/2\alpha$, where $\alpha$ is the dipole polarizability. The last energy term in the TS potential accounts for screening of the charge−dipole interactions that occurs at short distances between the charge on the silicon cation and the dipole on the oxygen anion[29−31]

$$U^S = -\sum_i \sum_{j\neq i} s(r_{ij}) Q_j \frac{\mathbf{r}_{ij} \cdot \boldsymbol{\mu}_i}{r_{ij}^3} \quad (13)$$

where $s(r_{ij})$ is some screening function.[32] The dipoles on the polarizable atoms are assumed to follow the Born−Oppenheimer surface and respond instantaneously to changes in the atomic configurations. In this case, the instantaneous values of the dipoles are found by minimizing the total energy with respect to the dipole position. The result is

$$\boldsymbol{\mu}_i = \alpha \sum_{j\neq i} \left[ Q_j \frac{\mathbf{r}_{ij}}{r_{ij}^3} - \mathbf{T}_{ij} \cdot \boldsymbol{\mu}_j + s(r_{ij}) Q_j \frac{\mathbf{r}_{ij}}{r_{ij}^3} \right] \quad (14)$$

which can be solved by iterating to self-consistency.

In our reparametrization of the TS model, the screening function is discretized in a manner similar to the pairwise force, eq 3. The parameters of the TS model include the short-range pairwise potential, the charge on the silicon and oxygen atoms, the dipole polarizability, and the screening function. The force is thus a nonlinear function of its parameters, since it depends on the product of charges, the product of charge and screening functions, and the polarizability term. It is possible to define a new set of parameters (e.g., $Q_{\alpha\beta} = Q_\alpha Q_\beta$) such that the force depends linearly on its parameters, use force matching to find these new parameters, and then reconstruct the original parameters using nonlinear methods.[17,18] However, we found that very small differences in the charge terms (e.g., due to sampling noise) can lead to large changes in the pairwise terms, without giving rise to a significant difference in the resulting structure. In some cases, it is even possible to leave out the charge terms altogether and only maintain short-range interactions without significantly affecting the structure. In effect, there is enough freedom in the pairwise terms to make up for any small deficiencies in the charges. For this reason, we have retained the original TS charge and polarizability parameters, and only reparametrized the short-range pairwise and screening terms.

The components of the force field that are dependent on the parameters to be fit are the pairwise force, $\mathbf{F}_{i_\alpha}^{\mathrm{PW}}$, given by eq 2, and the force due to charge−dipole screening, $\mathbf{F}_{i_\alpha}^{\mathrm{S}}$. The derivative of the multicomponent pairwise force with respect to its parameters, as required in eqs 6 and 7, is

$$\frac{\partial \mathbf{F}_{i_\alpha}^{PW}}{\partial f_{\gamma\epsilon}^{d}} = \delta_{\alpha\gamma}(1-\delta_{\alpha\epsilon})\sum_{k_\epsilon \neq i_\alpha}^{N_\epsilon} \delta_{\mathrm{D}}(r_{ik}-r_{d'})\hat{\mathbf{r}}_{ik} + \delta_{\alpha\epsilon}\sum_{k_\gamma \neq i_\alpha}^{N_\gamma} \delta_{\mathrm{D}}(r_{ik}-r_{d'})\hat{\mathbf{r}}_{ik} \quad (15)$$

The screening force is

$$\mathbf{F}_{i_\alpha}^{S} = \sum_{\beta=1}^{N_{\mathrm{T}}}\sum_{j_\beta \neq i_\alpha}^{N_\beta}\left[ s_{\alpha\beta}(r_{ij})(Q_\beta \mathbf{T}_{ij}\boldsymbol{\mu}_i - Q_\alpha \mathbf{T}_{ij}\boldsymbol{\mu}_j) + \frac{\mathrm{d}s_{\alpha\beta}(r_{ij})}{\mathrm{d}r_{ij}}\left( Q_\beta \frac{\mathbf{r}_{ij}\boldsymbol{\mu}_i}{r_{ij}^4}\mathbf{r}_{ij} - Q_\alpha \frac{\mathbf{r}_{ij}\boldsymbol{\mu}_j}{r_{ij}^4}\mathbf{r}_{ij}\right)\right] \quad (16)$$

The screening function is expressed in a discrete tabularized form similar to eq 3, in which case the derivative of the screening function can be obtained from

$$\frac{\mathrm{d}s_{\alpha\beta}(r_{ij})}{\mathrm{d}r_{ij}} = \frac{s_{\alpha\beta}(r_{ij}+\Delta r) - s_{\alpha\beta}(r_{ij})}{\Delta r} \quad (17)$$

$$= \frac{1}{\Delta r}\sum_{d=1}^{N_{\mathrm{d}}} s_{\alpha\beta}^{d}(\delta_{\mathrm{D}}(r_{ij}-r_{d-1}) - \delta_{\mathrm{D}}(r_{ij}-r_d)) \quad (18)$$

The screening force can then be written

$$\mathbf{F}_{i_\alpha}^{S} = \sum_{\beta=1}^{N_{\mathrm{T}}}\sum_{j_\beta \neq i_\alpha}^{N_\beta}\sum_{d=1}^{N_{\mathrm{d}}} s_{\alpha\beta}^{d}(\boldsymbol{\sigma}_{j_\beta i_\alpha}^{d} - \boldsymbol{\sigma}_{i_\alpha j_\beta}^{d}) \quad (19)$$

where

$$\boldsymbol{\sigma}_{j_\beta i_\alpha}^{d} = Q_\beta \mathbf{T}_{ji}\boldsymbol{\mu}_i \delta_{\mathrm{D}}(r_{ij}-r_d) + \frac{Q_\beta}{\Delta r}(\delta_{\mathrm{D}}(r_{ij}-r_{d-1}) - \delta_{\mathrm{D}}(r_{ij}-r_d))\frac{\mathbf{r}_{ij}\boldsymbol{\mu}_i}{r_{ij}^4}\mathbf{r}_{\mathrm{ij}} \quad (20)$$

The derivative of the multicomponent screening force with respect to its parameters is thus

$$\frac{\partial \mathbf{F}_{i_\alpha}^{S}}{\partial s_{\gamma\epsilon}^{d'}} = \delta_{\alpha\gamma}(1-\delta_{\alpha\epsilon})\sum_{k_\epsilon \neq i_\alpha}^{N_\epsilon}(\boldsymbol{\sigma}_{k_\epsilon i_\alpha}^{d'} - \boldsymbol{\sigma}_{i_\alpha k_\epsilon}^{d'}) + \delta_{\alpha\epsilon}\sum_{k_\gamma \neq i_\alpha}^{N_\gamma}(\boldsymbol{\sigma}_{k_\gamma i_\alpha}^{d'} - \boldsymbol{\sigma}_{i_\alpha k_\gamma}^{d'}) \quad (21)$$

Equations 15 and 21 are used to construct the matrix **A** of eq 6. The vector **b** of eq 7 is also constructed using the above results, but where all constant components of the force field (e.g., the Coulomb, long-range charge−dipole, and dipole−dipole components) are subtracted from the ab initio forces, $\mathbf{F}_{i_\alpha}^{\mathrm{AI}}$.

The derivatives of the short-range screening force depend on the instantaneous values of the dipoles, which are calculated by self-consistently solving eq 14, which in turn depends on the screening function. Self-consistent methods can also be used here to solve these circuitous dependencies. First, a guess at the screening function is made, from which the dipoles can be calculated. Force matching is used to calculate a new screening function, which can then be used to find new dipole positions. The process is repeated until the screening function no longer changes.

**Self-Consistent Force Matching.** In earlier force matching work,[16,17] configurations were sampled from ab initio molecular dynamics simulations. Such simulations are particularly slow, and can be especially challenging in the case of amorphous silica, since long simulation times are necessary to obtain adequate equilibration. An alternative approach, initially adopted by Tangney and Scandolo,[11] is to use a self-consistent force matching (SCFM) procedure. Figure 1 shows a schematic of how SCFM works. In step 1, a fast classical molecular dynamics trajectory is simulated, giving an ensemble of initial configurations (step 2). Density functional theory (DFT) calculations are then performed on these configurations (step 3) to obtain the ab initio forces on the atoms. In step 4, these ab initio forces are then used to generate a new potential via standard force matching. This new potential is then used to generate a new ensemble of configurations, and the process is repeated until the force field has converged. SCFM thus simultaneously produces a classical force field and an ensemble of configurations that are consistent with the ab initio atomic forces. Furthermore, SCFM is an "embarrassingly" parallel method, in that there is no communication between the individual simulations during the most computationally intensive parts of the process (the MD and DFT calculations).

**Molecular Dynamics Simulations.** All classical molecular dynamics simulations were performed using the DL_POLY simulation package Ver. 2.17,[33] which was modified to

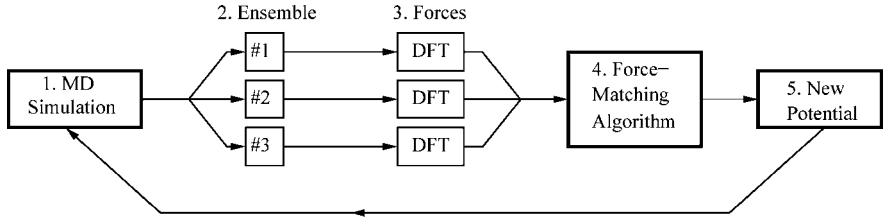



***Figure 1.*** Self-consistent force matching schematic.

incorporate the Wilson and Madden polarizability model[32] used in the TS force field.[11] Dipoles were approximated as small rigid rods (length of 0.02 Å) with "massless" atoms on the ends of the rods. In practice, these atoms were given a small mass so that the integration routines did not need to be modified (and the mass of the central atom was adjusted so that the total mass was the correct mass of the ion); however, the orientation and magnitude of the dipoles were solved according to eq 14, which is not affected by the mass of the "massless" atoms. The DL_POLY program was modified so that at every step, eq 14 was solved self-consistently to give the dipoles on the central ions, subsequently giving the charges on the "massless" atoms and the orientation of the rigid rod. All silica simulations discussed here were performed on systems of 24 $SiO_2$ units in a cubic box of size 10.286 Å (giving the experimental density, 2.20 g/cm$^3$, of room temperature amorphous silica silica). The particle−mesh Ewald method with a tolerance of $10^{-8}$ was used to calculate the long-range electrostatics. The short-range cutoff was set to 5.0 Å and a 0.5 fs time step was used.

In the TS force field, the Si−Si force is attractive and diverges at short distances. As suggested by Tangney and Scandolo,[11] a hard restraining potential (of the form $Cr^{-12}$, where $C = 1.55$ eV·Å$^{12}$) was summed into the Si−Si force to avoid these divergent attractions. The restraining potential was chosen to have a negligible effect on the Si−Si pairwise potential (i.e., the sum of the short-range pairwise and Coulomb components) in all regions except near the peak in the potential that occurs around 1.6 Å.

For the SCFM calculations, 112−224 configurations were used and the number of gridpoints $N_d$ was set to 200−400. These initial configurations were generated by starting with a 72-atom α-quartz configuration and heating four systems, using different random seeds, to 5000 K for 1 ns using the BKS force field.[6] Hundreds of 5000 K configurations were then annealed to 3000 K over 100 ps. This resulted in configurations free from the original crystalline order. Between SCFM iterations, the systems were annealed by heating up to 5000 K for 10 ps, annealing to 3000 K over 25 ps, and then equilibrating at the final temperature for another 5 ps. This annealing procedure helped to generate new configurations during subsequent SCFM iterations. No force information is obtained in regions that are not sampled by the original set of configurations, but it is possible that the system may sample these regions on a subsequent SCFM iteration. Therefore, hard restraining potentials needed to be added to the system at short distances to prevent the systems from collapsing into highly unlikely configurations. However, the restraining potentials may prevent the system from sampling regions just beyond those sampled in a previous iteration and therefore bias the sampling. In order to reduce this bias, a short (∼0.2 Å) "soft" region, where $r_{ij} \cdot f_{\alpha\beta}(r_{ij})$ was set to be constant, was added so that the system could reasonably sample these regions in the next iteration.

**DFT Calculations.** The DFT calculations were performed using the CPMD program.[34] The Perdew, Burke, and Ernzerhof functional with the generalized gradient approximation[35] was used in combination with the appropriate plane wave pseudopentials for Si and O supplied with CPMD. The energy cutoff for the plane wave basis was set to 130 Ry. Configurations for the CPMD calculations were generated from classical molecular dynamics simulations, and DFT forces were calculated by optimizing the wave function for these coordinates. For some of the 3000 K configurations sampled (∼30%), the wave function optimization failed to converge. This is likely due to the fact that some of the configurations sampled rare nuclear configurations that require more basis functions to converge. The failure rate could be reduced to about 20% by increasing the energy cutoff to 180 Ry. However, this incurred a much larger computational cost but did not significantly affect the resulting force field. Therefore, a 130 Ry cutoff was used for the calculations presented in this article.

## III. Results and Discussion

This section describes three different force fields that were constructed using SCFM. The first was a purely short-range pairwise force field that was parametrized using forces obtained from the TS force field, rather than DFT forces. The second was a purely short-range pairwise force field that was parametrized using DFT forces. The third was a polarizable force field similar to the TS force field, but where the short-range pairwise forces and the screening function have been reparametrized from DFT forces using linear tabularized forces. This section ends with a description of the amorphous silica structure that results following annealing the polarizable force fields from 3000 to 300 K.

**Force Matching TS onto a Purely Pairwise Force Field.** One of the main goals of this work was to investigate how well a short-range pairwise (PW) force field of the form of eqs 2 and 3 could reproduce the structure of bulk amorphous silica resulting from a silica model with explicit many-body effects. To this end, SCFM was used to construct a PW force field, but instead of using DFT forces, the PW force field was parametrized using forces obtained from the many-body TS force field. This new force-matched force field will be referred to using the shorthand notation TS → PW.

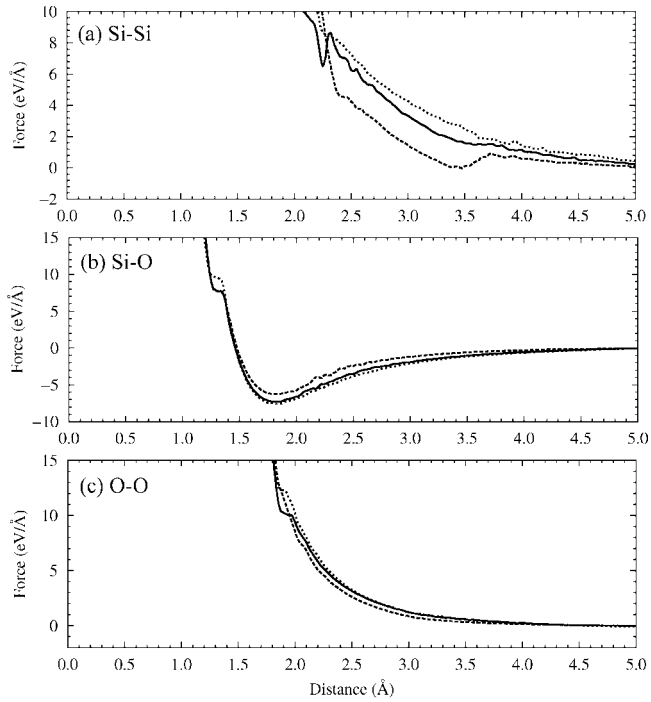



***Figure 2.*** Shows the (a) Si−Si, (b) Si−O, and (c) O−O pairwise force at the end of the first (long dashes), second (short dashes), and final (solid) SCFM iterations for the TS → PW force field. The "soft" regions of the hard restraining potential for the Si−O and O−O forces can be seen at ∼1.3 and ∼1.9Å, respectively.

Starting configurations for the TS → PW SCFM parametrization were sampled from 3000 K simulations of silica using the BKS force field.[6] The force-matched force field was considered to be converged when the difference in the forces between iterations was no greater than the noise in the force. Ten SCFM iterations were required to achieve convergence in this case.

The process of force field convergence serves to highlight some interesting aspects of the force matching method. Figure 2 shows the pairwise forces for the first two and last SCFM iteration. While the Si−O and O−O pairwise forces did not vary much over SCFM iterations, there were large changes in the Si−Si force. These changes can be understood by considering the changes in the structure, specifically the Si−O−Si angle, as the force field converges. The initial configurations were sampled from BKS simulations, where the Si−O−Si angle has a mean of 150°. The TS force field, from which this force field is being parametrized, has a mean Si−O−Si angle of 130°. The final converged angle is in between these two values at 140° (see Figure 3). So at the beginning of the SCFM procedure, the Si−O−Si angle is relatively flat, and at the end it is more bent. This angle consequently affects the magnitude of the induced dipole on the intervening oxygen (as calculated using the TS force field). In the initial flat configurations, the average magnitude of the dipole is 1.09(±0.01) D, whereas in the final more bent configurations, the average magnitude of the dipole is larger and has a value of 1.39(±0.01) D (errors reported at the 95% confidence interval).

Intuitively, these results appear to contradict the force data presented in Figure 2. If one considers a Si−O−Si system in isolation, then a large dipole on the oxygen should stabilize the Si−Si interaction and make the effective Si−Si force more attractive. Figure 2 shows just the opposite: in the flatter configurations where the dipole is weaker, the force is more attractive; in the more bent configurations, the dipole is stronger but the force is more repulsive. In fact, this is precisely the behavior required for the system to converge to the more bent configurations of the TS force field starting from the flatter BKS configurations. The configurations start flat and force matching gives a Si−Si force that is attractive, allowing the Si atoms to approach each other on the subsequent molecular dynamics simulations. After the first iteration, the system "overshoots" the optimal angle, giving configurations that are too bent. This resulted in a Si−Si force that was more repulsive, pushing the Si atoms away. The process repeated until convergence was obtained. But it still seems somewhat counterintuitive that the more bent configurations would give repulsive forces while the straight configurations give attractive forces. This result demonstrates how force matching incorporates many-body correlations into the effective pairwise force. The Si−Si atoms are more attractive in the straight configurations, not because of something to do with the dipole on the intervening O, but because the rest of the system pushes the Si atoms together toward an angle distribution more representative of the thermodynamic state.

Figure 4 compares the radial and angular distributions resulting from the TS force field with those from the force-matched force field, TS → PW. In many respects, the purely short-range PW force field actually does remarkably well at reproducing the structure of the TS system. The Si−O and O−O radial distribution functions, as well as many of the angular distributions (see Figure 3), are in very close agreement. The silica tetrahedra that result from the TS force field are accurately reproduced by the PW force field. The largest discrepancies are observed in the Si−Si radial and the Si−O−Si angular distributions. In the PW force field, the Si atoms are about 0.1 Å farther away from each other and the Si−O−Si is about 10° larger than that observed in TS. Furthermore, the small peak occurring between 80° and 100° in the TS Si−O−Si distribution is completely absent in the PW system. This peak is a sign of the existence of two-membered silica rings.[24] As discussed above, there is no prior experimental or simulation work suggesting the formation of two-membered rings at room temperature. However, the Si−O−Si angle distributions obtained from ab initio molecular dynamics simulations[11] at 3000 K do indicate the presence of rings. The TS force field thus appears to reproduce the structure of silica quite well at 3000 K. For the TS configuration, the magnitude of the dipole on oxygens belonging to such rings is distinctly different from the rest of the atoms (1.93(±0.04) D in the two-membered rings versus 1.46(±0.01) D for all the other oxygens). The discrepancies between the PW and TS force fields are a result of the fact that the PW force field is incapable of reproducing the forces needed to describe such a large difference in local environments.

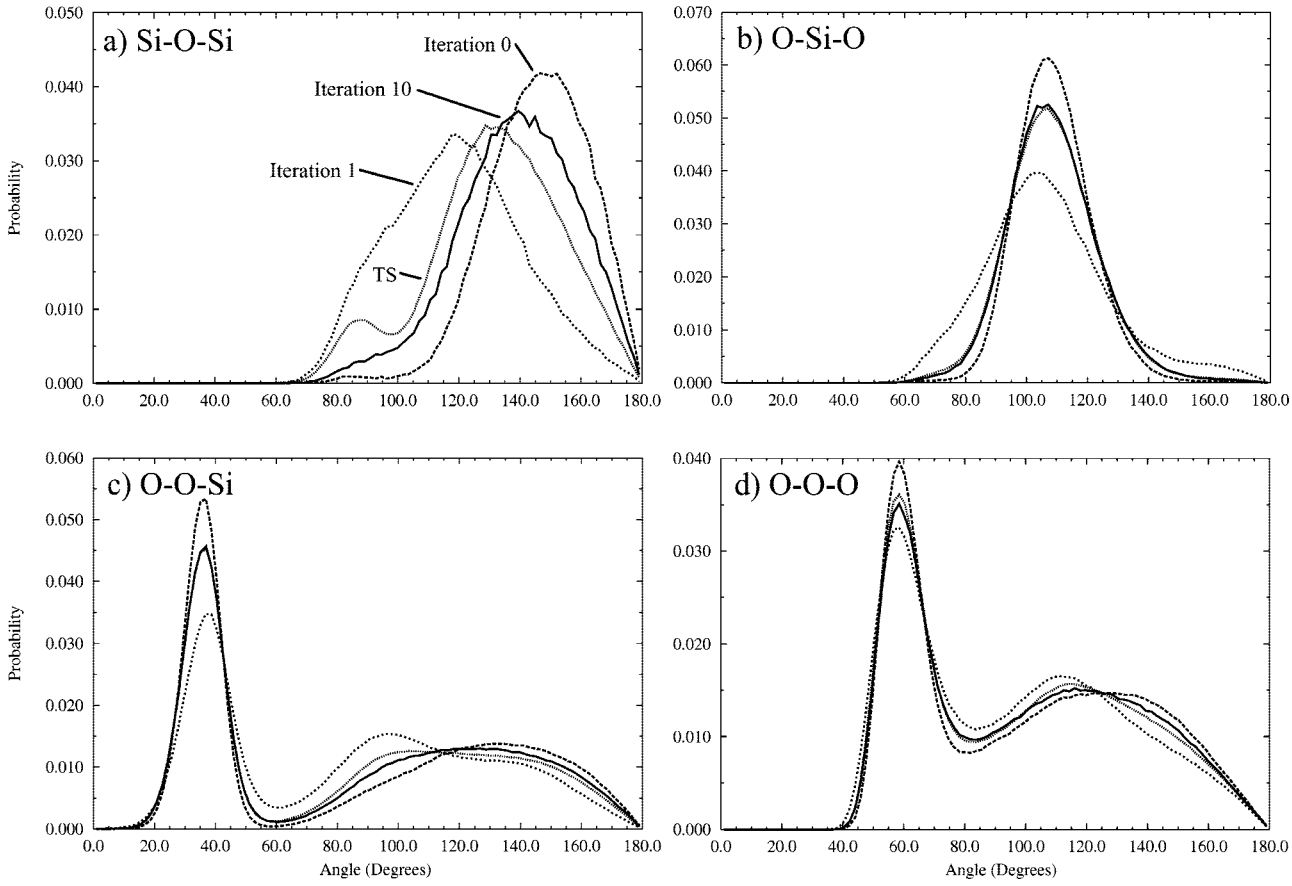



***Figure 3.*** Several angle distributions for the initial set of configurations sampled from BKS simulations (iteration 0, long dashes), the configurations resulting from the first force field (iteration 1, medium length dashes), the final configurations from the converged force field (iteration 10, solid), and configurations sampled from TS simulations.

**Force Matching DFT Forces onto a Pairwise Force Field.** Two previous studies have attempted to use force matching to construct a PW potential for silica from DFT forces.[11,36] In these studies, the PW potential resembled the functional form of the BKS potential,[6] and nonlinear force matching was used to find the parameters. However, in both cases, the resultant force fields failed to give reasonable silica structures. Using the methods described in this article, we were able to obtain a PW potential for silica parametrized from DFT forces, DFT → PW, that does give reasonable silica structures (see Figure 7, which compares this PW potential with a many-body potential discussed below). The failure of the previous models is likely a consequence of either the difficulty of minimizing the residual of eq 1 with nonlinear methods or the inability of the rigid functional form of the BKS potential to properly fit the DFT forces.

**Force Matching DFT Forces onto a Many-Body Force Field.** The TS force field was originally parametrized using nonlinear force matching, where both the short-range pairwise potential and screening functions were represented as fixed functional forms. To determine if the nonlinear fitting or the chosen functional forms may have given rise to any significant artifacts (as was observed in the PW case discussed above), we reparametrized the TS force field using tabularized forces and a screening function as given in eqs 3 and 19. This new force field will be referred to as the reparametrized Tangney−Scandolo (RTS) force field. Hard restraining potentials with "soft" regions were again used during SCFM. At very short distances that were not sampled, the screening function was set to −14.4 eV Å/$e^2$, which corresponds to complete screening, and linearly interpolated over 0.5 Å to its value at the shortest distances that were sampled.

As can be seen from Figure 5, the pairwise forces and screening function for the RTS force field closely match those of the original TS in the regions that were sampled in the SCFM simulations. The largest discrepancies are in the Si−Si pairwise force. The RTS Si−Si force is distinctly more attractive over nearly all distances sampled. This highlights one of the problems associated with picking a strict functional form and using a nonlinear force matching algorithm: the Si−Si pairwise force function used in TS gives a very poor fit to the forces observed using the tabularized potential. Nevertheless, the rather noticeable differences in the Si−Si force only leads to subtle differences in the resultant structure, as shown in Figure 6. The first peaks in the RTS radial distribution functions are shifted out by less than 0.1 Å, and the major peak in the Si−O−Si distribution is about 5° smaller. The RTS force field also shows a similar number of two-membered rings and is in good agreement with previous ab initio molecular dynamics simulations.[11]

Figures 7 and 8 compare the RTS force field with the DFT → PW force field discussed above, which were both parametrized from DFT calculations using the same procedure. Most of the structural parameters are in extremely close agreement. For perfect tetrahedra, the O−O−O angle is 60°,

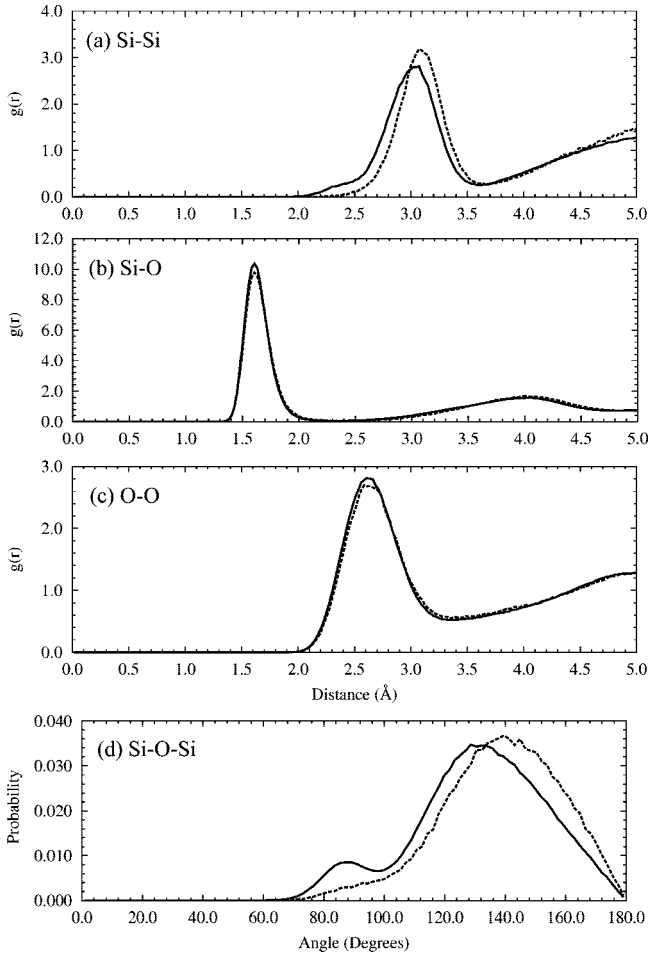



***Figure 4.*** Shows (a−c) radial and (d) angular distributions of the silica atoms obtained from the TS force field (solid lines) and the TS → PW force field (dashed lines).

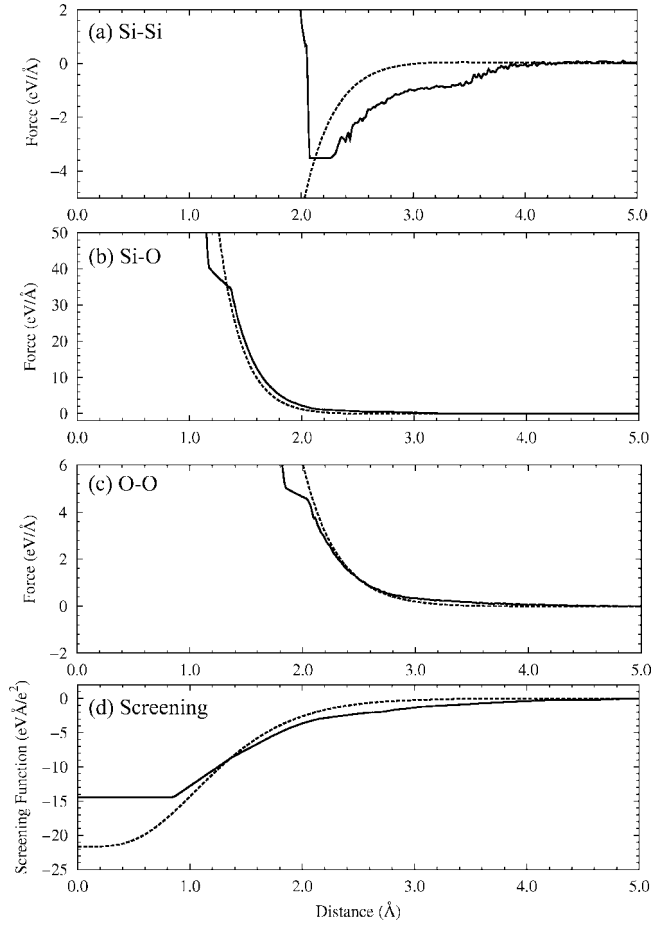



***Figure 5.*** Comparison of the (a−c) pairwise forces and (d) screening function between the reparameterized TS force field (solid lines) and the original (dashed lines).

the O−O−Si angle is 35°, and the O−Si−O angle is 109°.[12] Both the PW and many-body force fields give angles very close to these ideal values, indicating the presence of well-formed tetrahedra. However, similar to the differences observed between the TS and TS → PW force fields, the Si−Si radial, and Si−O−Si angular distributions differ significantly between the force fields. The PW force field gives larger Si−O−Si angles and does not show as strong of a tendency to form two-membered rings. Consequently, the average distance between Si atoms is larger with the PW force field. Again, these discrepancies have to do with the PW force field's inability to properly account for the large differences in forces that occur when two-membered rings form at 3000 K.

**Annealing TS and RTS to 300 K.** One of the original motivations for reparametrizing the TS force field concerned an anomaly we noticed when annealing TS to 300 K. Annealing was performed by taking configurations sampled at 3000 K and then setting the Nosé−Hoover[37] target temperature to 300 K and varying the thermostat relaxation time. The two-membered rings observed in the 3000 K simulations were again found in the annealed configurations (see Figure 9). Annealing times up to 500 ps were tested and gave results that were indistinguishable from much faster annealing times of 50 ps. While two-membered rings have been observed in ab initio molecular dynamics simulations of bulk silica at 3000 K[11] and are frequently observed as hydrolyzable defect sites in silica surface simulations,[8,25−28] we are not aware of any other bulk simulations or experiments that give two-membered rings at 300 K. We were able to remove these artifacts by adding a hard restraining potential to the Si−Si potential. However, the influence of the hard restraining potential had to be extended past the first peak in the Si−Si radial distribution function. This effectively replaced all the parametrized Si−Si interactions with an ad hoc potential, which we believe is an unacceptable method of removing the artifacts.

We therefore reparametrized the TS potential, with the expectation that the two-membered rings were a consequence of a poor fit of the TS potential to the ab initio forces at short distances and that a more flexible pairwise potential would remove the artifacts. Indeed, the original TS Si−Si force is not a good fit to the forces obtain using a tabularized potential (see Figure 5). However, as seen in Figure 9, the two-membered rings were also observed in the RTS simulations, although perhaps to a slightly smaller extent. While peaks in the Si−Si radial and Si−O−Si angular distribution functions corresponding to the two-membered rings appear small, about 1% of the oxygens belong to two-membered rings. For these 72 atom simulations, almost one out of every four configurations contain a two-membered ring.

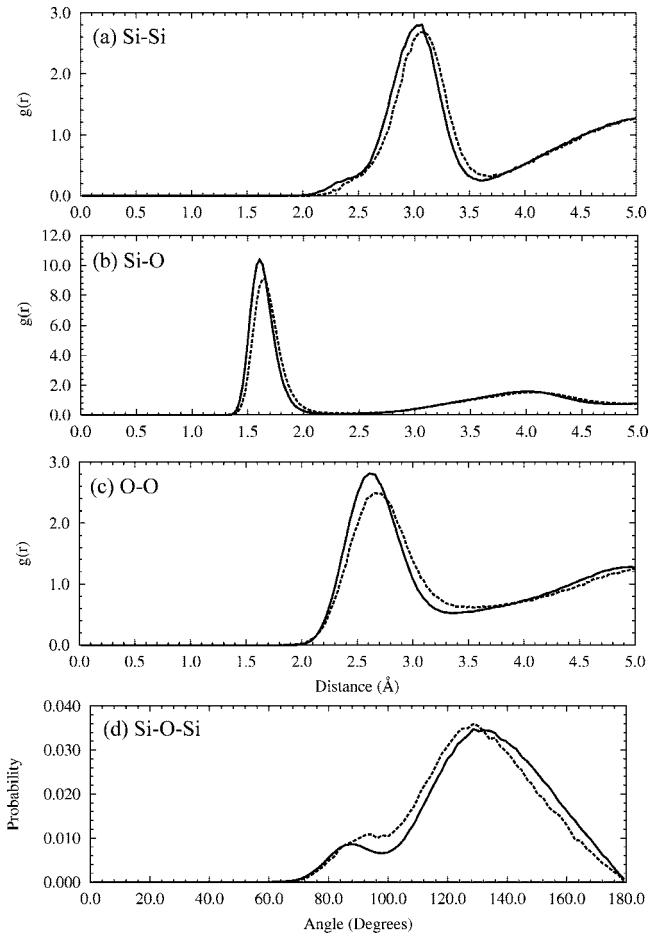



***Figure 6.*** Shows (a−c) radial and (d) angular distributions of the silica atoms obtained from the TS force field (solid lines) and the RTS force field (dashed lines).

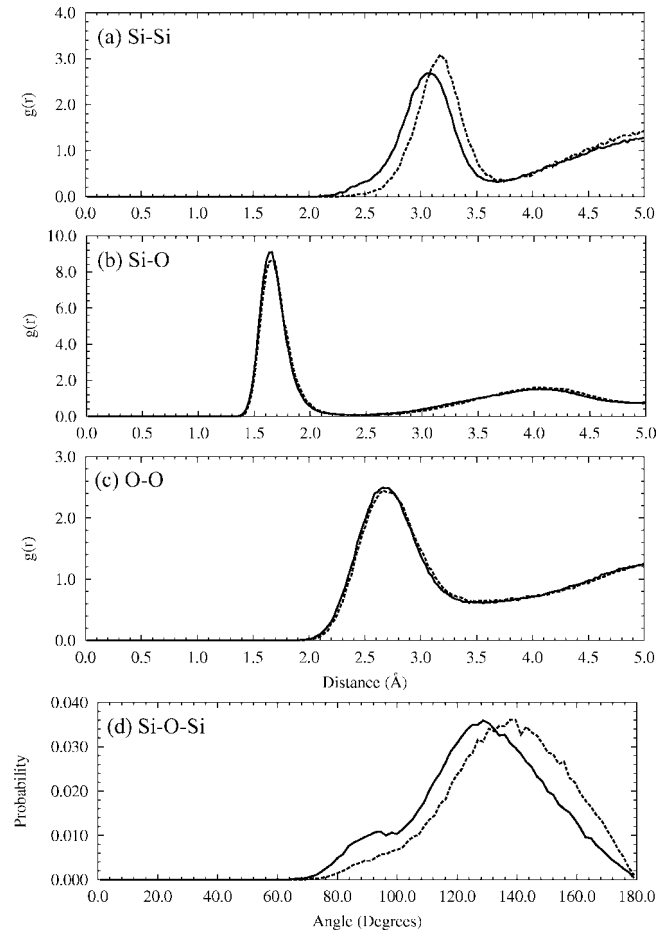



***Figure 7.*** Shows (a−c) radial and (d) angular distributions of the silica atoms obtained from the RTS force field (solid lines) and the DFT → PW force field (dashed lines).

Given the lack of any experimental data supporting the existence of two-membered rings in bulk silica, it seems likely that these are artifacts of the model. There are a few potential origins of the artifacts. One problem could be that the potential is essentially accurate, but that the annealing times are far too fast to allow for sufficient equilibration. Proper equilibration becomes particularly difficult when more degrees of freedom are added to the system (i.e., the dipoles). While we probed annealing times from 50 to 500 ps and did not observe significant differences in the formation of the rings, it is possible that computationally inaccessible annealing times could remove the two-membered rings. Another possibility is that the DFT calculations may not be giving completely accurate forces; at best, force matching can only give a force field that is as accurate as the underlying ab initio forces. Finally, it is possible that the polarizable ion model[32] may not be sufficient for describing silica under different thermodynamic conditions. For example, changes in the local density upon cooling could conceivably give rise to different effective charges on the Si and O atoms, an effect not captured by the present model.

## IV. Conclusions

This article explored how force matching can be used to construct classical force fields for silica. Three force fields were developed. The first force field was a purely short-range pairwise force that was parametrized from forces obtained from the TS potential. The pairwise potential

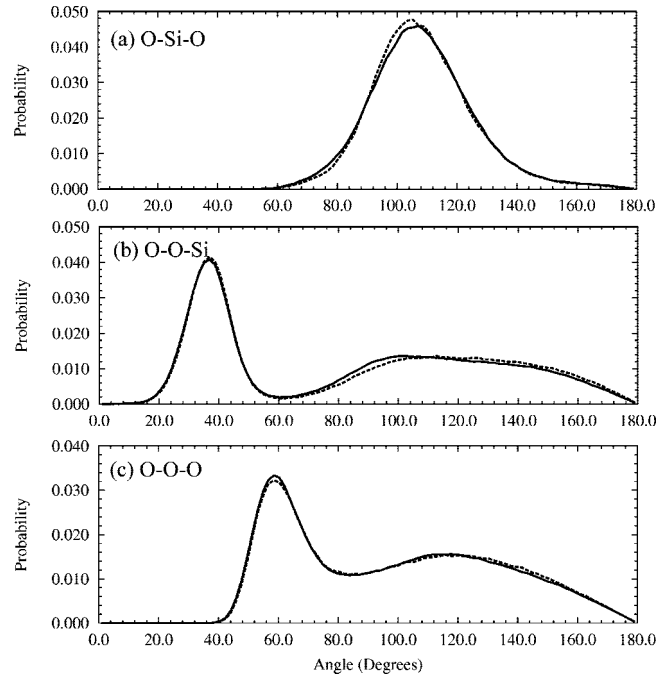



***Figure 8.*** Angular distributions for the RTS force field (solid lines) and the DFT → PW force field (dashed lines).

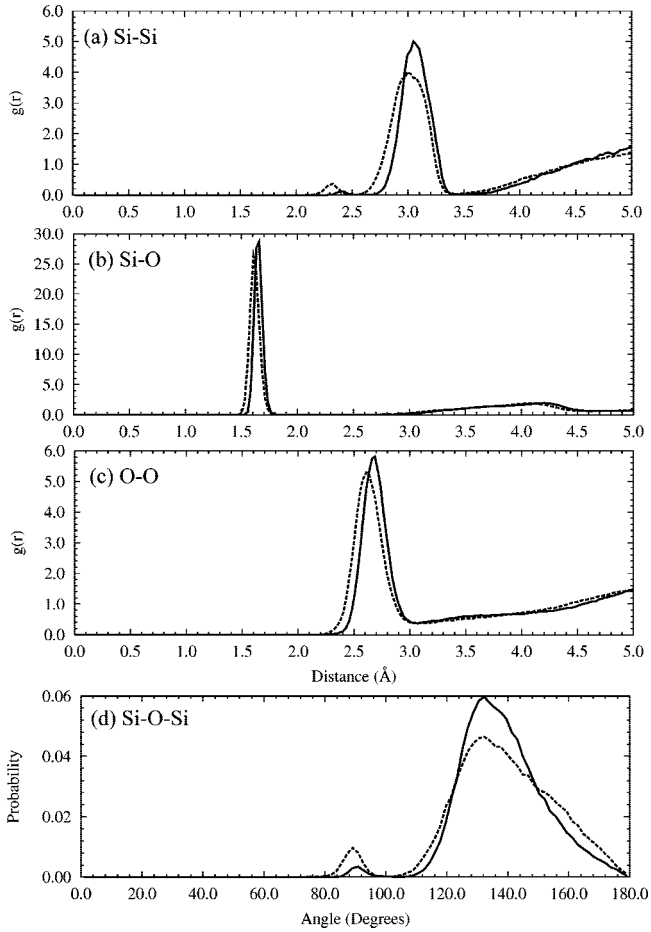



***Figure 9.*** Shows (a−c) radial and (d) angular distributions of the silica atoms obtained upon annealing to 300 K from the RTS force field (solid lines) and the TS force field (dashed lines).

reproduced the tetrahedral structure of the basic silica units and was in good agreement with several other structural parameters. However, the pairwise potential was unable to reproduce the correct Si−O−Si angle that connects the silica tetrahedra. This is because the Si−O−Si angle modulates the polarization of the oxygen and subsequently affects forces on nearby atoms. A pairwise model is incapable of reproducing this effect; but more importantly, these results demonstrate that the Si−O−Si angle is explicitly sensitive to many-body effects that cannot be "averaged out" using an effective pairwise form.

The second force field developed was a purely short-range pairwise force that was parametrized from forces obtained from DFT calculations. Previous attempts[11,36] at parametrizing a purely short-range pairwise force failed to give reasonable silica structures. This is likely due to the fact that this earlier work relied on strict functional forms for the potential and a nonlinear force matching algorithm. By employing a linear tabularized potential, we were able to develop a pairwise force field for silica that gives the correct tetrahedral structure. However, the Si−O−Si angle distribution is similar to that obtained using the first pairwise force field obtained from the TS forces. This suggests that the pairwise force field parametrized from DFT forces is also not properly reproducing the many-body effects involved in linking the silica tetrahedra.

The third force field developed was based on the TS force field, but where the short-range pairwise forces and screening function were replaced with linear tabularized functions (RTS). While the RTS Si−Si pairwise force is distinctly different from TS, only minor differences were observed in the resultant structure.

Lastly, our results also demonstrate some of the challenges involved in transferring a force field parametrized at one thermodynamic state to another. While the TS force field reproduces the structure of molten silica at 3000 K as determined from ab initio molecular dynamics calculations,[11] annealing the model to 300 K gives two-membered rings that have not been observed in other simulations or experiments. Failure of the model to anneal properly could indicate that the polarizable model needs to be augmented with more complex features, such as the ability to undergo charge transfer.

**Acknowledgment.** This research was supported by the National Science Foundation through a CRC grant (CHE 0628178) and an allocation of computer time on TeraGrid resources provided by NCSA and TACC.

## References

(1) Woodcock, L. V.; Angell, C. A.; Cheeseman, P. *J. Chem. Phys.* **1976**, *65*, 1565–1577.

(2) Sarnthein, J.; Pasquarello, A.; Car, R. *Phys. Rev. B* **1995**, *52*, 12690–12695.

(3) Ma, Y.; Foster, A. S.; Nieminen, R. M. *J. Chem. Phys.* **2005**, *122*, 144709.

(4) Karki, B. B.; Bhattarai, D.; Stixrude, L. *Phys. Rev. B* **2007**, *76*, 104205.

(5) Tsuneyuki, S.; Tsukada, M.; Aoki, H.; Matsui, Y. *Phys. Rev. Lett.* **1988**, *61*, 869–872.

(6) van Beest, B. W. H.; Kramer, G. J.; van Santen, R. A. *Phys. Rev. Lett.* **1990**, *64*, 1955–1958.

(7) Feuston, B. P.; Garofalini, S. H. *J. Chem. Phys.* **1988**, *89*, 5818–5824.

(8) Feuston, B. P.; Garofalini, S. H. *J. Chem. Phys.* **1989**, *91*, 564–570.

(9) Feuston, B. P.; Garofalini, S. H. *J. Appl. Phys.* **1990**, *68*, 4830–4836.

(10) Wilson, M.; Walsh, T. R. *J. Chem. Phys.* **2000**, *113*, 9180–9190.

(11) Tangney, P.; Scandolo, S. *J. Chem. Phys.* **2002**, *117*, 8898–8904.

(12) Vashishta, P.; Kalia, R. K.; Rino, J. P. *Phys. Rev. B* **1990**, *41*, 12197–12209.

(13) Hassanali, A. A.; Singer, S. J. *J. Phys. Chem. B* **2007**, *111*, 11181–11193.

(14) Bakaev, V. A.; Steele, W. A. *J. Chem. Phys.* **1999**, *111*, 9803–9812.

(15) Cruz-Chu, E. R.; Aksimentiev, A.; Schulten, K. *J. Phys. Chem. B* **2006**, *110*, 21497–21508.

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

(17) Izvekov, S.; Parrinello, M.; Burnham, C. J.; Voth, G. A. *J. Chem. Phys.* **2004**, *120*, 10896–10913.

(18) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 6573–6586.

(19) Youngs, T. G. A.; Del Pópolo, M. G.; Kohanoff, J. *J. Phys. Chem. B* **2006**, *110*, 5697–5707.

(20) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 4116–4127.

(21) Herzbach, D.; Binder, K.; Müser, M. H. *J. Chem. Phys.* **2005**, *123*, 124711.

(22) Anderson, E.; Bai, Z.; Bischof, C.; Blackford, S.; Demmel, J.; Dongarra, J.; Du Croz, J.; Greenbaum, A.; Hammarling, S.; McKenney, A.; Sorensen, D. *LAPACK Users' Guide*, 3rd ed.; Society for Industrial and Applied Mathematics: Philadelphia, PA, 1999.

(23) Iuchi, S.; Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2007**, *126*, 124505.

(24) Rino, J. P.; Ebbsjö, I. *Phys. Rev. B* **1993**, *47*, 3053–3062.

(25) Levine, S. M.; Garofalini, S. H. *J. Chem. Phys.* **1987**, *86*, 2997–3002.

(26) Garofalini, S. H. *J. Non-Cryst. Solids* **1990**, *120*, 1–12.

(27) Walsh, T. R.; Wilson, M.; Sutton, A. P. *J. Chem. Phys.* **2000**, *113*, 9191–9201.

(28) Du, J.; Cormack, A. N. *J. Am. Ceram. Soc.* **2005**, *88*, 2532–2539.

(29) Lekner, J. *Phys. Rev.* **1967**, *158*, 130–137.

(30) Stillinger, F. H. *J. Chem. Phys.* **1979**, *71*, 1647–1651.

(31) Martyna, G. J.; Berne, B. J. *J. Chem. Phys.* **1989**, *90*, 3744–3755.

(32) Wilson, M.; Madden, P. A. *J. Phys.: Condens. Mat.* **1993**, *5*, 2687–2706.

(33) Smith, W.; Forester, T. *J. Mol. Graph.* **1996**, *14*, 136–141.

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

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

(36) Carré, A.; Horbach, J.; Ispas, S.; Kob, W. *Europhys. Lett.* **2008**, *82*, 17001.

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

CT800244Q

JCTC Journal of Chemical Theory and Computation

# Role of Electrostatic Interactions on Engineering Reaction Barriers: The Case of CO Dissociation on Supported Cobalt Particles

Wai-Leung Yim and Thorsten Klüner*

*Institut für Reine and Angewandte Chemie, Theoretische Chemie, Carl von Ossietzky Universität Oldenburg, Carl-von-Ossietzky-Strasse 9-11, 26129 Oldenburg, Germany*

Received June 24, 2008

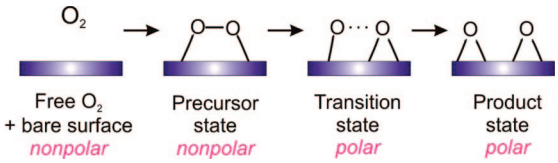
**Abstract:** We demonstrate a systematic optimization of the activation barrier of CO dissociation on cobalt surfaces on the basis of a chemical bonding picture of the corresponding transition structures. In particular, Co clusters adsorbed on MgO(100), graphene, and carbon nanotubes have been investigated. We discovered that the C−O moiety has a polar covalent character at the transition state which is feasibly stabilized by electrostatic interactions. This can be realized by replacing the $\beta$-Co atom with a less electronegative transition metal atom. The effect of 13 different substituting elements on CO dissociation has been investigated.

## Introduction

Modern experimental and theoretical methods have been used for understanding elementary industrial processes, such as CO oxidation,[1−3] CO activation,[4,5] oxygen reduction reaction (ORR),[6−8] and olefin metathesis.[9] In particular, the investigations can be consolidated by multiphysics approaches, e.g. using surface science techniques and first-principles calculations.[10−12] The knowledge thus obtained has impacted on the enhancement of power efficiency and the development of renewable energy resources.

Using modern apparatus, the screening of potential catalysts can be achieved without a detailed chemical understanding. For instance, a parallel screening of bimetallic catalysts for ORR was carried out using scanning electrochemical microscopy (SECM).[13] Despite this, the automatic screening process has certain limitations on the systems of great complexities, and detailed mechanistic studies on those cases would be desired. One of the examples is the CO-poisoning problem in fuel cells which is due to an inevitable contamination of hydrogen gas by carbon monoxide.[14] Another example is a debate about the initial activation mechanism in the Fischer−Tropsch (FT) process which converts CO and $H_2$ into liquid hydrocarbons using cobalt-containing catalysts: one proposed mechanism involved formation of C, O, and H adatoms, another one involved hydrogen-assisted C−O bond cleavage.[15]

***Scheme 1.*** Electronic Interaction between $O_2$ and a Transition Metal Surface during the $O_2$ Dissociation Process




First-principles calculations were proven to be powerful to resolve the influencing factors in the complicated processes. Based on the computed energetics results, the relative importance of reaction channels can be estimated, and the underlying mechanism can be rationalized by chemical bonding analyses. Recently, we have discovered that the short-range electrostatic effects were significant in tailoring the $O_2$ dissociation process, which was crucial for both the ORR and CO oxidation.[8] This was realized by a recent implementation of Bader analysis on charge density grids generated by planewave DFT calculations.[16,17]

The stabilization mechanism is illustrated in Scheme 1. The $O_2$ molecule exhibits no dipole moment in gas phase. Surprisingly, the $O_2$-species at the transition structure showed a considerable polar character.[17] So, an embedded ion at the reactive site can stabilize the transition state, while the precursor state is less affected. As a result, the $O_2$ dissociation

* Corresponding author phone: +49-441-798-3681; fax: +49-441-798-3964; e-mail: thorsten.kluener@uni-oldenburg.de.

10.1021/ct800243y CCC: $40.75 
Published on Web 09/27/2008

barrier can be manipulated. Short-range electrostatic effects may also be important for the gas reformation on the polarizable transition metal oxide substrates, such as ZnO,[18] and further confirmation will be required.

In this study, we examine the electrostatic effects on the CO activation on both cobalt nanoparticles and flat cobalt surfaces. In the case of CO, the carbon atom would be more negatively charged in the course of the dissociation process. Therefore, the CO activation is affected by the electrostatic effects. Similar effects for various adsorbate/substrate systems have been reported in the literature.[19−32] We selected a series of dopants of various electronegativities and found that the larger electronegativity difference between the doping element and cobalt was the stronger the stabilizing effect turned out to be. Our study illustrates the promoter effect by doping a reducing atom, which paves the way to enhance the efficiency of CO activation. Nevertheless, catalytic CO-activation is mechanistically very complicated, and our results can only be regarded as model studies in particular given the fact that we try to identify quite simple, i.e. electrostatic mechanisms. Although reality might be much more complex, we hope that this work stimulates further computational studies. Open questions concern the observation that the Pd-promoter effect is currently found not to be due to electrostatic interactions and the promoting behavior of Pd is not known yet.[33]

## Computational Details

**Model.** For isolated clusters, a 15 × 15 × 15 Å$^3$ cubic supercell with Γ-point sampling has been used. For supported clusters, we considered graphene, (6,6) single-walled carbon nanotube (SWNT), and MgO(100) as supports. For the cluster/graphene system, we used a hexagonal supercell ($a$ = 12.2 Å and $c$ = 16.0 Å), with a 2 × 2 × 1 Monkhorst-Pack (MP) mesh for k-space integration. For the cluster/(6,6) SWNT system, we used an orthorhombic supercell (18.0 × 23.9 × 12.3 Å$^3$) with a 1 × 1 × 2 MP k-point mesh. For the cluster/MgO(100) system, we used three layers of MgO in a tetragonal supercell ($a$ = 12.0 Å and $c$ = 20.3 Å) with a 2 × 2 × 1 MP k-point mesh. For each system, systematic convergence checks were performed with respect to slab thickness, k-point sampling, and cutoff.

To study CO dissociation on pure and substituted Co(0001) surfaces, we used a (3 × 3) Co(0001) surface slab of three atomic layers in a hexagonal supercell ($a$ = b = 7.47 Å, $c$ = 8 Å). A 2 × 2 × 1 MP mesh was chosen for k-space integration. The optimized structure of Co(0001)-CO was taken as a starting configuration of a MD-simulation, in which the sample was heated to 300 K for 250 fs, and then a constant temperature MD trajectory was calculated for a total simulation time of 1 ps, using Nosé-Hoover thermostat and a time step of 0.5 fs.[34] Constrained MD simulations were carried out subsequently, with the C−O distance of 1.2, 1.5, 1.8, 2.1, and 2.4 Å. Each trajectory was propagated at 50 K for 250 fs. The structure at the end of the last trajectory was optimized, and the minimum energy path was located by the climbing-NEB scheme.

**Theory.** We used the Vienna Ab Initio Simulation Package (VASP)[35−38] to perform spin-polarized DFT calculations. The PBE exchange-correlation functional within the generalized gradient approximation (GGA) was chosen.[39] Pseudopotentials constructed by the PAW method were adopted.[40] The planewave and augmentation charge cutoffs were set to 400 and 645 eV, respectively. Geometry optimizations were carried out by the conjugate gradient scheme in VASP. The geometrical parameter space was explored by starting at different initial cluster orientations and adsorption sites. The convergence threshold was set to $10^{-4}$ eV for both electronic structure calculations and geometry optimizations. The climbing-Nudged Elastic Band scheme was employed to search for transition structures.[41,42] Methods of charge topology analyses are referred to refs 16 and 17.

## Results and Discussion

To illustrate the electrostatic effects on CO activation, we model the cobalt surfaces by using an icosahedral $Co_{13}$ particle because of its small size and thermal stability.[43,44] We calculated the CO dissociation process on a bare $Co_{13}$ particle starting from their chemisorbed forms. On the surface of the bare $Co_{13}$ particle, CO adsorbs on a 3-fold hollow site with a chemisorption energy of −2.22 eV. The CO dissociation is slightly exothermic by 0.21 eV with a barrier of 1.89 eV. The apparent barrier, relative to the isolated reactants, is negative which means that the activation energy of the dissociation is smaller than the chemisorption energy of CO on the surface ($|E_a| < |E_{chemi}|$).

We have explored the substrate effects on the CO activation, by studying the reaction on the $Co_{13}$ particle deposited on several substrates—graphene, (6,6) single-walled carbon nanotube (SWNT), and MgO(100). The $Co_{13}$ particles on graphene-like substrates, graphene sheet and SWNT, have been considered. The strength of the $Co_{13}$-graphene interaction is considerable, resulting in an interaction energy of −1.55 eV. The $Co_{13}$ particle exhibits an even stronger interaction with a (6,6)-SWNT surface (−2.35 eV). This fact is due to the enhanced reactivity with increasing surface curvature of the SWNT surface as compared to a flat graphene sheet.[45] In both cases, the net spin population was reduced by about 20% as compared with the free cobalt particle. Interestingly, despite strong interactions between the $Co_{13}$ particles and graphenic surfaces, the CO dissocation barriers are only slightly reduced by 0.06−0.14 eV (cf. Table 1). Furthermore, the minor role of graphenic substrates was confirmed by substrate deformation, where the mechanical force changes the barrier height by only 0.05 eV on $Co_{13}$/graphene, and the effect is negligible (±0.01 eV) on $Co_{13}$/(6,6)-SWNT.

We also examined CO dissociation on an ionic MgO(100) support for comparison. The $Co_{13}$ particle is strongly bound to the MgO(100) surface, with an adsorption energy of −3.78 eV. The Co atoms are mainly attached to the oxygen sites of MgO, which is in accord with a former study by Xu et al.[46] Due to the symmetry mismatch, the triangular facet of the $Co_{13}$ particle attached to MgO(100) is distorted and opened. On $Co_{13}$/MgO(100), the CO dissociation barrier is further reduced to 1.69 eV. When the MgO(100) surface is compressed or stretched, the interaction with the distorted Co facet is changed: the barrier height is increased by 0.36

***Table 1.*** Energetics of CO Dissociation on Supported $Co_{13}$

| | $\Delta E$ (eV)[a] | $E_a$ (eV)[b] | $\Delta E_a$ (eV)[c] |
|---|---|---|---|
| bare $Co_{13}$ | −0.21 | 1.95 | - |
| $Co_{13}$/graphene | −0.23 | 1.81 | - |
| $Co_{13}$/graphene[d] | −0.13 | 1.85 | 0.04 |
| $Co_{13}$/graphene[e] | −0.17 | 1.86 | 0.05 |
| $Co_{13}$/(6,6)SWNT | −0.02 | 1.89 | - |
| $Co_{13}$/(6,6)SWNT[f] | −0.04 | 1.88 | −0.01 |
| $Co_{13}$/(6,6)SWNT[g] | 0.03 | 1.9 | 0.01 |
| $Co_{13}$/MgO | −0.16 | 1.69 | - |
| $Co_{13}$/MgO[h] | 0.16 | 2.05 | 0.36 |
| $Co_{13}$/MgO[l] | 0.06 | 1.92 | 0.23 |

[a] Energy change: $\Delta E = E$[C(ad)···O(ad)] − $E$[CO(ad)]. [b] Activation barrier: $E_a = E$(TS) − $E$[CO(ad)]. [c] Change of activation barrier ($\Delta E_a$) is referenced to the $Co_{13}$ deposited on undistorted support. [d] 2% compression, $E$(graphene) is increased by 3.0 eV/cell. [e] 4% expansion, $E$(graphene) is increased by 2.3 eV/cell. [f] 2% compression, $E$(SWNT) is increased by 1.4 eV. [g] 4% expansion, $E$(SWNT) is increased by 5.1 eV. [h] 5% compression. [l] 5% expansion.

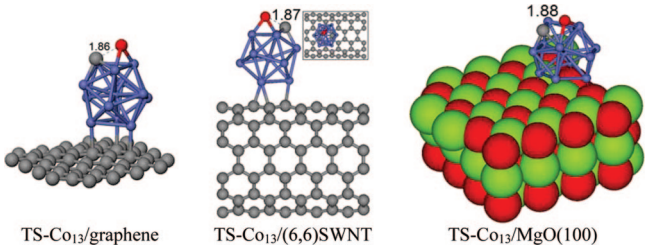



***Figure 1.*** Transition structures of CO dissociation on supported $Co_{13}$ particles.

eV (5% compression) and 0.23 eV (5% expansion), respectively. These changes are more pronounced than those on $Co_{13}$/graphenic surfaces.

To understand the trend of the CO dissociation barriers on different supported $Co_{13}$ particles, we illustrate the differences by showing the structural parameters and chemical bonding pictures of the transition structures. The precursor state contains a CO molecule adsorbing at the 3-fold hollow site of the $Co_{13}$ facet, while the C and O atoms as reaction products will adsorb at the hollow sites of the neighboring triangular facets. At the TS, the carbon atom has moved to the next-neighboring triangular facet and is coordinated to three Co atoms, while the O atom is adsorbed at the bridge site with the C−O distances ranging from 1.86 to 1.88 Å, respectively (cf. Figure 1). Interestingly, the structure of the transition state is virtually identical for bare and supported $Co_{13}$ particles, indicating that the influence of the support on the CO-dissociation barrier is not due to a simple structural deformation. The transition structures for CO dissociations on $Co_{13}$ particles supported by deformed substrates are also shown in Figure 2. All geometrical details of the adsorbed particles will be provided to the interested reader in the Supporting Information.

Instead, we found that the stabilization of the TSs was influenced by the surface polarity. We used the electron-localization-function (ELF)[47] and Bader analysis[16,17,48] to illustrate this phenomenon. The ELFs for the transition structures of CO dissociation on the supported $Co_{13}$ particles are shown in Figure 3. At the initial states, the hollow-site CO exhibits a localized orbital between the carbon atom and

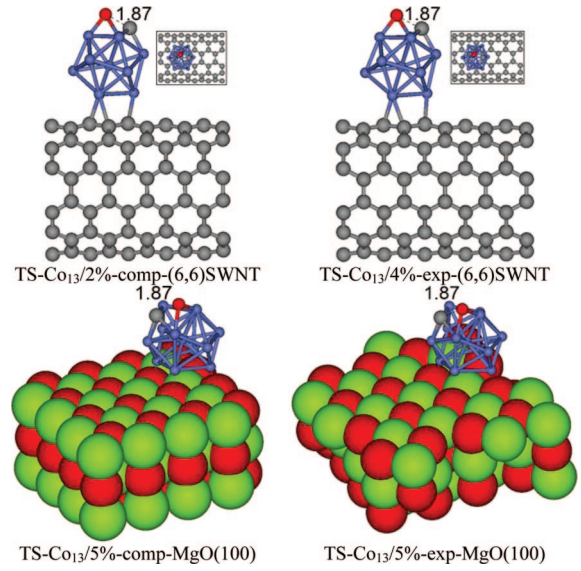



***Figure 2.*** Transition structures of CO dissociation on $Co_{13}$ on deformed supports.

the $Co_{13}$ surface, which refers to the 5$\sigma$ orbital of CO. The 4$\sigma$-lone pair orbital localized on the oxygen atom of CO and the C−O bond are visible as well.

At the TSs, the C−O bonding character greatly changes (cf. Figure 3). The electrons surrounding the C and O atoms exhibit a spherical shape, without a significant electron density between the C and O atoms. This indicates that the polar covalent character of the C−O bond increases at the TS, which is further confirmed by characterizing bond critical points (BCPs).[17,48] On the supported $Co_{13}$ particles, the Laplacians of the electron density at the BCPs of C−O are positive, revealing a dominating polar covalent character between them (cf. Table 2).[49] Moreover, the C and O atoms are bonded to the $Co_{13}$ surfaces by polar covalent bonds.

In addition, the surface polarity at the TS is demonstrated by the electric charges and atomic dipoles (cf. Table 2).[16] At the TS, both C and O atoms carry a negative charge, which agrees with the finding of ELF. The magnitude of these charges is not sensitive to the choice of support since the differences are beyond the accuracy of a population analysis. In contrast, a significant influence of the support on the atomic dipole moments in the TS-structure can be observed. Comparing the oxygen atomic dipole moment of the CO-molecule on both supports in fact suggests that the difference of atomic dipole moments on the two substrates might be due to the significantly different electric field of the supporting materials.

By chemical intuition, the transition state can be stabilized by embedding positive ions at the surface, while the chemisorbed CO molecule is less affected due to its charge neutrality. As a result, the CO dissociation barrier is adjustable. As mentioned above, the CO dissociation barrier is slightly lower when $Co_{13}$ is deposited on the MgO(100) surface, as compared to the (6,6)-SWNT. By comparing the difference of the Laplacian of the density at the BCP, we identified that the bond number 3 was the most important in stabilizing the transition state (cf. Table 2).

We hypothesize that the CO dissociation barrier can be modified chemically by substituting a less electronegative

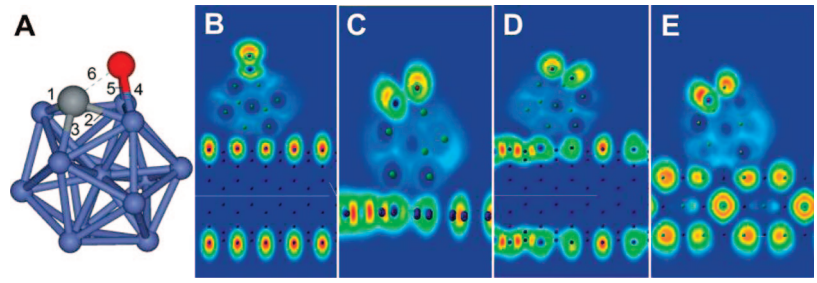



***Figure 3.*** Contour plots of the electron localization function: [a] connectivity of CO/bare-$Co_{13}$ at the TS; [b] ELF of CO/$Co_{13}$/(6,6)SWNT at the precursor state; [c] ELF of CO/$Co_{13}$/graphene at the TS; [d] ELF of CO/$Co_{13}$/(6,6)SWNT at the TS; and [e] ELF of CO/$Co_{13}$/MgO at the TS.

***Table 2.*** Properties of Bond Critical Points (BCPs) and Charge Distribution at Transition States[a]

| | | $Co_{13}$/(6,6)SWNT | | $Co_{13}$/MgO(100) | |
|---|---|---|---|---|---|
| bond | label | $\rho_b$ (au) | $\nabla^2\rho$ (au) | $\rho_b$ (au) | $\nabla^2\rho$ (au) |
| Co−C | 1 | 0.14 | 0.33 | 0.13 | 0.31 |
| Co−C | 2 | 0.15 | 0.36 | 0.16 | 0.36 |
| Co−C | 3 | 0.12 | 0.37 | 0.13 | 0.09 |
| Co−O | 4 | 0.09 | 0.58 | 0.09 | 0.54 |
| Co−O | 5 | 0.12 | 0.71 | 0.12 | 0.71 |
| C−O | 6 | 0.10 | 0.18 | 0.10 | 0.18 |

| | charge distribution | | | |
|---|---|---|---|---|
| | $Co_{13}$/(6,6)SWNT[b] | | $Co_{13}$/MgO(100)[c] | |
| label | charge (e) | $\mu_{atom}$ (D) | charge (e) | $\mu_{atom}$ (D) |
| Co atom | +0.17[d] | 1.11 | +0.06[d] | 1.07 |
| C atom | −0.57 | 1.73 | −0.56 | 1.60 |
| O atom | −0.84 | 0.25 | −0.82 | 0.47 |

[a] Atomic units are expressed as e bohr$^{-3}$ and e bohr$^{-5}$ for $\rho_b$ and Laplacian, respectively. Atom numbering refers to Figure 3a. [b] (6,6)-SWNT, chemical formula: $C_{120}$. [c] MgO(100), chemical formula: $Mg_{48}O_{48}$. [d] Average charge on Co atoms.

***Table 3.*** Energetics of CO Dissociation on Substituted M-$Co_{12}$ Particles

| particle | $\Delta E$ (eV)[a] | $E_a$ (eV)[b] | $E_{mono}$[c] ($1/4\pi\varepsilon_0$) | $E_{dipole}$[d] ($1/4\pi\varepsilon_0$) | charge$_M$ (e) |
|---|---|---|---|---|---|
| $LiCo_{12}$ | −0.01 | 1.67 | | | |
| $NaCo_{12}$ | −0.34 | 1.72 | | | |
| $YCo_{12}$ | −0.15 | 1.45 | −0.74 | 0.09 | +1.34 |
| $LuCo_{12}$ | −0.09 | 1.51 | −0.78 | 0.09 | +1.40 |
| $TiCo_{12}$ | 0.02 | 1.72 | −1.02 | 0.12 | +1.75 |
| $ZrCo_{12}$ | 0.09 | 1.66 | −1.59 | 0.13 | +2.42 |
| $HfCo_{12}$ | 0.14 | 1.61 | −2.08 | 0.10 | +2.80 |
| $MoCo_{12}$ | −0.55 | 1.84 | −0.49 | 0.07 | +0.82 |
| $WCo_{12}$ | −0.16 | 1.87 | −0.61 | 0.07 | +1.00 |
| $RuCo_{12}$ | −0.48 | 1.79 | −0.35 | 0.07 | +0.14 |
| $OsCo_{12}$ | −0.63 | 1.81 | −0.35 | 0.07 | +0.03 |
| $Co_{13}$ | −0.21 | 1.89 | −0.37 | 0.08 | |
| $PdCo_{12}$ | 0.07 | 2.14 | −0.49 | 0.09 | −0.41 |
| $PtCo_{12}$ | −0.02 | 1.77 | −0.58 | 0.07 | −0.69 |

[a] Energy change: $\Delta E = E$[C(ad)···O(ad)] − $E$[CO(ad)]. [b] Activation barrier: $E_a = E$(TS) − $E$[CO(ad)]. [c] Energy contribution due to the electric monopole term at the transition state. [d] Energy contribution due to the electric dipole term at the transition state.

metal atom at the $\beta$ position relative to the CO-attached cobalt facet. After substitution the geometry has been reoptimized using various initial structures which yield virtually identical results. We have selected transition metal elements of different groups and different periods of the periodic table, and also two main group elements (Li and Na), as shown in Table 3. From this, the effect of atomic/ionic size and electronegativity can be elucidated. It is noteworthy that we chose elements exhibiting a larger covalent radius than that of Co on purpose; otherwise, the core Co atom might move toward the surface and the skeleton of the model would be distorted.

Figure 4a reveals that the electronegativity of the substituting metal atoms plays an important role on the structure of the transition state and the modification of the corresponding energy barriers. When the difference in electronegativity relative to Co increases, the extent of stabilization will increase and the CO dissociation barrier height will decrease (cf. Figure 4a). The stabilization effect is also revealed by the Laplacian of the density: the more positive the Laplacian of the density at the BCP is, the smaller the dissociation barrier turns out to be.

We estimated the contributions of the electric monopoles and electric dipoles in the overall electronic interactions, by using the following formulas:[50]

Monopole interaction:

$$E_{el}(R^{AB}) = \frac{Q^A Q^B}{4\pi\varepsilon_0 R^{AB}}$$

Dipole interaction:

$$E_{el}(R^{AB}) = \frac{\mu^A \mu^B}{4\pi\varepsilon_0 (R^{AB})^3}(\cos\chi - 3\cos\alpha_A \cos\alpha_B)$$

$Q^A$ and $Q^B$ are estimated by the net charges of the atoms. The definitions of other symbols are shown in Scheme 2. To calculate the dipole−dipole interaction, the charge density difference is used to obtain the net dipole moments. The calculated results are also shown in Table 3 and Figure 4b. The electrostatic interaction is mainly determined by the monopole term. This can also be reflected by the charges of the substituents. Consistent with our prediction, the dissociation barrier decreases with increasing electrostatic interaction. It is of particular interest that Pd, which exhibits a promoter effect, does not activate CO dissociation by electrostatic interactions. Some exceptions are observed in Figure 4b, including substituting the elements Y, Lu, and Pd, which will be explained below.

Some of the data points in Figure 4 do not lie along the trend-curve, where the cases of Li, Na, Pd, and Pt turn out to be exceptions. These cases have very different transition structures as shown in the insets of Figure 4. Other transition

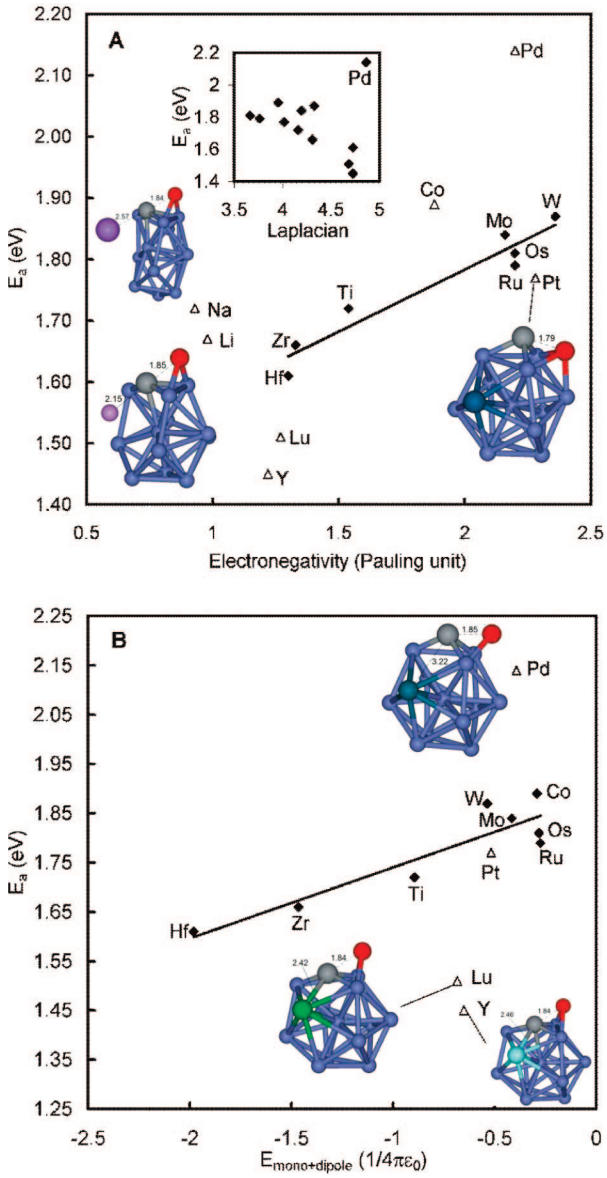
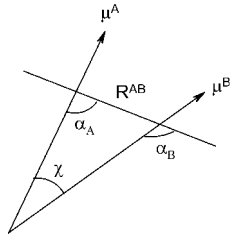



***Figure 4.*** [a] Relation of dissociation barrier and the electronegativity of the substituted metal atom and [b] dependence of dissociation barrier on the electrostatic interactions.

***Scheme 2.*** Parameters Required To Calculate the Energy Contribution Due to Monopole and Dipole Interactions




structures can also been found in Figure 5. In case of Li and Na, a charge transfer from the alkaline atom to the Co particle occurs, resulting in a significant distortion of the Co particle. For Pd and Pt, the C atoms are bonded to the hollow-site where the precursor-CO is located, instead of bonding to the $\beta$-Pd or the $\beta$-Pt atom, respectively. These structural differences explain the scattering of data points in Figure 4. Moreover, the barriers have been significantly reduced using Lu and Y. When these two low-valent transition metal elements are used, the triangular facet of M-$Co_{12}$ can be opened which is attributed to the weaker M-Co covalent interaction. As a result, the C atoms have higher coordination at the TS structures, leading to increased stability.

The abovementioned chemical tailoring is applied to the doped $MCo_{12}$ model. Under experimental conditions, however, the assumed structural skeleton may be distorted depending on the thermal stability of the nanoparticles. We have used *ab initio* molecular dynamics simulations to test the structural stability of the particles. In these simulations, the optimized geometry was heated to the desired temperature (300 or 500 K) for 1 ps, followed by a longer molecular dynamics simulation for 10 ps using a Nosé-Hoover thermostat.[34,51] The time interval was set to 2 fs.

Our prediction of CO dissociation on icosahedral particles is valid in the low-temperature regime (300 K) (cf. Figure 6); at higher temperature (500 K), the nanosized particles exhibit different structures (cf. Figure 7). Nevertheless, the unravelled stabilization mechanism of CO at the transition state should be applicable for other surfaces of different morphology as well.

We have also considered the flat Co surfaces of different chemical composition. In detail, we selected Hf-, Lu-, Y-, and Zr-substituted Co(0001) surfaces, as these dopants exhibited a larger promoter effect on the CO activation on substituted-$Co_{13}$ particles. Here, the surface slabs keep their structures at elevated temperature.

For comparison, we have calculated the reaction profile of the CO dissociation on the pure $(3 \times 3)$-Co(0001) surface (cf. Table 4 and Figure 8). In this case, the chemisorbed CO exhibits a dissociation barrier of 2.47 eV, forming C(ad) and O(ad), and the corresponding reaction energy is found to be 0.93 eV. Compared to the energetics of CO dissociation on the $Co_{13}$ particle, the barrier and reaction energy of CO dissociation on the flat Co(0001) surface are raised by 0.58 and 1.14 eV, respectively. The enhanced reactivity of the $Co_{13}$ particle is due to its higher curvature where the effect is even more pronounced in the product states [C(ad)···O(ad)]. Curvature effects can also be found on single-walled carbon nanotubes and low-coordinated transition metal surfaces.[52,53] In addition, the apparent barrier is found to be 0.71 eV, which is comparable to the previous theoretical value of 1.04 eV as reported on a $(2 \times 2)$-Co(0001) surface.[15,54]

In agreement with our prediction by the $Co_{13}$ model, the CO dissociation barrier is significantly reduced on the substituted-Co(0001) surfaces (cf. Table 4). With a transition metal substituent, the barrier is significantly reduced by 0.76−1.14 eV, resulting in a negative apparent barrier varying from −0.1 eV to −0.5 eV. We have also found a correlation between the activation energies obtained from the substituted cobalt surfaces of different curvatures (cf. Figure 9). Interestingly, a nice correlation is observed among pure cobalt, Hf- and Zr-substituted cobalt surfaces, which supports our intuition that the CO dissociation barrier can be engineered by introducing a less electronegative transition metal atom.

The explanation as mentioned above for the extraordinary low CO dissociation barriers for $YCo_{12}$ and $LuCo_{12}$ can also

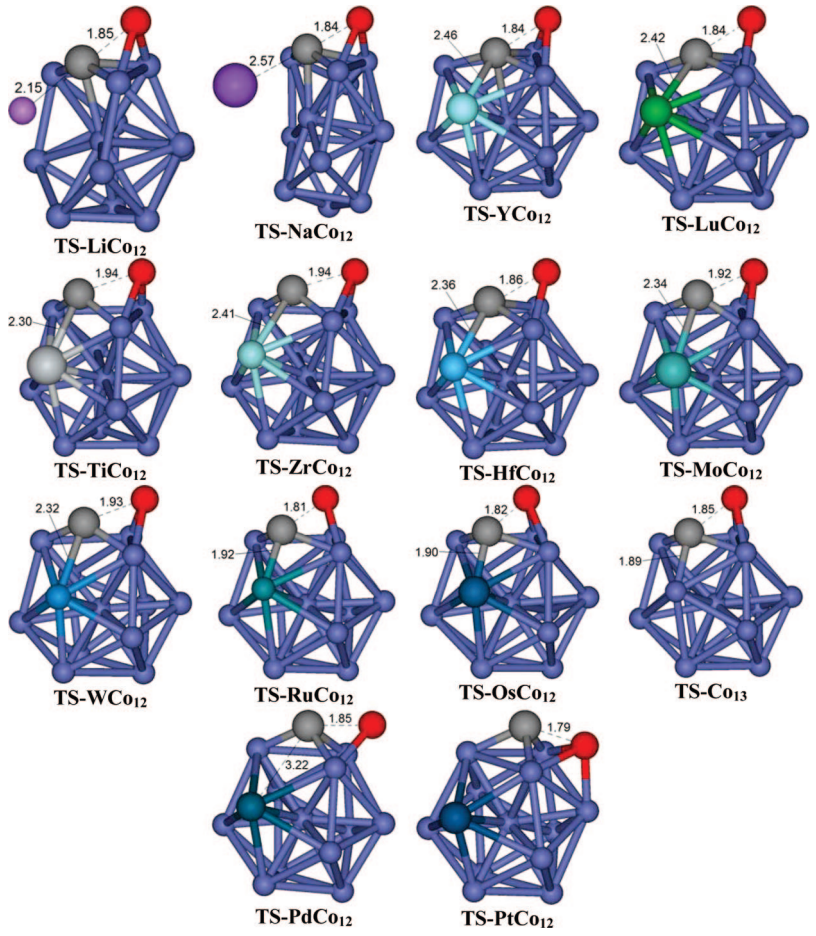



***Figure 5.*** Transition structures for CO dissociations on $MCo_{12}$ particles.

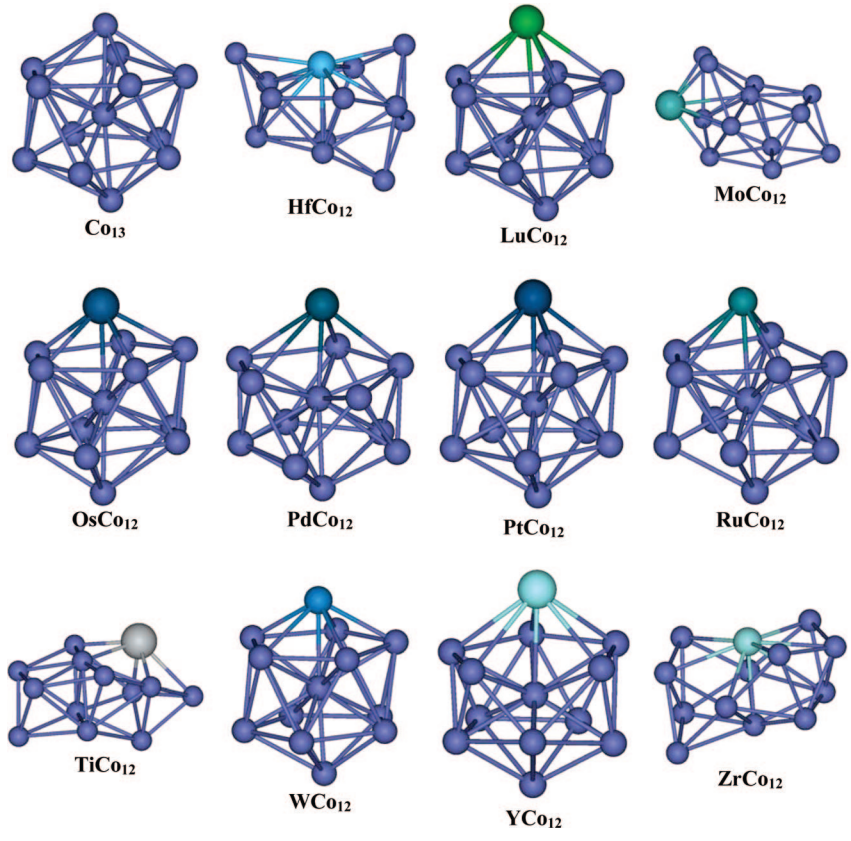



***Figure 6.*** Structures of $MCo_{12}$ particles after 300 K MD runs.

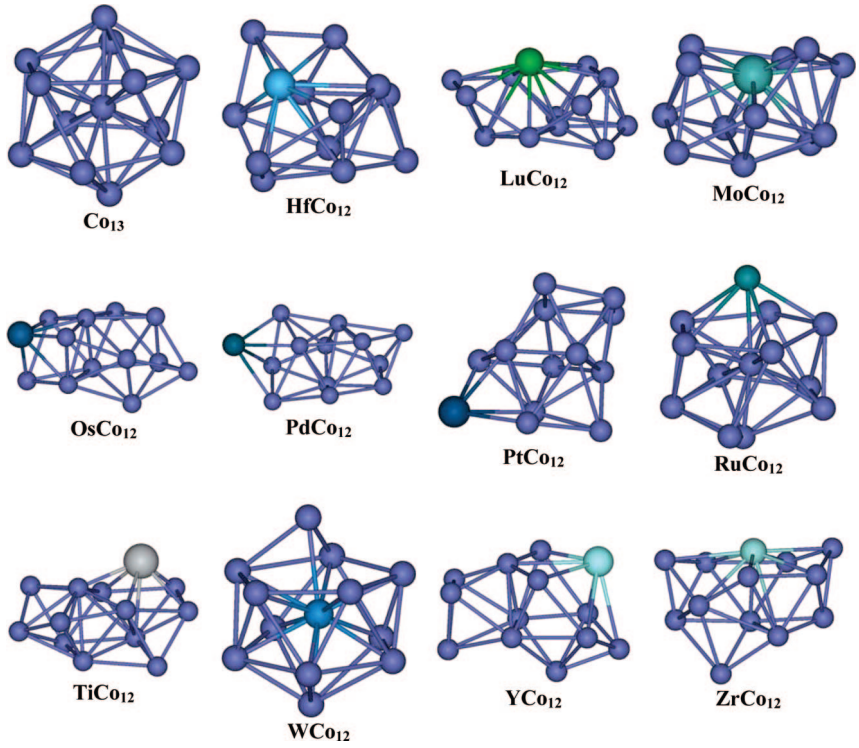



***Figure 7.*** Structures of $MCo_{12}$ particles after 500 K MD runs.

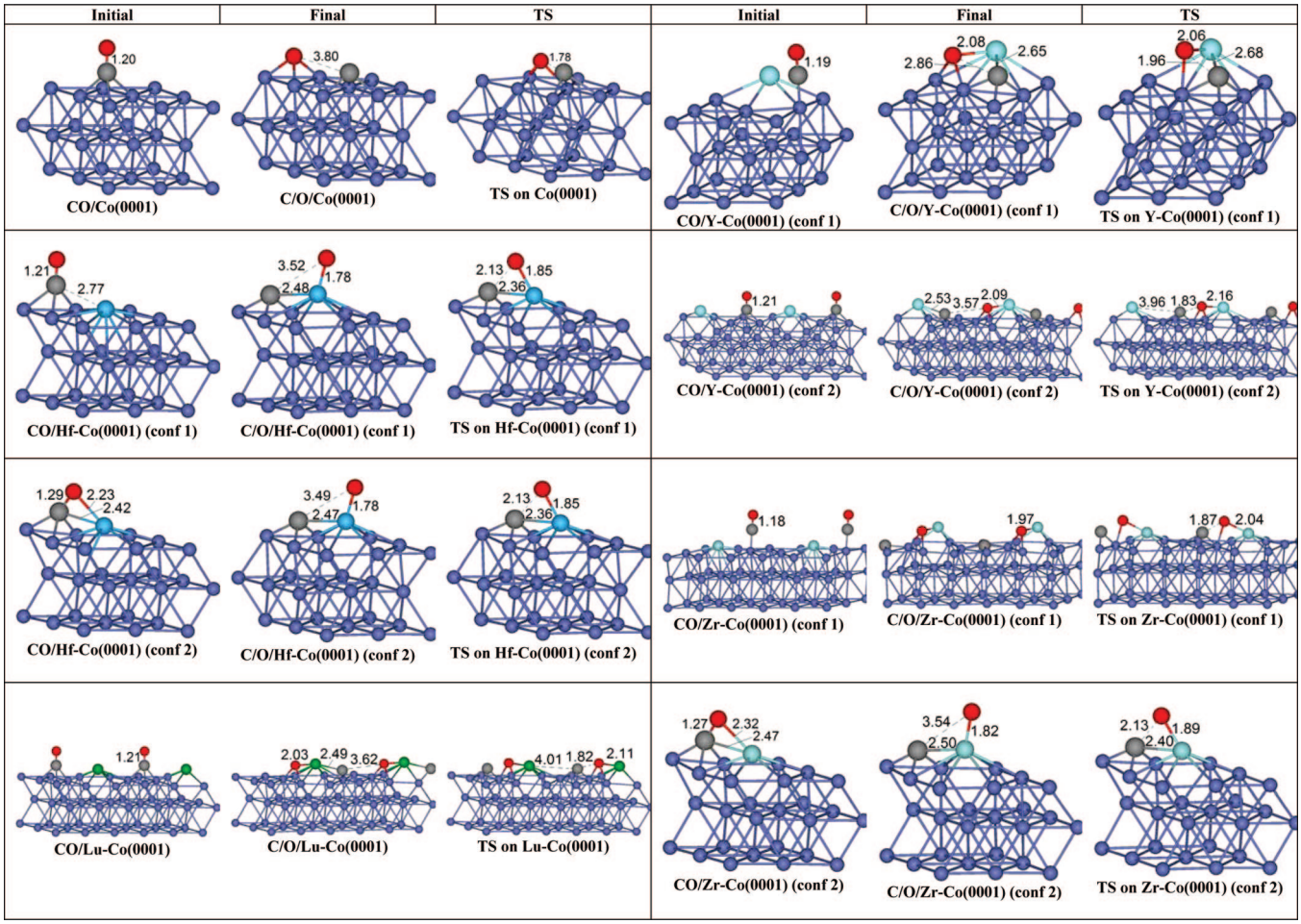



***Figure 8.*** Initial, final, and transition structures for CO dissociations on substituted Co(0001) surfaces are shown in left, middle, and right columns, respectively.

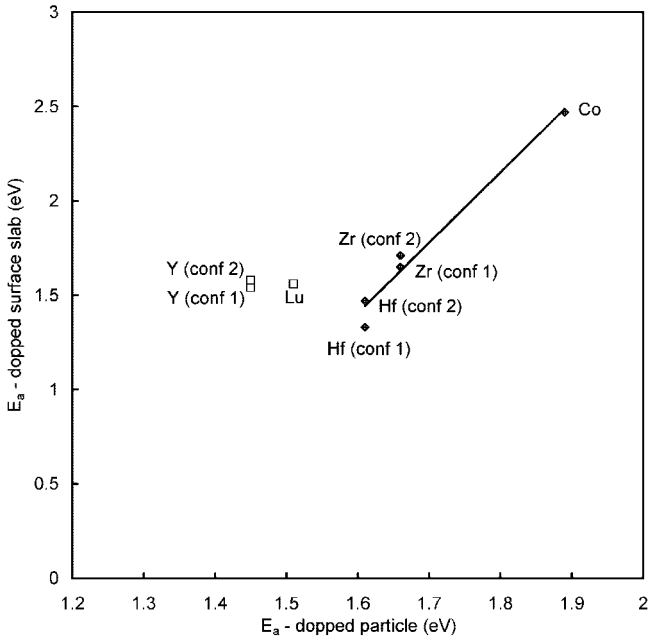



***Figure 9.*** $E_a$ (particle) vs $E_a$ (flat surface).

***Table 4.*** Energetics of CO Dissociation on Substituted Co(0001) Slabs

| surface | $E_{chemi}$ (eV)[a] | $\Delta E$ (eV)[b] | $E_a$ (eV)[c] |
|---|---|---|---|
| pure Co(0001) | −1.76 | 0.93 | 2.47 |
| Hf−Co(0001) (conf 1) | −1.83 | 0.84 | 1.33 |
| Hf−Co(0001) (conf 2) | −1.97 | 0.98 | 1.47 |
| Lu−Co(0001) | −1.82 | −0.27 | 1.56 |
| Y−Co(0001) (conf 1) | −1.76 | 0.17 | 1.54 |
| Y−Co(0001) (conf 2) | −1.83 | −0.27 | 1.58 |
| Zr−Co(0001) (conf 1) | −1.75 | 0.40 | 1.65 |
| Zr−Co(0001) (conf 2) | −2.01 | 1.17 | 1.71 |

[a] Chemisorption energy: $E_{chemi} = E$(CO-surface) − $E$(free CO) − $E$(bare surface). [b] Energy change: $\Delta E = E$[C(ad)···O(ad)] − $E$[CO(ad)]. [c] Activation barrier: $E_a = E$(TS) − $E$[CO(ad)].

be used for the scattered data of Y- and Lu-substituted cobalt surfaces. It should be mentioned that the minimum energy pathways lead to different oxygen coordination in the product states. For Y- and Lu-substituted Co(0001) surfaces, the oxygen adatoms occupy the hollow sites of the triangular Y−Co−Co and Lu−Co−Co facets, respectively, while the oxygen adatoms occupy the on-top sites of the dopants of other substituted-Co(0001) surfaces. Therefore, the CO dissociation reactions are found to be exothermic on the models Lu−Co(0001) and Y−Co(0001) (conf 2) but endothermic on other models.

## Conclusion

In conclusion, we performed large-scale DFT calculations to study the CO activation on $Co_{13}$ particles. We discovered the role of electrostatic interactions on stabilizing the transition states, while the initial structures were less affected due to the charge neutrality of the adsorbed CO-molecule. Our results revealed a controllable CO activation by replacing the $\beta$-Co atom. This strategy is also applicable on substituted-Co(0001) surfaces. In addition, as shown in this work, the Pd-dopant (one of the promoters in the FT process)[2] is unable to activate the CO dissociation via electrostatic effects. This motivates a further investigation on the CO activation mechanism on Pd-doped cobalt surfaces, and a detailed mechanistic study is currently being undergone in our group.

**Acknowledgment.** We thank financial support from Fonds der Chemischen Industrie, Alexander von Humboldt Foundation (W.L.Y.), Hanse Wissenschaftskolleg (W.L.Y.), and the EWE AG (T.K.). The simulations were performed on the national supercomputer NEC SX-8 at the High Performance Computing Center Stuttgart (HLRS) under the grant number WLYIM.

**Supporting Information Available:** All geometrical details of the adsorbed particles. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Bond, G. C.; Thompson, D. T. *Catal. Rev. - Sci. Eng.* **1999**, *41*, 319.

(2) Diehl, R. D.; McGrath, R. *Surf. Sci. Rep.* **1996**, *23*, 43.

(3) Alavi, A.; Hu, P. J.; Deutsch, T.; Silvestrelli, P. L.; Hutter, J. *Phys. Rev. Lett.* **1998**, *80*, 3650.

(4) Huo, C. F.; Ren, J.; Li, Y. W.; Wang, J. G.; Jiao, H. J. *J. Catal.* **2007**, *249*, 174.

(5) Andreoni, W.; Varma, C. M. *Phys. Rev. B* **1981**, *23*, 437.

(6) Xu, Y.; Mavrikakis, M. *J. Phys. Chem. B* **2003**, *107*, 9298.

(7) Xu, Y.; Ruban, A. V.; Mavrikakis, M. *J. Am. Chem. Soc.* **2004**, *126*, 4717.

(8) Yim, W.-L.; Klüner, T. *J. Catal.* **2008**, *254*, 349.

(9) Siaj, M.; Oudghiri-Hassani, H.; Maltais, C.; McBreen, P. H. *J. Phys. Chem. C* **2007**, *111*, 1725.

(10) Sorescu, D. C.; Rusu, C. N.; Yates, J. T. *J. Phys. Chem. B* **2000**, *104*, 4408.

(11) Carlisle, C. I.; King, D. A.; Bocquet, M. L.; Cerda, J.; Sautet, P. *Phys. Rev. Lett.* **2000**, *84*, 3899.

(12) Johnson, K.; Ge, Q.; Titmuss, S.; King, D. A. *J. Chem. Phys.* **2000**, *112*, 10460.

(13) Fernandez, J. L.; White, J. M.; Sun, Y. M.; Tang, W. J.; Henkelman, G.; Bard, A. J. *Langmuir* **2006**, *22*, 10426.

(14) Koper, M. T. M.; Shubina, T. E.; van Santen, R. A. *J. Phys. Chem. B* **2002**, *106*, 686.

(15) Inderwildi, O. R.; Jenkins, S. J.; King, D. A. *J. Phys. Chem. C* **2008**, *112*, 1305.

(16) Henkelman, G.; Arnaldsson, A.; Jónsson, H. *Comput. Mater. Sci.* **2006**, *36*, 354.

(17) Yim, W.-L.; Klüner, T. *J. Comput. Chem.* **2008**, *29*, 1306.

(18) Sanchez-Escribano, V.; Vargas, M. A. L.; Finocchio, E.; Busca, G. *Appl. Catal., A* **2007**, *316*, 68.

(19) Morgan, G. A.; Kim, Y. K.; Yates, J. T. *Surf. Sci.* **2007**, *601*, 3548.

(20) Mortensen, J. J.; Hammer, B.; Nørskov, J. K. *Phys. Rev. Lett.* **1998**, *80*, 4333.

(21) Kim, Y. K.; Morgan, G. A.; Yates, J. T. *Chem. Phys. Lett.* **2006**, *431*, 317.

(22) Kim, Y. K.; Morgan, G. A.; Yates, J. T. *Chem. Phys. Lett.* **2006**, *422*, 350.

(23) Pratt, S. J.; King, D. A. *Surf. Sci.* **2003**, *540*, 185.

(24) Liu, Z. P.; Hu, P. *J. Am. Chem. Soc.* **2001**, *123*, 12596.

(25) Jenkins, S. J.; King, D. A. *J. Am. Chem. Soc.* **2000**, *122*, 10610.

(26) Vaari, J.; Lahtinen, J.; Talo, A.; Hautojarvi, P. *Surf. Sci.* **1991**, *251*, 1096.

(27) Wu, M. C.; Dong, S. Z.; Zhu, A. R. *Surf. Sci.* **1989**, *216*, 420.

(28) Chen, J. G.; Crowell, J. E.; Ng, L.; Basu, P.; Yates, J. T. *J. Phys. Chem.* **1988**, *92*, 2574.

(29) Ng, L.; Uram, K. J.; Xu, Z.; Jones, P. L.; Yates, J. T. *J. Chem. Phys.* **1987**, *86*, 6523.

(30) Paul, J. *Nature* **1986**, *323*, 701.

(31) Konsolakis, M.; Yentekakis, I. V. *Appl. Catal., B* **2001**, *29*, 103.

(32) Hammer, B.; Jacobsen, K. W.; Nørskov, J. K. *Surf. Sci.* **1993**, *297*, L68.

(33) Zhang, J.; Chen, J.; Li, Y.; Sun, Y. *J. Nat. Gas Chem.* **2002**, *11*, 99.

(34) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511.

(35) Kresse, G.; Hafner, J. *Phys. Rev. B* **1993**, *47*, 558.

(36) Kresse, G.; Hafner, J. *Phys. Rev. B* **1994**, *49*, 14251.

(37) Kresse, G.; Furthmüller, J. *Phys. Rev. B* **1996**, *54*, 11169.

(38) Kresse, G.; Furthmüller, J. *Comput. Mater. Sci.* **1996**, *6*, 15.

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

(40) Kresse, G.; Joubert, D. *Phys. Rev. B* **1999**, *59*, 1758.

(41) Henkelman, G.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9978.

(42) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9901.

(43) Reboredo, F. A.; Galli, G. *J. Phys. Chem. B* **2006**, *110*, 7979.

(44) Ma, Q. M.; Xie, Z.; Wang, J.; Liu, Y.; Li, Y. C. *Phys. Lett. A* **2006**, *358*, 289.

(45) Lu, X.; Tian, F.; Xu, X.; Wang, N. Q.; Zhang, Q. *J. Am. Chem. Soc.* **2003**, *125*, 10459.

(46) Xu, L. J.; Henkelman, G.; Campbell, C. T.; Jónsson, H. *Phys. Rev. Lett.* **2005**, *95*, 146103.

(47) Savin, A.; Jepsen, O.; Flad, J.; ersen, O. K.; Preuss, H.; Vonschnering, H. G. *Angew. Chem., Int. Ed.* **1992**, *31*, 187.

(48) Bader, R. F. W. *Chem. Rev.* **1991**, *91*, 893.

(49) Costales, A.; Kandalam, A. K.; Pendas, A. M.; Blanco, M. A.; Recio, J. M.; Pandey, R. *J. Phys. Chem. B* **2000**, *104*, 4368.

(50) Jensen, F. Introduction to Computational Chemistry, 1st ed.; Wiley: West Sussex, 1999.

(51) Nosé, S. *Mol. Phys.* **1984**, *52*, 255.

(52) Lu, X.; Zhang, L.; Xu, X.; Wang, N.; Zhang, Q. *J. Phys. Chem. B* **2002**, *106*, 2136.

(53) Liu, Z.-P.; Hu, P.; Alavi, A. *J. Am. Chem. Soc.* **2002**, *124*, 14770.

(54) Gong, X. Q.; Raval, R.; Hu, P. *Surf. Sci.* **2004**, *562*, 247.

CT800243Y

JCTC Journal of Chemical Theory and Computation

# Molecular Modeling of Geometries, Charge Distributions, and Binding Energies of Small, Druglike Molecules Containing Nitrogen Heterocycles and Exocyclic Amino Groups in the Gas Phase and in Aqueous Solution

Brian R. White,[†] Carston R. Wagner,[†,‡] Donald G. Truhlar,[‡] and Elizabeth A. Amin*,[†]

*Department of Medicinal Chemistry, College of Pharmacy, and Department of Chemistry and Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455*

Received March 5, 2008

**Abstract:** We have tested a variety of approximate methods for modeling 30 systems containing mixtures of nitrogen heterocycles and exocyclic amines, each of which is studied with up to 31 methods in one or two phases (gaseous and aqueous). Fifteen of the systems are protonated, and fifteen are not. We consider a data set consisting of geometric parameters, partial atomic charges, and water binding energies for the methotrexate fragments 2-(aminomethyl)pyrazine and 2,4-diaminopyrimidine, as well as their cationic forms 1*H*-2-(aminomethyl)pyrazine and 1*H*-2,4-diaminopyrimidine. We first evaluated the suitability of several density functionals with the 6−31+G(d,p) basis set to serve as a benchmark by comparing calculated molecular geometries to results obtained from coupled-cluster [CCSD/6−31+G(d,p)] wave function theory (WFT). We found that the M05-2X density functional can be used to obtain reliable geometries for our data set. To accurately model partial charges in our molecules, we elected to use the well-validated charge model 4 (CM4). In the process of establishing benchmark values, we consider gas-phase coupled cluster and density functional theory (DFT) calculations, followed by aqueous-phase DFT calculations, where the effect of solvent is treated by the SM6 quantum mechanical implicit solvation model. The resulting benchmarks were used to test several widely available and economical semiempirical molecular orbital (SE-MO) methods and molecular mechanical (MM) force fields for their ability to accurately predict the partial charges, binding energies to a water molecule, and molecular geometries of representative fragments of methotrexate in the gaseous and aqueous phases, where effects of water were simulated by the SM5.4 and SM5.42 quantum mechanical implicit solvation models for SE-MO and explicit solvation was used for MM. In addition, we substituted CM4 charges into the MM force fields tested to observe the effect of improved charge assignment on geometric and energetic modeling. The most accurate MM force fields (with or without the CM4 charges substituted) were validated against gas-phase and aqueous-phase geometries and charge distributions of a larger set of 16 druglike ligands, both neutral and cationic. This process showed that the Merck Molecular Force Field (MMFF94) with or without CM4 charges substituted, is, on average, the most accurate force field for geometries of molecules containing nitrogen heterocycles and exocyclic amino groups, both protonated and unprotonated. This force field was then applied to the complete methotrexate molecule, in an effort to systematically explore its accuracy for trends in geometries and charge distributions. The most accurate force fields for the binding energies of nitrogen heterocycles to a water molecule are OPLS2005 and AMBER.

## 1. Introduction

Continuing advances in molecular modeling and computational chemistry have greatly facilitated the structure-based design of small-molecule inhibitors of proteins.[1−15] Although molecular mechanics (MM) force fields[16−20] can model protein structure, they often lack parameters that accurately represent the heteroatomic groups present in pharmaceuticals.[21−23] Density functional theory[24] (DFT) and wave function theory (WFT)[25] do not require new parameters for each type of atom; however, current technology still limits the calculations to smaller molecules and exploratory studies

* To whom correspondence should be addressed. E-mail: eamin@umn.edu.

[†] Department of Medicinal Chemistry, College of Pharmacy.

[‡] Department of Chemistry and Minnesota Supercomputing Institute.

10.1021/ct8000766 CCC: $40.75 
Published on Web 09/18/2008

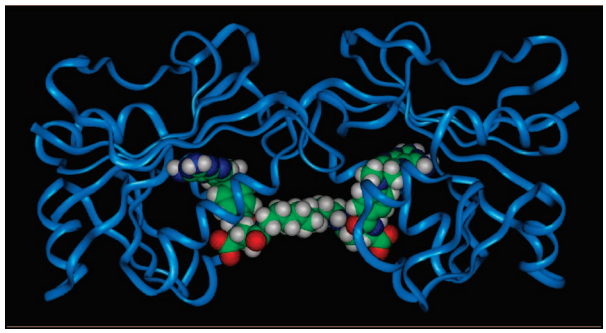


***Figure 1.*** $DHFR_2MTX_2$ chemically induced dimer.

on larger systems. Two viable approaches for simulating a protein bound to a druglike inhibitor are to obtain MM parameters for force fields that yield accurate molecular geometries and partial charges or to find a suitable level of combined QM/MM theory[26−28] in which a critical or active region of the system is treated by quantum mechanics (QM) and the surrounding areas by MM. An economical QM level for such calculations would be semiempirical molecular orbital[29−34] (SE-MO) theory. For small enough QM regions or short simulations, one can also use more reliable QM methods such as DFT. Reliable WFT calculations are, however, affordable only for the smallest systems.

A recent application of molecular modeling is the prediction of mutation effects on protein−protein interactions.[35−42] Protein multimer stability can be modified through the introduction of interfacial residue mutations, and it would be valuable to be able to predict the relative change in stability of a mutated protein multimer compared to the wild-type species. Such calculations would aid in understanding the functional evolution of proteins, as well as the development and control of stable, self-assembled protein structures, with applications ranging from nanoscale multiprotein constructs to drug delivery. With the advent of chemically induced dimerization (CID), our laboratory has demonstrated the ability to create self-assembled *E. coli* dihydrofolate reductase (DHFR) dimers from naturally existing DHFR monomers using a bivalent methotrexate dimerizer ($MTX_2$, complex−$DHFR_2MTX_2$) (Figure 1).[43] While the nature and effects of linker length have been examined previously, we are still in the early stages of in vitro and in silico observation of interfacial mutation effects on the dimer. We have found that the introduction of complementary interfacial mutations putatively leads to a stabilized DHFR heterodimer, which allows for a level of control over the assembly of such constructs.

One complicating issue present in our system, as well as other biological systems, is the protonation state of the ligand in solution and in complex with the protein. While the DHFR inhibitor MTX is unprotonated in solution, it is protonated on N1 (Figure 2) when bound to DHFR.[44,45] This raises the question of whether it is appropriate to use a single set of MM parameters to describe MTX both in solution and bound to the enzyme. To assist the in silico prediction of mutation effects on dimer stability, we have undertaken a study to develop a set of MM parameters that can accurately model the $DHFR_2MTX_2$ complex. To accomplish this goal, we will first try to establish an accurate method to model the single

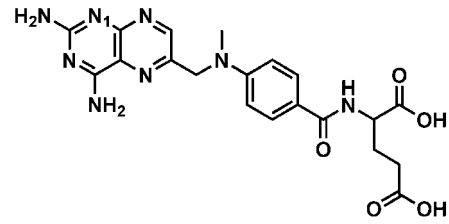



***Figure 2.*** Structure of methotrexate.

substrate, MTX, and then try to extend this method to the $DHFR_2MTX_2$ complex. During this process, we have tested a large variety of methods on a set of druglike molecules containing nitrogen heterocycles and exocyclic amino groups, and the results of these tests are presented in the present article because they should be of general interest for a variety of potential applications.

A critical issue in simulating systems with nitrogen heterocycles is modeling the charge distributions. Because partial charge distributions are not experimental observables, we will rely on theory to establish reasonable values. For this purpose, we first require accurate geometries, and we begin by establishing benchmark values using high-level WFT and DFT calculations on the MTX fragments 2-(aminomethyl)pyrazine (2-AMP), 1*H*-2-(aminomethyl)pyrazine (1*H*-2-AMP), 2,4-diaminopyrimidine (2,4-DAP), and 1*H*-2,4-diaminopyrimidine (1*H*-2,4-DAP). These fragments were chosen because of the role of the pteridine moiety in MTX binding to the DHFR active site. We then use DFT with class IV charges[46] to establish benchmark partial atomic charges. Coupled cluster theory[47,48] with single and double excitations (CCSD) and the M05-2X[49] density functional with the 6−31+G(d,p)[50] basis set are used for geometries, and charge model 4[51] (CM4) is used for partial atomic charges. The performance of widely available SE-MO and MM parametrizations is then surveyed for these four fragments to find the parametrized model that most accurately predicts the geometries, binding energies to water, and charge distribution of the unprotonated and protonated states of 2-AMP and 2,4-DAP. In addition, we consider the selected MM methods when CM4 charges are substituted for the force field's default charges in an effort to observe if increased accuracy in partial charge distribution leads to increased performance in geometric and energetic modeling. The most accurate methods are then used to calculate partial charges and geometries for a series of pharmacophorically similar molecules containing nitrogen heterocycles and exocyclic amines, and these results are compared to DFT and CM4 benchmarks to explore the validity of our chosen MM parameters more broadly.

## 2. Methods and Software

**2.1. Computational Methods.** For geometries, we consider three categories of QM theory plus MM. The QM categories are WFT, DFT, and SE-MO. For partial charges, we consider Mulliken population analysis,[52] MM[53−57] and the CM1,[46] CM2,[58] CM3,[59] and CM4[51] charge models. DFT calculations in the aqueous phase use the implicit solvation model SM6,[51] while solvation in SE-MO methods[30−32] was included by using the implicit SM5.4[60] or SM5.42[61] solvation

models. For aqueous-phase MM calculations, we employed explicit solvation using the respective programs' soak algorithms in conjunction with periodic boundary conditions (minimum cell size of 15 × 15 × 15 Å) to eliminate solvent−vacuum interfaces.

WFT calculations were carried out by CCSD with the 6−31+G(d,p)[50] basis set. These calculations were carried out with the *Gaussian03* computer program (Gaussian, Inc.).[62] The CCSD method was chosen over the popular Møller−Plesset second-order perturbation theory[63] as CCSD (or the closely related QCISD[64]) has been shown to yield more accurate geometries.[65]

The DFT methods examined are B3LYP,[66,67] mPW1PW[68,69] (which is also called mPW1PW91, mPW0, and MPW25), MPWB1K,[67,69,70] MPW1KCIS,[71] M06-L,[72] and M05-2X[49] with the 6−31+G(d,p)[50] basis set. Calculations were performed using a locally modified version of the *Gaussian03* program incorporating the MN-GSM 6.0[73] and MN-GFM 2.0.1[74] solvation and DFT modules. To select a density functional to generate benchmark values, gas-phase calculations were carried out, and the resulting geometries were evaluated relative to CCSD. The best functional was subsequently used to perform calculations in the aqueous phase using the SM6[51] solvation model for implicit solvation.

We have tested charge model 4 (CM4) partial charge assignments based on gas-phase and SM6 DFT calculations. CM4 charges, the fourth generation of class IV[46] charges, have a distinct advantage over the class II[52,75−77] and III[78−80] charges used in *Gaussian03*. Whereas the reliability of class III charges depends on the wave function and basis set used, class IV charges represent an extrapolation to full configuration interaction with a complete basis set.[46,51] Furthermore, class III charges are unstable with respect to buried charges,[81−84] while class IV charges provide a reliable method for obtaining buried charges. CM4 charges, in particular, have been parametrized against a large training set (398 molecules) and are well suited for modeling aliphatic functional groups, which makes them more suitable for modeling hydrophobic effects, a primary factor in protein−protein interactions.

SE-MO methods examined in the current study include AM1,[30] PM3,[31,32] and PDDG/PM3.[85] Calculations were performed for AM1 and PM3 using AMSOL 7.1[86] (a derivative of AMPAC 2.1) with Mulliken, CM1, CM2, or CM3 charges obtained from gas-phase calculations. For aqueous-phase AM1 and PM3 calculations, AMSOL 7.1 was used to obtain Mulliken, CM1, or CM2 charges within the SM5.4 (for Mulliken and CM1) or the SM5.42 (for CM2) solvation models. GAMESSPLUS[87] was used to obtain a second set of CM3 charges in the gas phase in an effort to test consistency in charge assignment across software. The notation used in this article for AM1 calculations with differing partial charge assignments is AM1, AM1-CM1, AM1-CM2, and AM1-CM3 for AM1 calculations with Mulliken, CM1, CM2, and CM3 charges, respectively. The notation for the PM3 calculations is analogous to that for AM1. It is important to note that because they are post-self consistent field (SCF) analysis tools, CM*x* or Mulliken charges of gas-phase wave functions do not alter an optimized molecule's geometry. Slight differences may be attributed to variances in the convergences of the SCF and geometry optimizations. In solution, each SM*x* model uses a particular choice of charge model; this choice, along with all the other SM*x* parameters, does affect the molecules' geometry. PDDG/PM3 gas-phase optimizations were performed using MOPAC 5.011mn,[88] and Mulliken partial atomic charges were obtained. In addition, PM3-Mulliken charge analyses were carried out with *Gaussian03* and MOPAC 5.011mn as part of a comparison between partial atomic charges assigned to optimized geometries calculated by PM3 and PDDG/PM3.

The MM force fields employed are AMBER,[55] AMBER*, CVFF,[53] CFF91,[89] MMFF94,[56] OPLS2005,[57] and Tripos.[54] The AMBER force field employed is the ff03 version,[90] in conjunction with the general atom force field[91] (GAFF) commonly utilized for small organic systems. The AMBER* force field contains additional atomic parameters as implemented in *MacroModel*. In most cases, we elected to use nonrigid water with each force field's default parameters for water molecules. However, the AMBER* force field was locally modified to use OPLS2005 nonrigid water (in OPLS2005, the nonrigid water has the same Lennard-Jones parameters as the rigid TIP3P water model), and the AMBER force field, via the SOLVATEOCT command, utilized the rigid TIP3P water model. Stretch, bend, Coulombic, and Lennard-Jones parameters for the water models used can be found in each of the force field's descriptions (see references above), except for AMBER*, for which the modified water parameters are described by the OPLS2005 reference.

In addition to the standard force fields, we also employ local modifications of the force fields that substitute CM4 charges for their default partial charges. This combination of a force field and CM4 charges is denoted as X-CM4, where X is the name of the original force field. In contrast to gas-phase SE-MO calculations, when CM*x* charges are used with MM, they can and do alter the optimized geometry. Note that when we use CM4 charges with MM calculations, we use gas-phase M05-2X/6−31+G(d,p)/CM4 charges calculated at gas-phase M05-2X/6−31+G(d,p) geometries for gas-phase MM calculations, and we use SM6/M05-2X/6−31+G(d,p)/CM4 aqueous-phase charges calculated at aqueous-phase SM6/M05-2X/6−31+G(d,p) geometries for aqueous-phase MM calculations.

General MM optimization conditions consisted of at least 1000 steps of conjugate gradient minimization with an energy gradient cutoff of 0.01 kcal $mol^{-1}$ $Å^{-1}$. In cases where water was included, the entire system was minimized, and although the stringent conditions for termination we have established were not typically reached, the system energy fluctuated only slightly around a stable potential energy. To test the effect of water position around the small molecules on geometries, six random orientations of 2-AMP and 2,4-DAP within the water box were used in minimizations with OPLS2005-CM4 and MMFF94-CM4. Because this process yielded only a slight standard deviation in geometries (on the order of less than 0.006 Å in bond length and 0.7 degrees in bond angle), we used only the default water orientation for all other aqueous-phase MM optimizations. We note, however, the

2-AMP and 2,4-DAP bond lengths and angles used for OPLS2005-CM4 and MMFF94-CM4 assessment represent the average of these six results. In binding energy studies, water molecules were placed at positions on 2-AMP, 1*H*-2-AMP, 2,4-DAP, and 1*H*-2,4-DAP, where the DHFR-MTX crystal structure denoted that hydrogen bonding takes place between the ligand and the enzyme.[45] The small molecule and water were optimized together, and the binding energy of the complex computed by subtracting the energies of the individual optimized molecules.

While testing the MM force fields, it was found that the CVFF and CFF91 force fields do not properly assign partial atomic charges to the 1*H*-2-AMP or 1*H*-2,4-DAP cations because the total charge assigned to the molecule is not +1.0. We therefore calculated the gas-phase Hartree−Fock molecular electrostatic potential with the 6−31+G(d,p) basis set for the neutral and cationic species and obtained ChelpG[80] electrostatic-potential-filling charges to substitute into the force fields. Geometry optimizations with the ChelpG partial charges were then carried out in both the gaseous and aqueous phases, and the resulting geometries, denoted CVFF-HF and CFF91-HF, were used in our evaluation.

**2.2. Platforms, Software, and Molecules.** Quantum mechanical calculations (WFT, DFT, and SE-MO) were performed on an IBM Power4 (p690 and p655) computer system running under the AIX operating system and an SGI Altix cluster running under the Linux operating system. Molecular mechanics calculations were performed on a Silicon Graphics O2 workstation running under the IRIX 6.5 operating system. Molecules were constructed for quantal calculations using the *GaussView 3.0* (Gaussian, Inc.) visualization program, and the generated *Z*-matrices were converted to Cartesian coordinates where appropriate. Molecules for MM calculations were constructed using *InsightII 2005* (Accelrys, Inc.) for the CFF91 and CVFF force fields, SYBYL 7.3 (Tripos, Inc.) for the MMFF94, AMBER and Tripos force fields, and *Maestro 7.5* (Schrodinger, Inc.) for the MMFF94, AMBER*, and OPLS2005 force fields (utilizing the *MacroModel* and *Impact* applications, respectively). These programs were also used to set up and run the MM minimizations except for AMBER minimizations, for which SYBYL was used only to generate molecular coordinates in Protein Data Bank or Mol2 formats, and the AMBER 9[92] suite was used for the minimizations. The small molecules included in the present study are illustrated in Figure 3. Each molecule was modeled in both its neutral and protonated form. When an exocyclic amine is present, the proton was added there. Otherwise, the proton was added on a heterocyclic amine.

## 3. Results and Discussion

**3.1. Error Analysis.** To rank the methods we chose to test, we calculated the unsigned residual between a calculated value with method *m* and the corresponding benchmark value

$$R_i^{x,y,m} = |x_i^{\text{calcd},m}(y) - x_i^{\text{benchmark}}(y)| \quad (1)$$

where *y* is the phase (gas phase or aqueous phase) and $x_i(y)$ signifies case *i* of molecular property *x* in phase *y*, for example, when *x* is bond length (*r*), $x_1(y)$ is the first bond

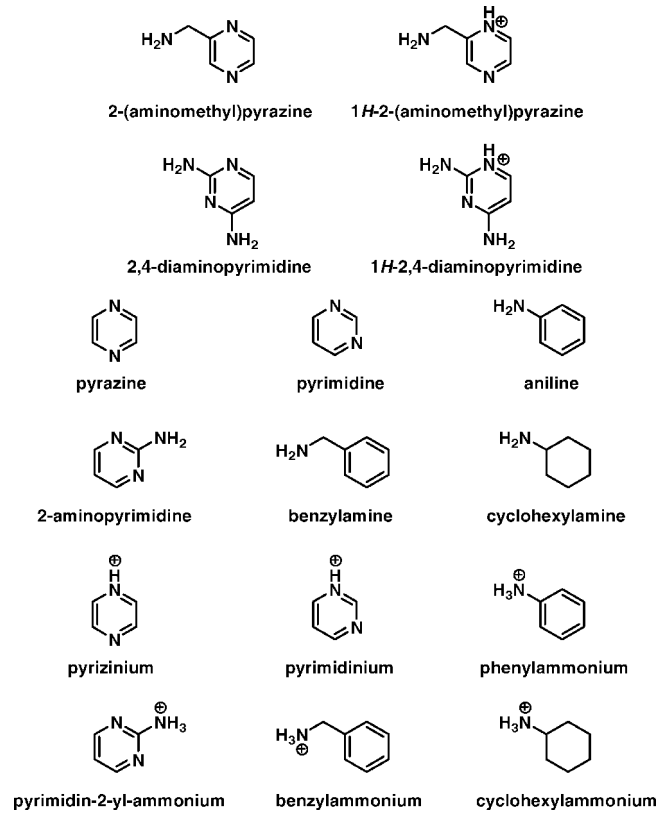



***Figure 3.*** Structures of small molecules used in current study.

length. Alternatively, *x* could stand for partial charge (*q*), bond angle ($\theta$), or binding energy ($E_b$). The overall error in a particular molecular property *x* for a particular method *m* and phase *y* is quantified by the mean unsigned error (MUE)

$$\text{MUE}_{x,y,m} = \frac{\sum_{i=1}^{n_{x,y}} R_i^{x,y,m}}{n_{x,y}} \quad (2)$$

where $n_{x,y}$ is the number of combinations of $x_i$ and *y* for which $R_i^{x,y}$ is evaluated.

We also calculate the average mean unsigned error (AMUE) for each property across *N* methods

$$\text{AMUE} = \frac{\sum_{m=1}^{N} \text{MUE}_{x,y,m}}{N} \quad (3)$$

Division of this value by a method's MUE yields the reduced (unitless) mean unsigned error (RMUE) for the method for partial charge, bond length, or bond angle.

$$\text{RMUE}_{x,y,m} = \frac{\text{MUE}_{x,y,m}}{\text{AMUE}_{x,y}} \quad (4)$$

The RMUE is a measure of each method's performance relative to the mean of the others for calculating a particular molecular property. A value of 1.0 indicates that the method is average. Lower values indicate better methods, while higher values indicate worse methods. The reason for introduction of these unitless reduced quantities is so that we can combine errors for *r* and $\theta$ (which have different

***Table 1.*** Mean Unsigned Error (Å and deg) for Gas-Phase Bond Lengths and Angles Between Those Calculated with CCSD and Each Functional

| | 2-AMP | | 1*H*-2-AMP | | 2,4-DAP | | 1*H*-2,4-DAP | |
|---|---|---|---|---|---|---|---|---|
| method | bond length | bond angle | bond length | bond angle | bond length | bond angle | bond length | bond angle |
| B3LYP | 0.003 | 0.73 | 0.003 | 0.37 | 0.003 | 0.82 | 0.003 | 0.27 |
| MPW1PW | 0.005 | 0.44 | 0.005 | 0.49 | 0.005 | 0.87 | 0.004 | 0.26 |
| mPWB1K | 0.011 | 0.52 | 0.010 | 0.42 | 0.012 | 0.99 | 0.010 | 0.26 |
| mPW1KCIS | 0.005 | 0.78 | 0.005 | 0.50 | 0.004 | 0.90 | 0.004 | 0.28 |
| M06-L | 0.006 | 0.52 | 0.005 | 0.33 | 0.004 | 0.54 | 0.004 | 0.22 |
| M05-2X | 0.004 | 0.28 | 0.004 | 0.26 | 0.005 | 0.94 | 0.003 | 0.22 |
| AMUE[a] | 0.006 | 0.54 | 0.005 | 0.40 | 0.006 | 0.84 | 0.005 | 0.25 |

[a] Mean of entire column.

units) to make an overall assessment for combined geometric performance.

We define a reduced deviance ($D_{y,m}$) of a SE-MO or MM method from the average performance in either the gas or aqueous phase by averaging the RMUE for $r$ and $\theta$ in each method for both 2-AMP and 1*H*-2-AMP

$$D_{y,m} = \frac{1}{4}(\mathrm{RMUE}_{r,y,m}^{\text{2-AMP}} + \mathrm{RMUE}_{\theta,y,m}^{\text{2-AMP}} + \mathrm{RMUE}_{r,y,m}^{1H\text{-2-AMP}} + \mathrm{RMUE}_{\theta,y,m}^{1H\text{-2-AMP}}) \quad (5)$$

The reduced deviance for 2,4-DAP and 1*H*-2,4-DAP is averaged with the reduced deviance for 2-AMP and 1*H*-2-AMP to yield an overall performance. Reduced deviance in partial charge ($q$) assignment is calculated similarly. Reduced deviance for the validation of the most accurate methods is also calculated similarly, with all molecules taken into account.

**3.2. Establishing Geometry and Partial Charge Benchmark Sets.** The MUE of the gas-phase geometries calculated by DFT with respect to the CCSD-calculated geometries is given in Table 1. All DFT methods perform well, with mean unsigned errors within ~0.01 Å for bond length and less than one degree for bond angles. On the basis of its high degree of accuracy for both bond length and angle, we chose to proceed with the M05-2X level of theory to generate our geometric benchmark set.

We selected the well-validated[51,59,93−96] CM4 charge model to obtain benchmark partial atomic charges. To examine whether the CM4 charge model would assign partial charge similarly for each functional used, we compared the gas-phase CCSD partial charges generated by Mulliken population analysis (a class II[46] charge method) to CM4 partial charges assigned by each density functional tested (note that the CM4 charges are probably more accurate). The results are summarized in Figure 4. In this figure and in this whole article, charges are given in atomic units, in which the charge on a bare proton is 1.0. Overall, the mean unsigned deviations between the CCSD Mulliken analysis and the CM4-assigned partial charges vary by ≤0.01 charge units regardless of the functional or phase tested. We therefore applied the geometrically accurate M05-2X functional in conjunction with M05-2X/CM4 partial charges to obtain our benchmark set.

**3.3. Exploration of CVFF and CFF91 Atom Typing and Charge Distribution.** As mentioned in section 2.2, a deficiency in the CVFF and CFF91 force fields is that they

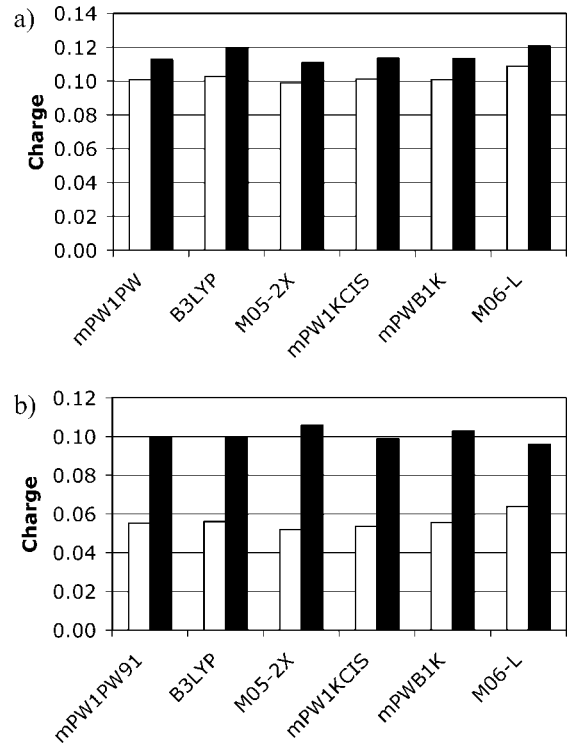



***Figure 4.*** Mean unsigned deviation of DFT/CM4 charges relative to CCSD/Mulliken charges for (a) 2-AMP and it is cation (white, 2-AMP; black, 1*H*-2-AMP) and (b) DAP and it is cation (white, DAP; black, 1*H*-DAP).

do not assign a total charge of +1.0 to the pyrazinium or pyridinium cations. Upon further exploration using 2-AMP, the default atom type assigned to N1 in 2-AMP by both force fields is "np" (an $sp^2$ nitrogen in a 5- or 6-membered ring), and the partial charge on this unprotonated nitrogen is −0.22 in CVFF and −0.48 in CFF91. Upon protonation and automated reassignment of atom types by *InsightII 2005*, the CVFF nitrogen atom type remains unchanged, and the CFF91 atom type changes to "nh+" (a protonated nitrogen in a 6-membered ring). The protons added have partial charges of +0.28 (CVFF) and +0.33 (CFF91), and the partial charges on the nitrogens change to −0.50 and −0.81 charge units, respectively. The partial charge on all other atoms in the molecule are unaffected by the addition of the proton. This charge balancing yields a total charge on the molecule of 0, which is incorrect for a cation.

Some of the partial atomic charges in the CVFF and CFF91 force fields are derived from fits to the Hartree−Fock molecular electrostatic potential,[97−99] and we took this as a cue for how to correct the problem in a way consistent with these force fields. In particular, we carried out single-point, gas-phase Hartree−Fock (6−31G(d,p) basis set) calculations on the optimized 2-AMP and 1*H*-2-AMP geometries and obtained partial atomic charges by electrostatic potential fitting with the ChelpG[80] algorithm. The resulting partial atomic charges were used in the CVFF and CFF91 force fields for geometry optimizations in both the gas and aqueous phases. We used the gas-phase partial charges in both phases because the original CVFF and CFF91 partial charges are based on gas-phase calculations (we note that all MM force fields considered in this article use partial charges that do

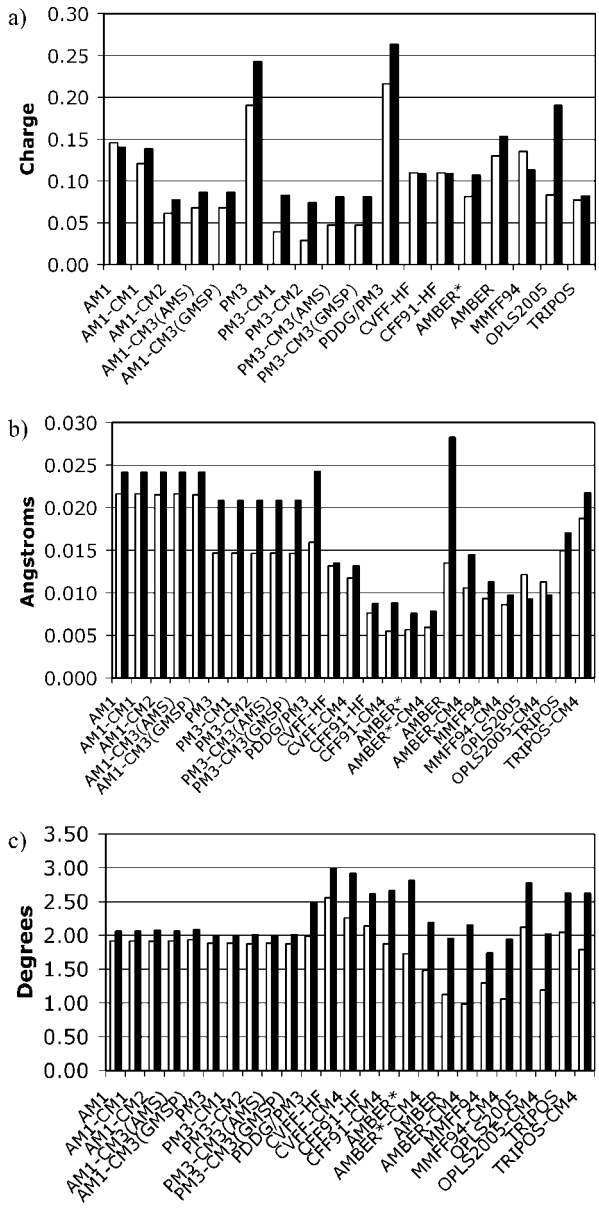



***Figure 5.*** MUE in (a) partial charge, (b) bond length, and (c) bond angle for selected SE-MO and MM methods in the gas phase relative to M05-2X/CM4 (white, 2-AMP; black, 1*H*-2-AMP). AMS denotes AMSOL, and GMSP denotes GAMESS-PLUS.

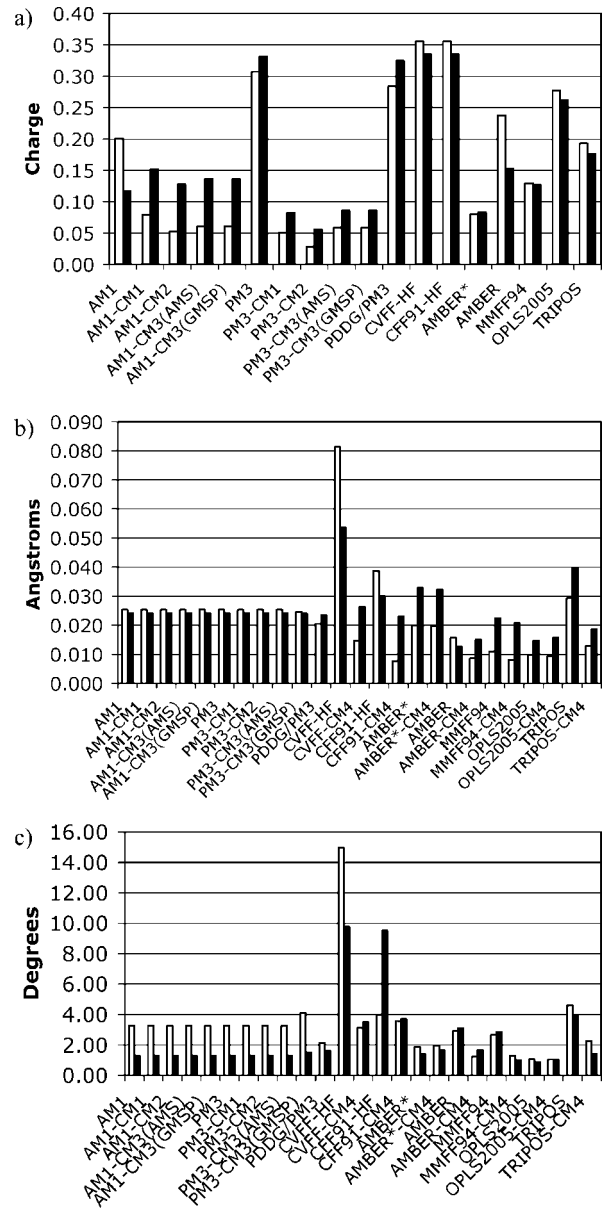



***Figure 6.*** MUE in (a) partial charge, (b) bond length, and (c) bond angle for selected SE-MO and MM methods in the gas phase relative to M05-2X/CM4 (white, 2,4-DAP; black, 1*H*-2,4-DAP).

not depend on the phase). To denote the new Hartree−Fock partial charges in the force fields, we name the force fields that use these newly assigned charges as CVFF-HF and CFF91-HF.

**3.4. Evaluation of SE-MO and MM Calculations.** We tested available SE-MO and MM parameter sets in an effort to select the most accurate method for modeling our small molecules both in the gas phase and in solvent. In addition, we substituted CM4 charges into the MM force fields tested to observe effects on geometric accuracy. A summary of the results of these tests is given in Figures 5 and 6 (gas phase) and 7 and 8 (aqueous phase). In the gas phase, the AM1 and PM3 SE-MO methods utilizing the CM1, CM2, and CM3 charge models, as well as the CVFF-HF, CFF91-HF, AMBER*, and MMFF94 force fields all predict the partial charges of the atoms in both sets of molecules to within 0.15 (Figures 5a and 6a). In most of the methods tested, the partial charge assignment becomes less accurate in the cationic species: the most notable examples, for which the average error increases by a factor of 2, are the OPLS2005 force field for 1*H*-2-AMP, the PM3-CM1, -CM2, and -CM3 methods for 1*H*-2-AMP, and the SE-MO methods utilizing the CM1, CM2, and CM3 charge models for 1*H*-2,4-DAP (with the exception of PM3-CM3). In contrast, the MMFF94 force field becomes more accurate for 1*H*-2-AMP, and the AMBER force field becomes more almost 2-fold more accurate for 1*H*-2,4-DAP, although AMBER's overall charge assignment is somewhat inaccurate (the mean MUE for the neutral and cationic species together is 0.20). The force fields with CM4 charges substituted are not included in the partial charge analysis since their mean unsigned error in partial charge is always zero.

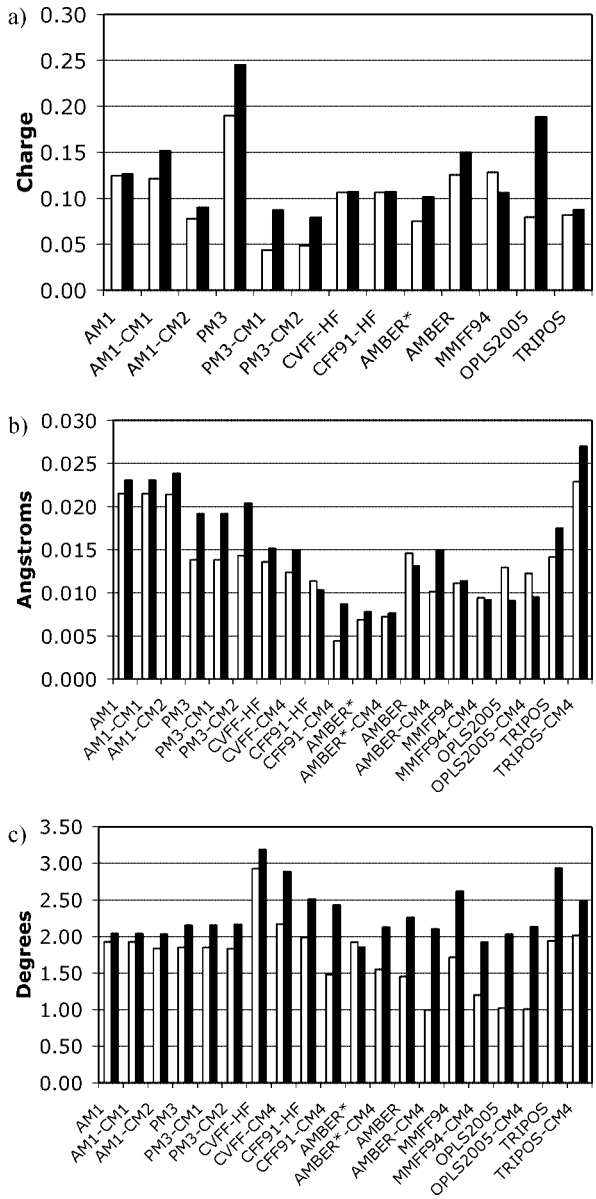



***Figure 7.*** MUE in (a) partial charge, (b) bond length, and (c) bond angle for selected SE-MO and MM methods in solution relative to M05-2X/CM4 (white, 2-AMP; black, 1*H*-2-AMP).

The gas-phase partial charges assigned by the PDDG/PM3 method have a mean unsigned error of 0.22 and 0.28 for the neutral species and 0.26 and 0.32 for the cationic species of 2-AMP and 2,4-DAP, respectively. Table 2 summarizes our comparison of the PDDG/PM3 charges to Mulliken population analysis of the PM3-optimized structure of 2-AMP as calculated by both *Gaussian03* and MOPAC 5.011mn (the numbering system is given in Figure S1 of the Supporting Information); this comparison verifies the consistency of the two programs (as a check). Figures 5 and 6 show that PDDG/PM3 is less accurate than PM3 for gas-phase partial charges, bond lengths, and bond angles for both 2-AMP and 1*H*-2-AMP and for gas-phase bond angles for both 2,4-DAP and 1*H*-2,4-DAP. Because the PDDG reparameterization of PM3 deteriorated the performance for these molecules, PDDG/PM3 was not considered in the aqueous-phase calculations that follow.

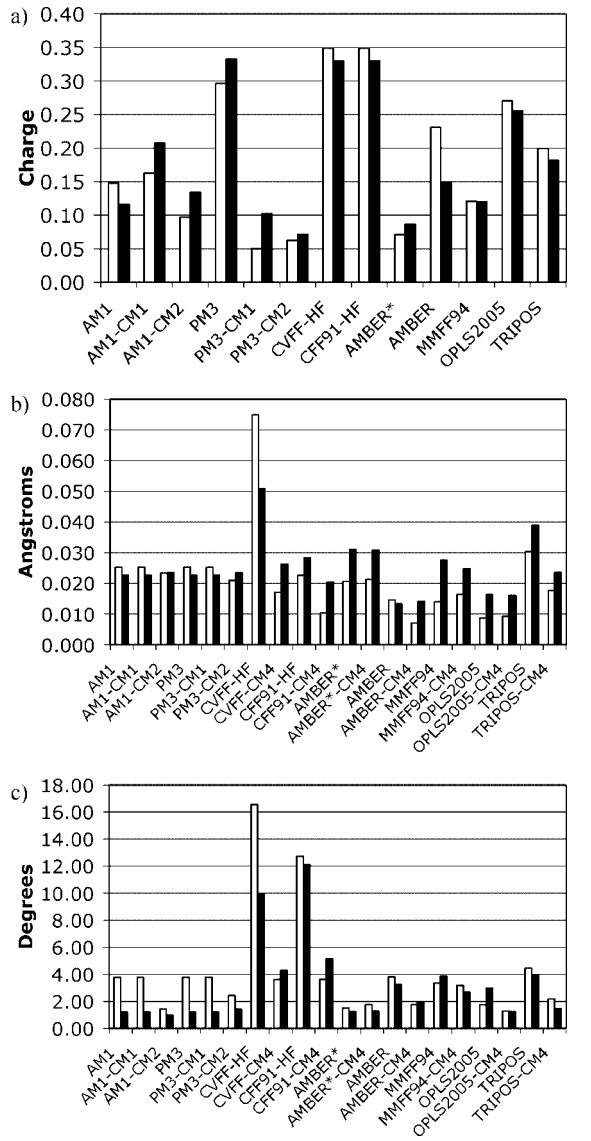



***Figure 8.*** MUE in (a) partial charge, (b) bond length, and (c) bond angle for selected SE-MO and MM methods in solution relative to M05-2X/CM4 (white, 2,4-DAP; black, 1*H*-2,4-DAP).

***Table 2.*** Gas-Phase Partial Charges of 2-AMP Calculated by PM3 (Mulliken Population Analysis) and PDDG/PM3

| atom | PM3(G03) | PM3(MOPAC) | PDDG/PM3(MOPAC) |
|---|---|---|---|
| C1 | −0.1310 | −0.1309 | −0.1316 |
| C2 | −0.1067 | −0.1067 | −0.1647 |
| N3 | −0.0382 | −0.0382 | −0.0274 |
| C4 | −0.1095 | −0.1094 | −0.1664 |
| C5 | −0.1085 | −0.1085 | −0.1581 |
| N6 | −0.0239 | −0.0239 | −0.0211 |
| H7 | 0.1344 | 0.1344 | 0.1814 |
| H8 | 0.1319 | 0.1319 | 0.1798 |
| H9 | 0.1316 | 0.1316 | 0.1789 |
| C10 | −0.0569 | −0.0569 | −0.1491 |
| H11 | 0.0677 | 0.0677 | 0.1151 |
| H12 | 0.0830 | 0.0830 | 0.1296 |
| N13 | −0.0282 | −0.0282 | −0.0928 |
| H14 | 0.0301 | 0.0301 | 0.0665 |
| H15 | 0.0242 | 0.0242 | 0.0598 |

For 2-AMP and 1*H*-2-AMP, gas-phase bond lengths (Figure 5b) are predicted to within ∼0.01 Å only by the

CFF91-HF, AMBER*, MMFF94, and OPLS2005 force fields. All methods except OPLS2005 become less accurate when modeling the cationic species. Notably, the AMBER force field becomes 2-fold less accurate when modeling the cation, and it clearly benefits from revised charges because the AMBER-CM4 force field yields a 2-fold performance improvement when modeling bond lengths. On average, force fields with CM4 charges substituted tend to perform slightly better than those with default charges, except in the case of TRIPOS-CM4. Gas-phase bond angles (Figure 5c) are predicted to within ~2.0 degrees by all the SE-MO methods, while the majority of the MM force fields are ~0.5−1° less accurate. Exceptions include the MMFF94 and AMBER force fields, which predict them to within ~1.5°. All methods are slightly more inaccurate dealing with the cation. As with bond lengths, when CM4 charges are substituted, bond angle predictivity is increased slightly in most methods. OPLS2005-CM4 is notable in that it has a 2-fold improvement in performance over OPLS2005 when modeling the neutral species and a smaller, yet significant (~0.5°), increase in accuracy for the cationic species.

For 2,4-DAP and 1*H*-2,4-DAP, bond lengths (Figure 6b) are predicted to an accuracy of ~0.025 Å by all methods except CVFF-HF, CFF91-HF, and TRIPOS. The MMFF94 and OPLS2005 force fields both display accuracy less than 0.01 Å for 2,4-DAP, while AMBER and OPLS2005 are accurate to better than 0.015 Å for 1*H*-2,4-DAP. MMFF94 is 2-fold less accurate for the 1*H*-2,4-DAP cation. With regard to 2,4-DAP and 1*H*-2,4-DAP gas-phase bond angles, the best performing methods include the AMBER*, AMBER, MMFF94, and OPLS2005 force fields, predicting bond angle to within 1.9°, 3.1°, 2.9°, and 1.1° for both 2,4-DAP and 1*H*-2,4-DAP, respectively. With respect to force fields with CM4 charges substituted, the clearest example of a force field benefiting from revised charges is the CVFF-HF and CFF91-HF force fields which, as previously discussed, have major partial charge assignment problems. When CM4 charges are substituted into these force fields, both methods show a dramatic improvement in performance for modeling bond lengths, with CVFF-CM4 yielding a 4-fold more accurate bond length prediction overall, and CFF91-CM4 yielding the same improvement in accuracy when modeling the neutral species. The TRIPOS force field also benefits, with TRIPOS-CM4 displaying a 2-fold improvement in accuracy for both molecules.

All methods except CVFF-HF, CFF91-HF, and TRIPOS predict bond angles to within ~4.0 degrees for both species (Figure 6c). Both CVFF-HF and CFF91-HF presented a challenge when attempting to minimize structure with respect to the fact that 1*H*-2,4-DAP could not be assigned proper atomic partial charges. As described earlier, we substituted the default charges with ChelpG charges derived from HF calculations on 2,4-DAP and 1*H*-2,4-DAP. When these charges were substituted into the force field, the charges on the protons on the exocyclic amines were made largely more positive than the default charges. This effect, in conjunction with the largely negatively charged heterocyclic amines, caused the minimization to distort the $sp^3$ structure of the exocyclic amines and pull the protons toward the pyrimidine ring. The result is a large error in the assignment of bond angle. When the minimization is performed with the default charges, no distortion of bond angle takes place; however, these charges correspond to a nonphysical total molecular charge. As observed in tests already described, CM4 charge substitution greatly improves geometric modeling performance. The CVFF-CM4 and CFF91-CM4 force fields essentially eliminate the problems seen with these force fields, causing both to model bond angles with accuracy on par with the rest of the methods used. In addition, the AMBER-CM4, MMFF94-CM4, and TRIPOS-CM4 force fields all yield accuracies at least 2-fold greater than their counterparts.

For 2-AMP and 1*H*-2-AMP in the aqueous phase, the AM1-CM2, PM3-CM1, and PM3-CM2 methods, as well as the CVFF-HF, CFF91-HF, AMBER*, MMFF94, and TRIPOS force fields retain a high degree of accuracy for partial charge prediction (MUE ≤ 0.13, Figure 7a). Among these methods, the MMFF94 force field is again the only method to become more accurate when modeling the protonated species. On the other hand, the OPLS2005 force field becomes 2-fold less accurate when predicting the charge of the cation. Bond lengths (Figure 7b) are calculated with an error similar to that of the gas phase, with the CFF91-HF, AMBER*, MMFF94, and OPLS2005 force fields maintaining a mean unsigned error of ~0.01 Å. Interestingly, while it becomes much less accurate when assigning partial charge to the ionic species, the OPLS2005 force field becomes more accurate in bond length prediction. As with 2-AMP and 1*H*-2-AMP in the gas phase, CM4 partial charge substitution generally produces a modest increase in bond length accuracy. A notable exception is the CFF91-CM4 force field, with an MUE <0.005 Å. When compared to calculations in the gas phase, all methods either retain their accuracy or become slightly less accurate when predicting bond angle in solution except for the OPLS2005 force field. Except for the CVFF-HF, AMBER*, and TRIPOS force fields, all methods are still accurate to ~2.5° or less (Figure 7c). Again, on average, CM4 charge substitution slightly increases bond angle accuracy for both species.

In the aqueous-phase treatment of 2,4-DAP and 1*H*-2,4-DAP, partial charge performance is quite varied (Figure 8a). The best performing methods are the PM3-CM1 and -CM2 SE-MO methods and the AMBER* and MMFF94 force fields, which all model the partial charges to an accuracy of 0.12 or better. In the treatment of these molecules, all the force fields except AMBER* become more accurate to varying degrees when dealing with the cationic species. The quality of modeling the bond lengths (Figure 8b) is similar for most methods; however, the CVFF-HF force field is particularly inaccurate, with a MUE in bond length for 2,4-DAP of 0.075 Å. The AMBER and OPLS2005 force fields perform particularly well for both species, with MUEs of ≤0.016 Å overall. However, the OPLS2005 force field becomes 2-fold less accurate when modeling the cationic species. MMFF94 also performs relatively well, modeling the neutral species with a MUE ≤ 0.014 Å, but, like OPLS2005, becomes 2-fold less accurate when modeling the cation. CVFF-CM4, CFF91-CM4, AMBER-CM4 (for the

***Table 3.*** Reduced Deviance ($D_{y,m}$) in Partial Charge for Each SE-MO and MM Method Tested

| | 2-AMP and 1*H*-2-AMP | | | 2,4-DAP and 1*H*-2,4-DAP | | | all molecules | | |
|---|---|---|---|---|---|---|---|---|---|
| method | gas | aqueous | mean | gas | aqueous | mean | gas | aqueous | mean |
| PM3-CM2 | 0.45 | 0.56 | 0.50 | 0.25 | 0.36 | 0.31 | 0.35 | 0.46 | 0.40 |
| PM3-CM1 | 0.54 | 0.56 | 0.55 | 0.40 | 0.41 | 0.40 | 0.47 | 0.49 | 0.48 |
| AM1-CM2 | 0.63 | 0.75 | 0.69 | 0.54 | 0.62 | 0.58 | 0.58 | 0.68 | 0.63 |
| AMBER* | 0.85 | 0.78 | 0.81 | 0.49 | 0.42 | 0.46 | 0.67 | 0.60 | 0.64 |
| MMFF94 | 1.15 | 1.06 | 1.11 | 0.77 | 0.65 | 0.71 | 0.96 | 0.85 | 0.91 |
| TRIPOS | 0.73 | 0.75 | 0.74 | 1.12 | 1.03 | 1.07 | 0.92 | 0.89 | 0.91 |
| AM1-CM1 | 1.18 | 1.21 | 1.19 | 0.69 | 1.00 | 0.84 | 0.93 | 1.10 | 1.02 |
| AM1 | 1.32 | 1.12 | 1.22 | 0.97 | 0.71 | 0.84 | 1.14 | 0.92 | 1.03 |
| AMBER | 1.29 | 1.22 | 1.25 | 1.19 | 1.02 | 1.11 | 1.24 | 1.12 | 1.18 |
| OPLS2005 | 1.20 | 1.15 | 1.17 | 1.63 | 1.42 | 1.52 | 1.41 | 1.28 | 1.35 |
| CVFF-HF | 1.00 | 0.96 | 0.98 | 2.09 | 1.83 | 1.96 | 1.54 | 1.39 | 1.47 |
| CFF91-HF | 1.00 | 0.96 | 0.98 | 2.09 | 1.83 | 1.96 | 1.54 | 1.39 | 1.47 |
| PM3 | 1.96 | 1.92 | 1.94 | 1.92 | 1.70 | 1.81 | 1.94 | 1.81 | 1.88 |
| PM3-CM3(GMSP)[b,a] | 0.57 | ND | ND | 0.43 | ND | ND | 0.50 | ND | ND |
| PM3-CM3(AMS)[b,a] | 0.57 | ND | ND | 0.43 | ND | ND | 0.50 | ND | ND |
| AM1-CM3(GMSP)[b,a] | 0.70 | ND | ND | 0.58 | ND | ND | 0.64 | ND | ND |
| AM1-CM3(AMS)[b,a] | 0.70 | ND | ND | 0.58 | ND | ND | 0.64 | ND | ND |
| PDDG/PM3[a] | 2.17 | ND | ND | 1.83 | ND | ND | 2.00 | ND | ND |

[a] ND = not determined. [b] AMS = calculated using AMSOL7.1, GMSP = calculated using GAMESPLUSS.

***Table 4.*** Reduced Deviance ($D_{y,m}$) in Combined Geometry for Each SE-MO and MM Method Tested

| | 2-AMP and 1*H*-2-AMP | | | 2,4-DAP and 1*H*-2,4-DAP | | | all molecules | | |
|---|---|---|---|---|---|---|---|---|---|
| method | gas | aqueous | mean | gas | aqueous | mean | gas | aqueous | mean |
| OPLS2005-CM4 | 0.73 | 0.76 | 0.75 | 0.45 | 0.45 | 0.45 | 0.59 | 0.60 | 0.60 |
| AMBER-CM4 | 0.78 | 0.80 | 0.79 | 0.52 | 0.49 | 0.50 | 0.65 | 0.65 | 0.65 |
| MMFF94-CM4 | 0.66 | 0.71 | 0.68 | 0.50 | 0.85 | 0.68 | 0.58 | 0.78 | 0.68 |
| OPLS2005 | 0.96 | 0.76 | 0.86 | 0.43 | 0.61 | 0.52 | 0.69 | 0.68 | 0.69 |
| AMBER*-CM4 | 0.67 | 0.71 | 0.69 | 0.86 | 0.77 | 0.82 | 0.77 | 0.74 | 0.75 |
| AMBER* | 0.76 | 0.73 | 0.75 | 0.84 | 0.74 | 0.79 | 0.80 | 0.74 | 0.77 |
| CFF91-CM4 | 0.78 | 0.70 | 0.74 | 0.97 | 0.96 | 0.97 | 0.88 | 0.83 | 0.85 |
| MMFF94 | 0.70 | 0.93 | 0.81 | 0.85 | 0.95 | 0.90 | 0.78 | 0.94 | 0.86 |
| AMBER | 1.03 | 0.94 | 0.98 | 0.85 | 0.80 | 0.82 | 0.94 | 0.87 | 0.90 |
| PM3-CM2 | 1.05 | 1.10 | 1.08 | 0.91 | 0.74 | 0.83 | 0.98 | 0.92 | 0.95 |
| PM3 | 1.05 | 1.07 | 1.06 | 0.91 | 0.85 | 0.88 | 0.98 | 0.96 | 0.97 |
| PM3-CM1 | 1.05 | 1.07 | 1.06 | 0.91 | 0.85 | 0.88 | 0.98 | 0.96 | 0.97 |
| TRIPOS-CM4 | 1.19 | 1.43 | 1.31 | 0.65 | 0.69 | 0.67 | 0.92 | 1.06 | 0.99 |
| AM1-CM2 | 1.24 | 1.27 | 1.25 | 0.91 | 0.68 | 0.79 | 1.07 | 0.98 | 1.02 |
| CVFF-CM4 | 1.04 | 1.10 | 1.07 | 1.03 | 1.03 | 1.03 | 1.04 | 1.06 | 1.05 |
| AM1-CM1 | 1.24 | 1.28 | 1.26 | 0.91 | 0.85 | 0.88 | 1.07 | 1.06 | 1.07 |
| AM1 | 1.24 | 1.28 | 1.26 | 0.91 | 0.85 | 0.88 | 1.07 | 1.06 | 1.07 |
| TRIPOS | 1.09 | 1.15 | 1.12 | 1.49 | 1.33 | 1.41 | 1.29 | 1.24 | 1.26 |
| CFF91-HF | 0.85 | 0.94 | 0.90 | 2.02 | 2.30 | 2.16 | 1.44 | 1.62 | 1.53 |
| CVFF-HF | 1.12 | 1.27 | 1.20 | 3.60 | 3.20 | 3.40 | 2.36 | 2.24 | 2.30 |
| PM3-CM3(AMS)[b,a] | 1.05 | ND | ND | 0.91 | ND | ND | 0.98 | ND | ND |
| PDDG/PM3[a] | 1.19 | ND | ND | 0.79 | ND | ND | 0.99 | ND | ND |
| PM3-CM3(GMSP)[b,a] | 1.05 | ND | ND | 0.99 | ND | ND | 1.02 | ND | ND |
| AM1-CM3(GMSP)[b,a] | 1.24 | ND | ND | 0.91 | ND | ND | 1.07 | ND | ND |
| AM1-CM3(AMS)[b,a] | 1.24 | ND | ND | 0.91 | ND | ND | 1.07 | ND | ND |

[a] ND = not determined. [b] AMS = calculated using AMSOL7.1, GMSP = calculated using GAMESPLUSS.

neutral species) and TRIPOS-CM4 all yield about a 2-fold increase in accuracy. Other force fields with CM4 charges in place show little to no difference. Bond angles (Figure 8c) are treated with an accuracy similar to that in the gas phase, with the CVFF-HF and CFF91-HF force fields performing very poorly (MUE $\geq$ 10 degrees). All methods except the TRIPOS, CVFF-CM4, and CFF91-CM4 force fields model bond angles to within $\sim$4.0°, although CVFF-CM4 and CFF91-CM4 again represent a great improvement over their parent force fields. Again, CM4 treatment of the force fields generally leads to an almost 2-fold increase in accuracy for at least one, if not both, of the molecules modeled by each method.

Comparison of Figures 5 and 6 to Figures 7 and 8 shows that the SE-MO methods and force fields are about equally accurate when comparing gas-phase to aqueous-phase results, so no one method in particular stands out as extremely ill-suited to work in either the gaseous or aqueous phase.

**3.5. Overall Geometric Assessment.** To make an overall geometric assessment, we consider the reduced deviances ($D_{y,m}$) in partial charge assignment (Table 3) and geometric modeling (Table 4). Reduced deviance in partial charge shows that seven of the eleven SE-MO methods tested (the exceptions being AM1, AM1-CM1, PM3, and PDDG/PM3) perform better than the average method. In fact, PM3-CM2 is a factor of 2.5 better than average. The only MM methods

that predict partial charge better than average are the AMBER*, MMFF94, and TRIPOS force fields. This assessment, however, also must take into account that the PM3-CM*x* and AM1-CM*x* methods use charge methods that serve as precursors to charge model 4. Since the training sets for the various CM*x* models share some molecules, it is perhaps not surprising that the partial atomic charges of the various CM*x* models show some agreement.

Across the molecules, reduced deviance is fairly consistent when the large errors in CVFF-HF and CFF91-HF for 2,4-DAP and 1*H*-2,4-DAP are taken into account. A slight skewing of the data may be occurring because of the aforementioned error because CVFF-HF and CFF91-HF actually perform as well as the average in partial charge assignment for 2-AMP and 1*H*-2-AMP but perform so poorly in the other cases that their overall reduced deviance is high. However, if we were to carry on with either the CVFF-HF or CFF91-HF force fields by using the original method[98] of partial charge assignment for these force fields, we would be forced to perform electrostatic potential fitting on each molecule we study, and it is known that one of the weaknesses of ChelpG is fitting to the molecular electrostatic potential (MEP) of larger systems, where buried atoms are screened from the points where the MEP is evaluated, and changes of the partial charges on these atoms have only small effects on the MEP. Attempting to fit these charges, then, often produces nonphysical results.[80−83]

Whereas the SE-MO charges and CM*x* charges are different in the gas phase and the aqueous phase, the MM charges (with the exception of those models using CM4 charges) are the same. Table 3 allows for further examination of an issue discussed briefly at the end of section 3.4, namely, the question of whether standard MM charges are more appropriate for the gas phase or for liquid-phase solution. Some force fields are explicit on this issue, for example, OPLS is explicitly named for its use in liquid-phase simulations. Others are implicitly designed for use in liquid phases simply because that is where the greatest number of applications of MM force fields occur. Table 3 shows that MMFF94, AMBER, OPLS2005, CVFF-HF, and CFF91-HF all perform significantly better for partial atomic charges in water than in the gas phase, and AMBER* and TRIPOS are slightly better in the aqueous phase than in the gas phase. The finding that the partial charges are more indicative of the charge distribution in the aqueous phase than the gas phase in all seven cases is quite remarkable and is encouraging.

The reduced deviance in geometric modeling is given in Table 4. Of particular interest are the excellent performance of the MM-CM4 methods and the repair of the CFF91-HF force field. Prior to partial charge replacement, the CFF91-HF force field has a geometric $D_{y,m}$ equal to 1.54. After CM4 treatment, the $D_{y,m}$ for CVFF-CM4 is 0.86, representing a significant increase in the geometric modeling capabilities of the force field by fixing partial charge assignment problems. Other MM-CM4 methods also perform very well, with OPLS2005-CM4, AMBER-CM4, and MMFF94-CM performing 39%, 36%, and 32% better than the average, respectively. As far as non-CM4 treated force fields are concerned, the OPLS2005 force field performs the best ($D_{y,m} = 0.68$), with AMBER*, MMFF94, and AMBER also modeling geometries better than average. The tested SE-MO methods all perform very similarly around a $D_{y,m}$ of ∼1.0.

Another interesting point in Table 4 is the comparison of performance in the gas phase with that in aqueous solution. Of the fourteen MM rows in Table 4, six show smaller errors for gas-phase geometries, and eight show smaller errors for aqueous geometries (of the eight other methods for which such comparison is possible, all show better agreement in the aqueous phase). The MMFF94 method is particularly noteworthy in showing (with either MMFF94 charges or CM4 charges) much better accuracy for gas-phase geometries than for aqueous ones.

On the basis of a combined assessment of partial charge assignment and geometric modeling performance, we find that the AMBER, AMBER*, MMFF94, and OPLS2005 force fields, along with their CM4-treated counterparts, would be suitable for carrying on into a validation step against a larger set of molecules. Each force field has some particular advantages. The AMBER force field has traditionally been highly regarded for its use in modeling biopolymers, and in this case, is used with the incorporated parameters of the General Amber Force Field (GAFF). Atom types in GAFF are designed to be more general than those of traditional AMBER force fields, in an effort to cover a larger portion of organic space. The parametrization of GAFF was developed in an effort to reproduce restrained electrostatic potential (RESP) charges[81,85] at the MP2/6−31G(d) level and reproduce MP2/6−31G(d) and crystallographic geometries.[91] The AMBER* force field has been implemented in *MacroModel* and been modified to better reproduce HF/6−31+G(d) data on peptides, as well as small organic molecules, especially those with nitrogen as a component.[100,101] The MMFF94 force field has been parametrized against a large (∼2800 structures), high-quality ab inito training set for use with both organic molecules and biopolymers, specifically in solution.[102]

**3.6. Comparison of Binding Energies.** We carried out gas-phase binding energy calculations between 2-AMP, 1*H*-2-AMP, 2,4-DAP, and 1*H*-2,4-DAP and water molecules placed at locations where hydrogen bonding has been described in the DHFR-MTX crystal structure (Figure 9). The results of this study are presented in Figure 10, wherein the MUE is the mean unsigned error for a method in binding energy for all neutral molecules, all charged molecules, or both sets together. Compared to the M05-2X benchmark binding energies, all methods except for CVFF-HF and CFF91-HF predict binding energies in both sets of species with an MUE of less than 7.0 kcal/mol. Fourteen of the 16 methods tested predict binding energy for the charged species less accurately than for the neutral species, with some methods, such as AMBER and AMBER*, displaying a 4-fold increase in MUE. Two methods, AM1 and CVFF-CM4, predict binding energy for protonated species more accurately than for neutral species, with MUE differences between the two species of 1.0 and 2.0 kcal/mol, respectively.

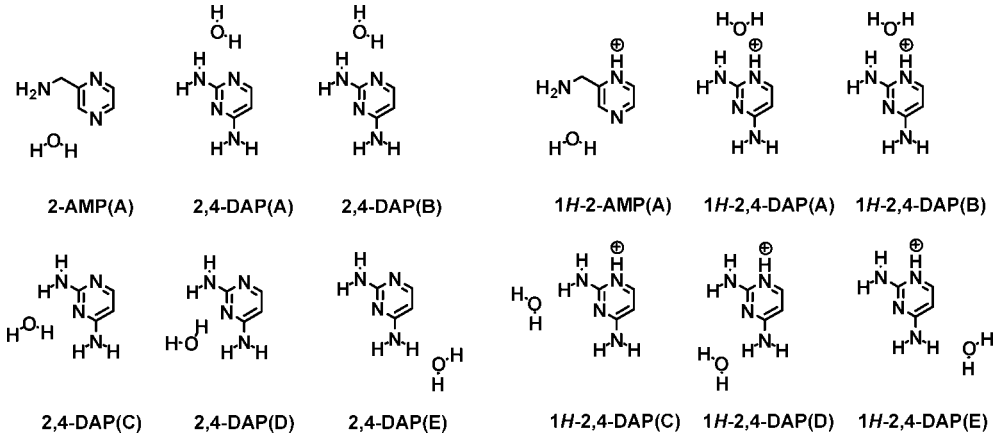



***Figure 9.*** Molecular systems used in binding energy calculations. Position of the water molecule reflects the optimized complex.

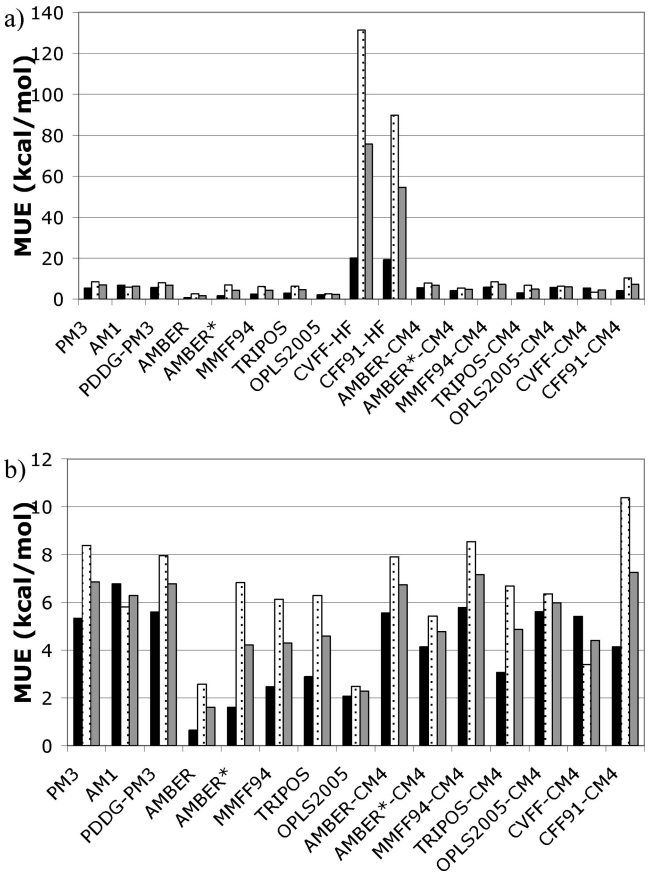



***Figure 10.*** MUE in prediction of binding energy between 2-AMP, 1*H*-2-AMP, 2,4-DAP, and 1*H*-2,4-DAP and water molecules place at hydrogen bonding locations denoted in the DHFR-MTX crystal structure. Panel a includes all methods tested, and panel b omits CVFF-HF and CFF91-HF for easier viewing (black, neutral species; dotted, charged species; gray, both sets of species).

The CVFF-HF and CFF91-HF force fields perform quite poorly, likely, because of the necessary substitution of CHELPG charges for the force fields' default charges. Notably, the substitution of CM4 charges into the CVFF and CFF91 force fields again improves their modeling accuracy; however, this is not the case with the remaining MM force fields. Out of all the methods tested, the most accurate are the OPLS2005 and AMBER force fields, which predict binding energy for both sets of molecules to an MUE of 2.28 and 1.61 kcal/mol, respectively. In addition to its excellent accuracy, the OPLS2005 force field shows consistency between the neutral and charged species, calculating binding energy to an MUE of 2.07 and 2.49, respectively. The TRIPOS, MMFF94, and AMBER* force fields all predict binding energies about 2-fold less accurately, with MUEs of 4.59, 4.29, and 4.22 kcal/mol, respectively.

***Table 5.*** Reduced Deviance ($D_{y,m}$) in Partial Charge for Force Field Validation For All Neutral and Charged Species in Figure *3*

| method | gas phase | aqueous phase | both phases |
|---|---|---|---|
| AMBER* | 0.68 | 0.69 | 0.68 |
| MMFF94 | 1.04 | 1.04 | 1.04 |
| AMBER | 1.05 | 1.07 | 1.06 |
| OPLS2005 | 1.18 | 1.19 | 1.18 |

***Table 6.*** Reduced Deviance ($D_{y,m}$) in Combined Geometry for Force Field Validation For All Neutral and Charged Species in Figure *3*

| method | gas phase | aqueous phase | both phases |
|---|---|---|---|
| MMFF94-CM4 | 0.73 | 0.79 | 0.76 |
| MMFF94 | 0.80 | 0.85 | 0.82 |
| OPLS2005-CM4 | 0.82 | 0.91 | 0.86 |
| OPL2005 | 0.95 | 0.88 | 0.92 |
| AMBER-CM4 | 0.90 | 0.93 | 0.92 |
| AMBER | 1.03 | 0.97 | 1.00 |
| AMBER* | 1.38 | 1.32 | 1.34 |
| AMBER*-CM4 | 1.39 | 1.35 | 1.37 |

**3.7. Validation of the AMBER, AMBER*, MMFF94, and OPLS2005 Force Fields.** Tables 5 and 6 summarize our validation of the AMBER, AMBER*, MMFF94, and OPLS2005 force fields (as well as their MM-CM4 counterparts) in the gaseous and aqueous phases against both the molecules already included in the study, as well as an additional set of pharmacophore-containing molecules and their cations shown in Figure 3. In contrast to its relatively poor performance assigning partial charges to 2-AMP, 1*H*-2-AMP, 2,4-DAP, and 1*H*-2,4,-DAP, the AMBER force field performs better in this case than OPLS2005, although the degree to which this increased performance is relevant is

***Table 7.*** Partial Charge Distribution of Gas-Phase Neutral and Cationic MTX as Calculated by M05-2X/CM4 and MMFF94

| | MTX | | | MTX+ | | | | MTX | | | MTX+ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| atom | CM4 | MMFF94 | residual | CM4 | MMFF94 | residual | atom | CM4 | MMFF94 | residual | CM4 | MMFF94 | residual |
| C1 | 0.44 | 0.72 | 0.28 | 0.51 | 0.77 | 0.26 | C29 | −0.12 | 0.09 | 0.21 | −0.08 | 0.09 | 0.17 |
| C2 | 0.29 | 0.41 | 0.12 | 0.27 | 0.41 | 0.14 | H30 | 0.07 | 0.15 | 0.08 | 0.07 | 0.15 | 0.08 |
| C3 | 0.07 | 0.31 | 0.24 | 0.09 | 0.31 | 0.22 | H31 | 0.09 | 0.15 | 0.06 | 0.10 | 0.15 | 0.05 |
| C4 | 0.29 | 0.62 | 0.33 | 0.35 | 0.67 | 0.32 | C32 | 0.00 | 0.37 | 0.37 | −0.01 | 0.37 | 0.38 |
| C5 | 0.14 | 0.17 | 0.03 | 0.18 | 0.17 | 0.02 | H33 | 0.06 | 0.00 | 0.06 | 0.07 | 0.00 | 0.07 |
| C6 | 0.14 | 0.16 | 0.02 | 0.21 | 0.16 | 0.05 | H34 | 0.06 | 0.00 | 0.06 | 0.06 | 0.00 | 0.06 |
| H7 | 0.07 | 0.15 | 0.08 | 0.09 | 0.15 | 0.06 | H35 | 0.06 | 0.00 | 0.06 | 0.06 | 0.00 | 0.06 |
| N8 | −0.59 | −0.90 | 0.31 | −0.52 | −0.90 | 0.38 | C36 | 0.36 | 0.54 | 0.19 | 0.33 | 0.54 | 0.22 |
| H9 | 0.31 | 0.40 | 0.09 | 0.33 | 0.40 | 0.07 | O37 | −0.38 | −0.57 | 0.19 | −0.36 | −0.57 | 0.22 |
| H10 | 0.32 | 0.40 | 0.08 | 0.34 | 0.40 | 0.06 | N38 | −0.43 | −0.73 | 0.30 | −0.43 | −0.73 | 0.30 |
| N11 | −0.61 | −0.90 | 0.29 | −0.55 | −0.90 | 0.35 | H39 | 0.26 | 0.37 | 0.11 | 0.26 | 0.37 | 0.11 |
| H12 | 0.32 | 0.40 | 0.08 | 0.35 | 0.40 | 0.05 | C40 | 0.11 | 0.36 | 0.26 | 0.11 | 0.36 | 0.25 |
| H13 | 0.32 | 0.40 | 0.08 | 0.34 | 0.40 | 0.07 | H41 | 0.07 | 0.00 | 0.07 | 0.07 | 0.00 | 0.07 |
| N14 | −0.44 | −0.62 | 0.18 | −0.41 | −0.62 | 0.21 | C42 | −0.08 | 0.00 | 0.08 | −0.07 | 0.00 | 0.07 |
| N15 | −0.44 | −0.62 | 0.18 | −0.40 | −0.18 | 0.22 | H43 | 0.07 | 0.00 | 0.07 | 0.07 | 0.00 | 0.07 |
| N16 | −0.32 | −0.62 | 0.30 | −0.31 | −0.62 | 0.31 | H44 | 0.07 | 0.00 | 0.07 | 0.07 | 0.00 | 0.07 |
| N17 | −0.29 | −0.62 | 0.33 | −0.28 | −0.62 | 0.34 | C45 | −0.10 | 0.06 | 0.16 | −0.10 | 0.06 | 0.16 |
| C18 | 0.03 | 0.51 | 0.49 | 0.03 | 0.51 | 0.48 | H46 | 0.09 | 0.00 | 0.09 | 0.09 | 0.00 | 0.09 |
| H19 | 0.05 | 0.00 | 0.05 | 0.06 | 0.00 | 0.06 | H47 | 0.09 | 0.00 | 0.09 | 0.09 | 0.00 | 0.09 |
| H20 | 0.06 | 0.00 | 0.06 | 0.07 | 0.00 | 0.07 | C48 | 0.29 | 0.66 | 0.37 | 0.29 | 0.66 | 0.37 |
| N21 | −0.33 | −0.84 | 0.51 | −0.34 | −0.84 | 0.50 | C49 | 0.29 | 0.66 | 0.37 | 0.29 | 0.66 | 0.37 |
| C22 | 0.18 | 0.10 | 0.08 | 0.16 | 0.10 | 0.06 | O50 | −0.35 | −0.57 | 0.22 | −0.35 | −0.57 | 0.22 |
| C23 | −0.11 | −0.15 | 0.04 | −0.12 | −0.15 | 0.03 | O51 | −0.34 | −0.57 | 0.23 | −0.33 | −0.57 | 0.24 |
| C24 | −0.11 | −0.15 | 0.04 | −0.10 | −0.15 | 0.05 | O52 | −0.33 | −0.65 | 0.33 | −0.32 | −0.65 | 0.33 |
| C25 | −0.05 | −0.15 | 0.10 | −0.04 | −0.15 | 0.11 | H53 | 0.32 | 0.50 | 0.18 | 0.33 | 0.50 | 0.18 |
| H26 | 0.06 | 0.15 | 0.09 | 0.05 | 0.15 | 0.10 | O54 | −0.36 | −0.65 | 0.29 | −0.36 | −0.65 | 0.29 |
| C27 | −0.02 | −0.15 | 0.13 | −0.01 | −0.15 | 0.14 | H55 | 0.32 | 0.5 | 0.18 | 0.32 | 0.50 | 0.18 |
| H28 | 0.07 | 0.15 | 0.08 | 0.07 | 0.15 | 0.08 | H56 | N/A[a] | N/A[a] | N/A | 0.33 | 0.46 | 0.13 |
| | | | | | | | MUE | | | 0.17 | | | 0.17 |

[a] Atom not present in the protonated molecule.

debatable. All three force fields excel at assigning partial charge to different types of molecules (data not shown), so it is not clear which is consistently more accurate.

Table 6 summarizes the reduced deviance for each method in modeling molecular geometries. Whereas Figure 4 showed large increases in performance when CM4 charges are substituted for the original charges of several MM methods, here we see a much more modest effect, although incorporating CM4 charges does improve the geometric modeling capabilities of three of the force fields. The MMFF94 and MMFF94-CM4 force fields are clearly superior when modeling the molecules included in the validation. With this in mind, we consider the MMFF94-CM4 force field as the most accurate force field for our system. The performance of the unmodified MMFF94 force field is also excellent, and it merits consideration because it is already well defined.

**3.8. Comparison of CM4 and MMFF94 Charge Distribution Calculated for Methotrexate's Neutral and Cationic Forms.** Table 7 contains the partial charge distribution of the gas-phase optimized structure of MTX (neutral and cationic) as calculated by M05-2X/6−31+G(d,p)/CM4 and MMFF94. The numbering systems for the molecules are given in the Supporting Information. Both the ionic and nonionic species' charge distributions are calculated to a MUE of 0.17 by MMFF94, validating this force field's good performance for such a large molecule.

## 4. Conclusions

We have studied 30 systems, each of which is examined with up to 31 methods in one or two phases (gaseous and aqueous). We found that the M05-2X density functional with the 6−31+G(d,p) basis set yields geometries very close to those obtained with coupled-cluster calculations. M05-2X is therefore useful in obtaining benchmark values for larger molecules involved in drug design.

The assignment of appropriate partial atomic charges is critical to accurate modeling of molecules by molecular mechanics. We found that substitution of CM4 charges for the original charge parameters of a given MM model improved the geometric accuracy of all seven force fields for which this substitution was tested, with some errors in geometry decreasing by factors of 3.5 and 4 in the two most dramatic cases. With the improved partial charge assignment, four of the MM methods come very close to reproducing coupled-cluster calculations.

Although the substitution of CM4 charges into our MM force fields improved geometric accuracy, it had the opposite effect on prediction of binding energies for the majority of the force fields tested. Thus, the overall improvement of a MM force field does not lie solely in the improvement of one aspect of that method, and charge substitution should be used with care.

We have found that the MMFF94-CM4 force field, in which CM4 charges are substituted for the MMFF94 default charges, yields the most accurate geometries for

representative fragments of methotrexate, as well for as an additional set of druglike molecules. Furthermore, the MMFF94 force field without charge substitution exhibits the second best geometric performance among the sixteen methods tested. However, in our binding energy studies, we find that excellent performance modeling geometries and charge distributions does not necessarily correlate directly to the prediction of energetics. Therefore, when the combined charge distribution, geometric, and energetic results are taken into account, we consider the MMFF94, AMBER/GAFF, AMBER*, and OPLS2005 force fields to be the most accurate and economical methods available for modeling small molecules containing nitrogen heterocycles and exocyclic amines. We expect these methods to be suitable for use in modeling more general nitrogen-containing small molecules as well as larger systems including the bis-methotrexate chemical inducer of dimerization, the protein−ligand complex, and the residues contained at the DHFR−DHFR interface.

**Acknowledgment.** We would like to thank David Ferguson, Casey Kelly, Benjamin Lynch, Nate Schultz, and Yuk Sham for their assistance during this study. B.R.W. was supported by a Chemistry and Biology NIH traineeship (T32 GM008700). This work was supported in part by the University of Minnesota Nanobiotechnology Initiative, the University of Minnesota Supercomputing Institute, and NSF grant CHE-0704974.

**Supporting Information Available:** Numbering system of 2-AMP, 1*H*-2-AMP, 2,4-DAP, 1*H*-2,4-DAP, MTX, and the MTX cation (Figures S1−S3), benchmark calculations of partial charge, bond length, and bond angle on 2-AMP, 1*H*-2-AMP, 2,4-DAP, and 1*H*-2,4-DAP utilizing CCSD/6−31+G(d,p) in the gas phase, as well as M05-2X/6−31+G(d,p) in both the gas and aqueous phases (Tables S1−S12), and cartesian coordinates of CCSD and M05-2X optimizations (Tables S13−S15). This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Gohlke, H.; Klebe, G. *Agnew. Chem. Int. Ed.* **2002**, *41*, 2644.

(2) Hessler, G.; Klabunde, T. *ChemBioChem* **2002**, *3*, 928.

(3) Lugovsky, A. A.; Degterev, A. I.; Fahmy, A. F.; Zhou, P.; Gross, J. D. *J. Am. Chem. Soc.* **2002**, *124*, 1234.

(4) Perez, J. J.; Concho, F.; Llorens, O. *Curr. Med. Chem.* **2002**, *24*, 2209.

(5) Rao, G. S.; Bhatnagar, S.; Ahuja, V. *J. Biomol. Struct. Dyn.* **2002**, *20*, 31.

(6) Shahripour, A. B.; Plummer, M. S.; Lunny, E. A.; Albrecht, H. P.; Hays, S. J. *Bioorg. Med. Chem.* **2002**, *10*, 31.

(7) Anderson, A. *Chem. Biol.* **2003**, *10*, 787.

(8) Berlicki, L.; Kafarski, P. *Curr. Org. Chem.* **2005**, *9*, 1829.

(9) Ikejiri, M.; Bernardo, M. M.; Meroueh, S. O.; Brown, S.; Chang, M. *J. Org. Chem.* **2005**, *70*, 5709.

(10) Zhang, D. A. W.; Zhang, J. Z. H. *Int. J. Quantum Chem.* **2005**, *103*, 246.

(11) Armstrong, K. A.; Tidor, B.; Cheng, A. C. *J. Med. Chem.* **2006**, *49*, 2470.

(12) Ortiz, A. R.; Gomez-Puentas, P.; Leo-Macias, A.; Lopez-Romero, P.; Lopez-Vinas, E. *Curr. Top. Med. Chem.* **2006**, *6*, 41.

(13) Rao, G. S.; Ramachandran, M. V.; Bajaj, J. S. *J. Biomol. Struct. Dyn.* **2006**, *23*, 377.

(14) Ragno, R.; Simeoni, S.; Castellano, S.; Vicidomini, C.; Mai, A. *J. Med. Chem.* **2007**, *50*, 1241.

(15) Strockbine, B.; Rizzo, R. C. *Proteins: Struct., Funct., Genet.* **2007**, *67*, 630.

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

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

(18) Cornell, W. D.; Cieplak, P. *J. Am. Chem. Soc.* **1995**, 5179.

(19) Petersson, I.; Liljefors, T. *Rev. Comput. Chem.* **1996**, 9.

(20) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D. *J. Phys. Chem. B* **1998**, *102*, 3586.

(21) Boyd, D. B.; Snoddy, J. D.; Lin, H. S. *J. Comput. Chem.* **1991**, *12*, 635.

(22) Jalaie, M.; Lipkowitz, K. *Rev. Comput. Chem.* **1999**, 14.

(23) Khandelwal, A.; Lukacova, V.; Comez, D.; Kroll, D. M.; Raha, S. *J. Med. Chem.* **2005**, *48*, 5437.

(24) Kohn, W.; Becke, A. D.; Parr, R. G. *J. Phys. Chem.* **1996**, *100*, 12974.

(25) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968.

(26) Mulholland, A. J. *Theor. Comput. Chem.* **2001**, *9*, 597.

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

(28) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185.

(29) Pople, J. A.; Beveridge, D. L. *Approximate Molecular Orbital Methods*; McGraw-Hill: New York, 1970.

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

(31) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.

(32) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221.

(33) Richards, N. J. *Molecular Orbital Calculations for Biological Systems*; Oxford University Press: New York, 1998.

(34) McKercher, S. R.; Lombardo, C. R.; Bobkov, A.; Jia, X.; Assa-Muut, N. *Proc. Natl. Acad. Sci. U.S.A* **2003**, *100*, 511.

(35) Elcock, A. H.; Sept, D.; McCammon, J. A. *J. Phys. Chem. B* **2001**, *105*, 1504.

(36) Prazulj, N.; Wigle, D. A.; Jurisica, I. *Bioinformatics* **2004**, *20*, 340.

(37) Gao, Y.; Wang, R.; Lai, L. *J. Mol. Model.* **2004**, *10*, 44.

(38) Gohlke, H.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 238.

(39) Zhao, H. X. *Curr. Med. Chem.* **2004**, *11*, 539.

(40) Hou, T.; Chen, K.; McLaughlin, W. A.; Lu, B.; Wang, W. *PLoS Comp. Biol.* **2006**, *2*, 46.

(41) Ababou, A.; van der Vaant, A.; Gogonea, V.; Merz, K. M. J. *Biophys. Chem.* **2007**, *125*, 221.

(42) Aslan, F. M.; Yu, Y.; Vajda, S.; Mohr, S. C.; Cantor, C. R. *J. Biotechnol.* **2007**, *128*, 213.

(43) Carlson, J. C. T.; Kanter, A.; Thudappathy, G. R.; Cody, V.; Pineda, P. E. *J. Am. Chem. Soc.* **2003**, *125*, 1501.

(44) Sutton, P. A.; Cody, V.; Smith, D. *J. Am. Chem. Soc.* **1986**, *108*, 4155.

(45) Bolin, J. T.; Filman, D. J.; Matthews, D. A.; Hamlin, R. C.; Kraut, J. *J. Biol. Chem.* **1982**, *257*, 13650.

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

(47) Cizek, J. *J. Chem. Phys.* **1966**, *45*, 4256.

(48) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910.

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

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

(51) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 1133.

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

(53) Dauber-Osguthorpe, P.; Roberts, V. A.; Osguthorpe, D. J.; Wolff, J.; Genest, M. *Proteins: Struct., Funct., Genet.* **1988**, *4*, 31.

(54) Clark, M.; Cramer III, R. D.; van Opdenhosch, N. *J. Comput. Chem.* **1989**, *10*, 982.

(55) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. J. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

(56) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490.

(57) Damm, W.; Frontera, A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Comput. Chem.* **1997**, *18*, 1955.

(58) Li, J.; Zhu, J.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1998**, *102*, 1820.

(59) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **2003**, *24*, 1291.

(60) Chambers, C. C.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *1000*, 16385.

(61) Zhu, T.; Li, J.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1998**, 109.

(62) Helgaker, T.; Gauss, J.; Jorgensen, P.; Olsen, J. *J. Chem. Phys.* **1997**, *106*, 6430.

(63) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Phys. A* **2003**, *107*, 1384.

(64) Wong, M. W.; Radom, L. *J. Phys. Chem. A* **1998**, *102*, 2237.

(65) Byrd, E. F. C.; Sherrill, C. D.; Head-Gordon, M. *J. Phys. Chem. A* **2001**, *105*, 9736.

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

(67) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040.

(68) Perdew, J. P.; Burke, K.; Wang, Y. *Phys. Rev. B* **1996**, *54*, 16533.

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

(70) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908.

(71) Zhao, Y.; Gonzalez-Garcia, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012.

(72) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.

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

(74) Zhao, Y.; Truhlar, D. G. *MN-GFM Minnesota Gaussian Functional Module*, version 2.0.1; University of Minnesota: Minneapolis, MN, 2006.

(75) Mulliken, R. S. *J. Chem. Phys.* **1935**, *3*, 564.

(76) Lowdin, P. O. *J. Chem. Phys.* **1950**, *18*, 365.

(77) Mulliken, R. S. *J. Chem. Phys.* **1962**, *36*, 3428.

(78) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comput. Chem.* **1986**, *7*, 230.

(79) Besler, B. H.; Merz, K. M.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431.

(80) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361.

(81) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269.

(82) Francl, M. M.; Carey, C.; Chirlian, L. E.; Gange, D. M. *J. Comput. Chem.* **1996**, *17*, 367.

(83) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129.

(84) Chirlian, L. E.; Francl, M. M. *J. Comput. Chem.* **1987**, *1987*, 894.

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

(86) Hawkins, G. D.; Giesen, D. J.; Lynch, G. C.; Chambers, C. C.; Rossi, I.; Storer, J. W.; Li, J.; Zhu, T.; Thompson, J. D.; Winget, P.; Lynch, B. J.; Rinaldi, D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *AMSOL*, version 7.1; University of Minnesota: Minneapolis, MN, 2004.

(87) Chamberlin, A. C.; Pu, J.; Kelly, C. P.; Thompson, J. D.; Xidos, J. D. *GAMESSPLUS*, version 4.7; University of Minnesota: Minneapolis, MN, 2005. Based on the General Atomic and Molecular Electronic Structure System (GAMESS) as described in 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.; Dupius, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.

(88) Stewart, J. J. P.; Zheng, J.; Rossi, I.; Hu, W.; Lynch, G. C.; Liu, Y.; Chuang, Y.; Pu, J.; Li, J.; Cramer, C. J.; Fast, P. L.; Truhlar, D. G. *MOPAC*,version 5.011mn; University of Minnesota: Minneapolis, MN, 2006.

(89) Maple, J. R.; Hwang, M. J.; Stockfish, T. P.; Dinur, U.; Waldman, M. *J. Comput. Chem.* **1994**, *15*, 162.

(90) Duan, Y.; Wu, C.; Choudhury, S.; Lee, M. C.; Xiong, G. *J. Comput. Chem.* **2003**, *24*, 1999.

(91) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157.

(92) 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.; Morgan, 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.

(93) Winget, P.; Thompson, J. D.; Xidos, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2002**, 106.

(94) Brom, J. M.; Schmitz, B. J.; Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, 107.

(95) Kalinowski, J. A.; Lesyng, B.; Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, 108.

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

(97) Momany, F. J. *J. Phys. Chem.* **1978**, *85*, 592.

(98) Dinur, U.; Hagler, A. T. *J. Am. Chem. Soc.* **1989**, *111*, 5149.

(99) Hobza, P.; Kabelec, M.; Sponer, J.; Mejzlik, P.; Vondrasek, J. *J. Comput. Chem.* **1996**, *118*, 1136.

(100) McDonald, D. Q.; Still, W. C. *Tetrahedron Lett.* **1992**, *33*, 7747.

(101) Chakravorty, S.; Reynolds, C. H. *J. Mol. Graphics Modell.* **1999**, *17*, 315.

(102) Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 730.

CT8000766

JCTC Journal of Chemical Theory and Computation

# Accurate Calculations of Binding, Folding, and Transfer Free Energies by a Scaled Generalized Born Method

Harianto Tjong and Huan-Xiang Zhou*

*Department of Physics and Institute of Molecular Biophysics, Florida State University, Tallahassee, Florida 32306*

Received May 14, 2008

**Abstract:** The Poisson−Boltzmann (PB) equation is widely used for modeling solvation effects. The computational cost of PB has restricted its applications largely to single-conformation calculations. The generalized Born (GB) model provides an approximation at substantially reduced cost. Currently the best GB methods reproduce PB results for electrostatic solvation energies with errors at ∼5 kcal/mol. When two proteins form a complex, the net electrostatic contributions to the binding free energy are typically of the order of 5 to 10 kcal/mol. Similarly, the net contributions of individual residues to protein folding free energy are <5 kcal/mol. Clearly in these applications the accuracy of current GB methods is insufficient. Here we present a simple scaling scheme that allows our GB method, GBr$^6$, to reproduce PB results for binding, folding, and transfer free energies with high accuracy. From an ensemble of conformations sampled from molecular dynamics simulations, five were judiciously selected for PB calculations. These PB results were used for scaling GBr$^6$. Tests on the binding free energies of the barnase-barstar, GTPase-WASp, and U1A-U1hpII complexes and on the folding free energy of FKBP show that the effects of point mutations calculated by scaled GBr$^6$ are accurate to within 0.3 kcal/mol of PB results. Similar accuracy was also achieved for the free energies of transfer for ribonuclease Sa and insulin from the crystalline phase to the solution phase at various pHs. This method makes it possible to thoroughly sample the transient-complex ensemble in predicting protein binding rate constants and to incorporate conformational sampling in electrostatic modeling (such as done in the MM-GBSA approach) without loss of accuracy.

## I. Introduction

Electrostatic interactions make important contributions to fundamental properties such as protein binding free energy, protein folding stability, and protein solubility. Point mutations and variations of salt concentration and pH are often used to probe such contributions. Computational methods that reliably predict those energetic effects are highly desirable, both for elucidating the underlying physical principles and for protein design. Developing such methods is a formidable task, since the free energy of binding, for example, is a small difference between two large quantities, namely the free energy due to interactions within the complex in the solvent environment and the solvation energy of the subunits in the unbound state. The effects of point mutations on the free energy of binding or folding are even smaller. The Poisson−Boltzmann (PB) equation[1−8] has found great success in modeling electrostatic contributions to mutational effects on protein binding free energy[9−12] and protein folding stability.[13−15] We have also been able to use the PB equation to model the effects of salt and pH on protein solubility.[16,17] However, the computational cost of solving the PB equation presents a major stumbling block.

There have been some efforts devoted to the development of fast PB methods.[4,6] An alternative that holds great promise is the generalized Born (GB) model,[18] which approximates the PB equation at substantially reduced computational cost. Considerable efforts have been invested in developing GB methods that would hold down the computational cost but have great accuracy in reproducing the PB results.[19−26]

* Corresponding author phone: (850)645-1336; fax: (850)644-7244; e-mail: zhou@sb.fsu.edu.

10.1021/ct8001656 CCC: $40.75 
Published on Web 09/06/2008

The reduced computational cost of the GB model opens new avenues for more realistic modeling of electrostatic effects, such as the inclusion of protein conformational sampling.[27] However, the small magnitudes of electrostatic contributions to binding and folding free energies pose a challenging demand on the accuracy of calculations. To date, the errors of the best GB methods in reproducing PB solvation energies are about 0.5% in relative terms or roughly several kcal/mol in absolute terms.[26,28] The scatter of these errors among different proteins also appears to be random.

The electrostatic contributions to protein−protein binding free energies typically have an order of magnitude of 10 kcal/mol.[9,10] Therefore even the currently most accurate GB methods (when benchmarked against PB results) may be inadequate for predicting binding free energy. One idea for improving accuracy is to reparameterize GB methods against PB results. One such attempt was made in calculating protein−ligand binding free energy,[29] using PB results for a subset of ligands as training data, but the error (as measured against PB results) in that work was still relatively large, about 5 kcal/mol. In fact, our recently developed GB method, GBr$^6$, has even smaller errors on protein−protein binding free energies (∼1 kcal/mol for protein complexes studied here) without any reparameterization.[26]

Here we take a different approach to GB reparameterization. The aim of the reparameterization is to allow for conformational sampling in electrostatic modeling. For each protein or protein complex, we generate an ensemble of conformations from molecular dynamics simulations. The raw GB result for the solvation energy of each conformation in the ensemble is calculated, and five conformations are selected to be representative of the variations in solvation energy of the ensemble. PB results for the solvation energies of the five representative conformations are then calculated and used for scaling the corresponding raw GB results. The scaling factor is finally applied to the rest of the conformation ensemble. Accuracy of the scaling method is assessed by comparing the scaled GB results and PB results for the whole ensemble. While the scaling method is applicable to any GB method, we report here results for GBr$^6$. We refer to the reincarnation scaled GBr$^6$ or sGBr$^6$ in short. The promise of reparameterizing GB against PB results calculated for a small subset of conformations was demonstrated in an early study.[30]

We test scaled GBr$^6$ on a number of applications. Mutational and salt effects on the binding free energies are calculated for four protein−protein and protein-RNA complexes. To assess the accuracy of scaled GBr$^6$ for calculating mutational effects on folding free energy, 26 mutants of the FK506-binding protein (FKBP) are studied. In addition, scaled GBr$^6$ is used to calculate the pH dependence of solubility of two proteins, ribonuclease Sa and insulin.

As another important application, we use scaled GBr$^6$ to calculate the electrostatic interaction energy of the transient complex along the protein−protein association pathway, which predicts the electrostatic enhancement of the protein binding rate.[31−34] The transient complex consists of an ensemble of configurations, each of which presents a slightly different relative separation and relative orientation between the two subunits in a complex. Four pairs of binding proteins are studied.

In all these applications, we show that the relevant small differences in electrostatic solvation energy obtained by scaled GBr$^6$ are accurate to within 0.3 kcal/mol of the corresponding PB results. These diverse applications illustrate the wide utility of our GB scaling approach.

## II. Theory

In previous studies, we have described the theoretical models for calculating electrostatic contributions to free energies of protein binding[9,10] and folding,[13,14] for calculating the electrostatic rate enhancement of protein binding,[31−34] and for calculating the pH dependence of protein solubility.[17] Here we give brief outlines of these models.

**2.1. Binding Free Energy.** When two proteins, A and B, bind to form a complex C, the electrostatic contribution to the binding free energy is calculated as

$$\Delta G_b = G_{el}(C) - G_{el}(A) - G_{el}(B) \quad (1)$$

where $G_{el}(X)$, X = A, B, or C, is the electrostatic free energy of molecule X. $G_{el}(X)$ can be decomposed into a Coulomb term and a solvation term:

$$G_{el}(X) = G_{Coul}(X) + G_{solv}(X) \quad (2)$$

By this decomposition, $\Delta G_b$ has a Coulombic component and a solvation component. These will be denoted as $\Delta G_{b;Coul}$ and $\Delta G_{b;solv}$, respectively. When a point mutation is introduced, $\Delta G_b$ can be calculated for the wild-type (wt) pair of proteins and for the mutant (mt) pair. The change in $\Delta G_b$ by the mutation is

$$\Delta\Delta G_b = \Delta G_b(mt) - \Delta G_b(wt) \quad (3)$$

**2.2. Folding Free Energy.** For protein folding, we are interested in mutational effects on the electrostatic contribution to the folding free energy. It is assumed that, in the unfolded state, individual residues do not interact, and hence the contributions of residues other than the one under mutation are the same in the wild-type protein and the mutant. (Residual charge−charge interactions in the unfolded state have been modeled previously.[35]) Neglecting the electrostatic free energy of the "other" residues in the unfoled state (which does not affect the change in folding free energy by the mutation), the electrostatic contribution to the folding free energy is

$$\Delta G_f = G_{el}(protein) - G_{el}(residue) \quad (4)$$

where the two terms on the right-hand side represent the electrostatic free energies of the protein in the folded state and the residue under mutation, respectively. The solvation component of $\Delta G_f$ will be denoted as $\Delta G_{f;solv}$. The mutational effect on $\Delta G_f$ is given by

$$\Delta\Delta G_f = \Delta G_f(mt) - \Delta G_f(wt) \quad (5)$$

Notice that even $\Delta G_f(wt)$ is mutant-specific; we will come back to this point in subsection 3.4.

**2.3. pH Dependence of Solubility.** The effect of pH on protein solubility can be calculated from the pH dependence

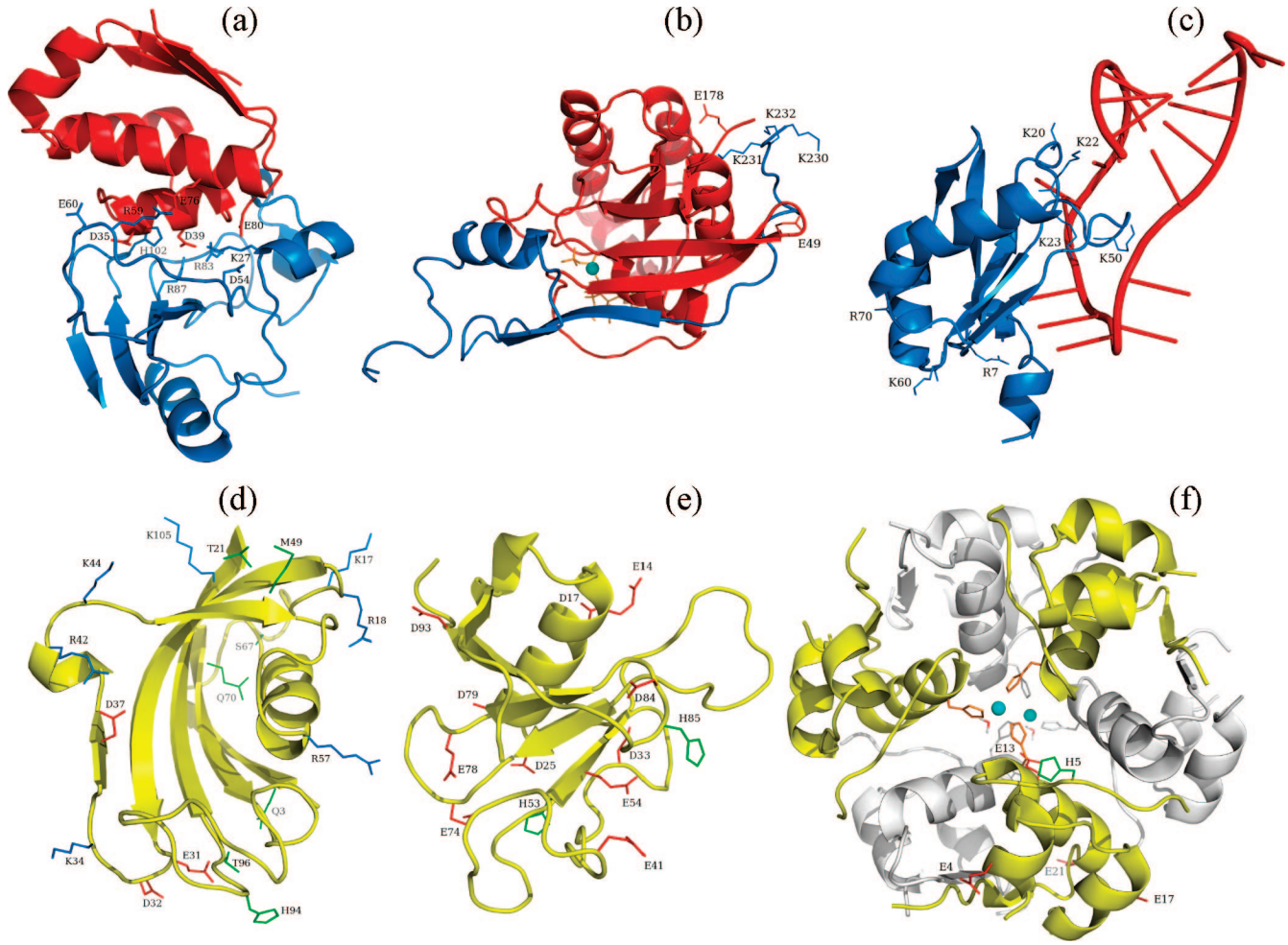



***Figure 1.*** Systems studied in the present work. (a)−(c) Complexes of barnase and barstar, WASp and Cdc42, and U1A and U1SLII. The first and second subunits are shown in blue and red, respectively. (d) FKBP. (e) Ribonuclease Sa. (f) Insulin hexamer. Mutated residues in (a)−(d) and titrated residues in (e) and (f) are labeled.

of the transfer free energy of a protein from the condensed phase to the solution phase. Both the solution phase and the condensed phase are modeled as continuum dielectrics, with the dielectric constants taking a value appropriate for water (equal to 78.5 at room temperature) for the former and a value (∼55) intermediate between those for water and for the protein solute. The electrostatic component of the transfer free energy, which is what is relevant for determining the pH dependence of solubility, is

$$\Delta G_t = G_{solv}(\varepsilon_s = 78.5) - G_{solv}(\varepsilon_s = 55) \quad (6)$$

where $G_{solv}(\epsilon_s)$ is the electrostatic solvation energy of the solute protein in a continuum solvent with dielectric constant $\epsilon_s$. This result uses the fact that the Coulomb term is the same in both phases. $G_{solv}(\epsilon_s)$ at each pH was calculated by averaging over conformations sampled from constant-pH molecular dynamics simulations.

**2.4. Electrostatic Rate Enhancement.** According to the transient-complex theory,[31,33,34] the rate constant for the association of two proteins under diffusion control can be predicted as

$$k_a = k_{a0} e^{-\Delta G_{el}^*/k_B T} \quad (7)$$

where $k_{a0}$ is the basal rate constant, and $\Delta G_{el}^*$ is the electrostatic interaction energy of the proteins in the transient-complex ensemble. $\Delta G_{el}^*$ is calculated as the average over configurations representing the transient complex. For each configuration, the electrostatic interaction energy is as defined by eq 1, except that C now represents a transient-complex configuration.

## III. Computation Details

**3.1. Generation of Conformational Ensembles.** For calculating electrostatic contributions to protein binding and folding free energies by scaled GBr[6], conformations were generated by explicit-solvent molecular dynamics (MD) simulations. Four protein−protein and protein-RNA complexes were studied: an enzyme−inhibitor complex formed by barnase and barstar; complexes formed by the Wiskott-Aldrich Syndrome protein (WASp) with two homologous Rho GTPases, Cdc42 and TC10; and a protein-RNA complex formed by the U1A protein and stem/loop II of the U1 small nuclear RNA (U1SLII). The first and last complexes were previously studied by PB calculations.[9,11] The two GTPase-WASp complexes have been studied experimentally.[36] These and other systems studied here are shown in Figure 1.

The simulation of the barnase-barstar complex was carried out as follows. Starting from the X-ray structure of the complex (Protein Data Bank entry 1brs[37]), hydrogens, four $Na^+$ ions as neutralizing counterions, and 9558 TIP3P water

molecules were added by the LEAP program in the Amber package.[38] The side chains of arginine and lysine residues were positively charged, side chains of aspartate and glutamate residues were negatively charged, and side chains of histidine residues were neutral (appropriate for a nominal pH of 7). The force field was ff99SB.[39] The solvent (water molecules plus counterions) was relaxed first by 200 steps of energy minimization and then by 100 ps of MD simulation at a constant pressure MD, all the while the protein complex was fixed. Next, the whole system was energy minimized with decreasing harmonic constraints applied to the protein complex, from 50 kcal/mol/Å$^2$ to 0, for a total of 2500 steps. The cutoff for the nonbonded interactions was 9 Å, and the particle mesh Ewald method was used to treat long-range electrostatic interactions. The whole system was then heated for 40 ps at constant volume to a final temperature of 298 K. Finally the simulation was continued at constant pressure and temperature. Bond lengths involving hydrogen atoms were restrained by the SHAKE algorithm[40] throughout the simulation, allowing a time step of 2 fs. The first 2 ns of the constant pressure and temperature simulation was discarded; thereafter conformations were sampled at every 10 ps, and a total of 548 conformations were collected.

For the Cdc42-WASp complex, an NMR structure with 20 models was available (Protein Data Bank entry 1cee).[41] We randomly picked five (models 1, 5, 9, 13, and 17) of the 20 to generate conformational ensembles. No structure for the TC10-WASp complex was available. We modeled its structure by aligning the structure of unbound TC10 (Protein Data Bank entry 2atx[36]) to Cdc42 in the five NMR models of the Cdc42-WASp complex. From these 10 starting structures (five for each complex), MD simulations were carried out following the protocol for the barnase-barstar complex. For each of the 10 simulations, 100 conformations were sampled at every 10 ps after discarding the first 1 ns of the constant pressure and temperature simulation.

The starting structure for the U1A-U1SLII complex was as prepared in our previous study.[11] 100 conformations were sampled at every 20 ps after discarding the first 2 ns of the constant pressure and temperature simulation.

FKBP was chosen as a model system for folding because it is a protein being studied experimentally in our laboratory.[42] We have accumulated experimental data for the effects of a large number of charge mutations on the folding stability (J. Batra, HT, and HXZ, to be published). The same mutations are the targets of the present study. Starting from the X-ray structure (Protein Data Bank entry 1fkb[43]), the MD simulation of FKBP followed the protocol for the barnase-barstar complex. 200 conformations were sampled at every 15 ps after discarding the first 1 ns of the constant pressure and temperature simulation.

Previously we have used PB calculations to predict the pH dependences of the solubility of two proteins, ribonuclease Sa and insulin.[17] The transfer free energy from the condensed phase to the solution phase was calculated by averaging over conformations sampled from constant-pH MD simulations, which were based on the GB model.[44] For testing scaled GBr$^6$, here we simply took the conformations generated in our previous work and used the PB results calculated there as the benchmark. We just point out that 100 conformations were collected at each of the nine pH values (2.3, 2.9, 3.6, 4.0, 4.5, 4.8, 5.0, 5.2, and 5.4) for ribonuclease Sa and at each of the eight pH values (4.0, 4.5, 5.0, 5.5, 5.75, 6.25, 6.75, and 7.0) for insulin.

We have obtained PB results for the electrostatic interaction energies of the transient complexes of four protein pairs, formed by barnase and barstar, interleukin-4 (IL4) and interleukin-4 binding protein (IL4BP), colicin E9 and immunity protein 9 (Im9), and acetylcholinesterase (AChE) and fasciculin (fas).[31−33] Each transient complex was represented by 100 configurations, generated by sampling in the six-dimensional space of relative translation and rotation (the conformations of the two subunits were held fixed). Here we used the same configurations and the PB results to test scaled GBr$^6$.

**3.2. Mutation Protocol.** In the present work mutational effects on binding free energy were calculated for the barnase-barstar complex, the two GTPase-WASp complexes, and the U1A-U1SLII complex. Similarly mutational effects on folding free energy were calculated for FKBP. Mutations were modeled on each of the conformations collected from MD simulations of a "wild-type" protein or protein complex, and the protocol for mutation was the same as developed previously in single-conformation studies.[9,11,13,14] Briefly, for each single mutation, the LEAP program was used to replace the wild-type side chain with the mutant side chain. The new side chain was then energy minimized in vacuum (while holding the rest of protein or protein complex fixed) up to 50,000 steps. Multiple mutations were decomposed into a series of single mutations.

The 17 mutations on the barnase-barstar complex studied previously[9] were also the target of the present work. Of these, 11 are single mutations (bnK27A, bnD54A, bnR59A, bnE60A, bnR83Q, bnR87A, bnH102A, bsD35A, bsD39A, bsE76A, and bsE80A; bn and bs refer to barnase and barstar, respectively). The remaining 6 are double mutations (bnK27A/bsD39A, bnR59A/bsD35A, bnR59A/bsE76A, bnR59A/bsE80A, bnR83Q/bsD39A, and bnR87A/bsD39A). Similarly, seven single mutations (R7Q, K20Q, K22Q, K23Q, K50A, K60Q, and R70Q) on the U1A in its complex with U1SLII, studied previously,[11] were the target of the present work.

For the Cdc42-WASp complex, the present study covered a total of 10 mutations. Of these, 7 were single mutations (E49K on Cdc42, K230A, K230E, K231A, K231E, K232A, and K232E on WASp), 1 was a double mutation (E49K/E178K on Cdc42), and 2 were triple mutations (K230A/K231A/K232A and K230E/K231E/K232E on WASp). Two mutations were studied on the TC10-WASp complex: K63E and K63E/T192E; both are on TC10. For the purpose of making error assessment on scaled GBr$^6$, we group together the 12 mutations on the two related GTPase-WASp complexes.

We studied 26 mutations on FKBP to find their effects on the folding free energy. These include 17 single, 5 double, 2 triple, 1 quadruple, and 1 quintuple mutation (listed in Figure 6).

**3.3. PB and GB Calculations.** The PB results for benchmarking scaled GBr$^6$ on pH dependence of protein solubility and on electrostatic interaction energy of transient

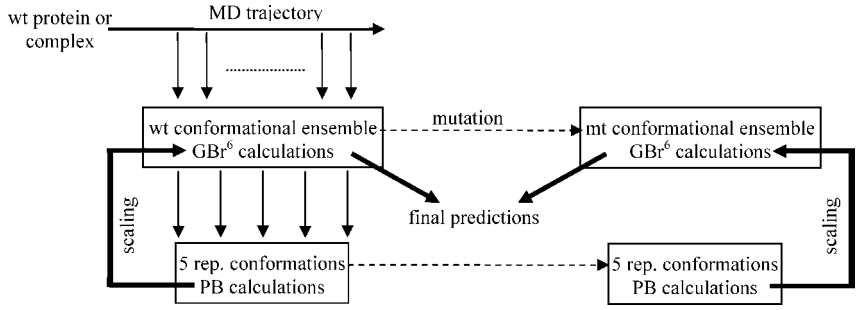



***Figure 2.*** Illustration of the scaling protocol.

complex were available from previous studies.[17,32] PB results for electrostatic contributions to binding and folding free energies were newly calculated here on the conformational ensembles described above. The calculations were done by the UHBD program[2] following a previously established protocol.[9,11,13,14] In particular, the dielectric boundary between the protein or RNA low dielectric and the solvent high dielectric was specified by the van der Waals surface instead of the more popularly used molecular surface. We have had the most success with the former surface in comparing PB calculations with experimental results for mutational effects on protein folding and binding free energies[9−11,13,14] and for electrostatic enhancement of protein binding rates.[33,34] The van der Waals surface is also what is modeled into our GB method, GBr$^6$.[26]

Mutational effects on the barnase-barstar binding free energy were calculated for an ionic strength of 25 mM. In addition, PB and GB calculations were also carried out for the wild-type complex at ionic strengths of 50, 125, 225, 325, 525, and 1000 mM to test scaled GBr$^6$ on salt effects. The ionic strengths for PB and GB calculations were 100 mM for the GTPase-WASp complexes, 160 mM for the U1A-U1SLII complex, and 150 mM for FKBP. All calculations used the Amber94 atomic partial charges[45] and Bondi radii (H, 1.2 Å; O, 1.5 Å; N, 1.55 Å; C, 1.7 Å; S and P, 1.8 Å).[46] Except for the U1A-U1SLII complex, all PB results were obtained by solving the linearized PB equation. For the protein-RNA complex, the full PB equation was solved due to the high charges on the system. We have developed a version of GBr$^6$ called GBr$^6$NL,[47] which mimics the full nonlinear PB equation. However, since the nonlinear PB equation has not been modeled by other GB methods and our scaling-based reparameterization was intended for GB methods in general, even for the U1A-U1SLII complex we used raw GBr$^6$ results for scaling. (We did use the nonlinear PB results for the U1A-U1SLII complex to test the scaling method on GBr$^6$NL. As expected, the raw GBr$^6$NL results had slightly less deviations from the PB targets than the raw GBr$^6$ results, but the scaling method worked equally well for the two GB versions.)

**3.4. GB Scaling Protocol.** Rather than refine GBr$^6$ internally, our strategy for reparameterization was to take raw GBr$^6$ results and postprocess them by scaling. The scaling was based on PB results for a small number (five) of conformations judiciously selected as representatives of the conformation ensemble. The selection varied somewhat from application to application, but followed the same overall design, as illustrated in Figure 2.

Let us illustrate the scaling protocol using mutational effects on the binding free energy of a protein complex. First, for each conformation in the ensemble, the raw GBr$^6$ result for $\Delta G_b$(wt) was calculated. Then, five conformations were selected for scaling purpose, based on the mean ($m$) and standard deviation ($\sigma$) of the raw GBr$^6$ results within the conformational ensemble. The representative conformations had raw GBr$^6$ results closest to the following target values: $m$, $m \pm 0.5\sigma$, $m \mp 0.75\sigma$, $m \pm 1.0\sigma$, and $m \mp 1.5\sigma$. The first or second set of signs was chosen depending on whether the actual value of the first conformation was above or below $m$. Other target values were experimented with, but the above values resulted in the best overall performance. We now refer to the raw GBr$^6$ results for the solvation component of $\Delta G_b$(wt) as $\Delta G_{GB}(i)$, $i = 1$ to 5, for the five conformations.

The corresponding PB results, $\Delta G_{PB}(i)$, $i = 1$ to 5, for the five representative conformations were also obtained. The average of the five individual ratios, $\Delta G_{PB}(i)/\Delta G_{GB}(i)$, was finally taken as the factor ($\lambda$) for scaling the raw GBr$^6$ results of the whole conformational ensemble. The same five representative conformations were also used for calculating the scaling factors of all mutants of the same protein complex. For each mutant, calculating the scaling factor involved again obtaining PB results on the five representative conformations. The ratio $\Delta G_{PB}(i)/\Delta G_{GB}(i)$ could spuriously deviate significantly from 1 when the magnitudes of $\Delta G_{b;solv}$ calculated by GBr$^6$ or PB on the representative conformations were too small. To prevent the scaling factor from being biased by such spurious ratios, we filtered out ratios outside the range of 0.75 to 1.25. This filtering was triggered only rarely in the present study.

After the scaling factors were separately found for the wild-type complex and each mutant, the mutational effect on the electrostatic contribution to the binding free energy was calculated as

$$\Delta\Delta G_b = [\lambda(\mathrm{mt})\Delta G_{GB}(\mathrm{mt}) + \Delta G_{b;Coul}(\mathrm{mt})] - [\lambda(\mathrm{wt})\Delta G_{GB}(\mathrm{wt}) + \Delta G_{b;Coul}(\mathrm{wt})] \quad (8)$$

This equation was applied to the individual conformations in the ensemble, all using the same scaling factors $\lambda$(mt) and $\lambda$(wt). The final, ensemble-averaged, prediction for the mutational effect was taken as an average of the results for $\Delta\Delta G_b$ calculated on the individual conformations. The same procedure was also adopted for studying salt effects on the binding free energy. In that case 'mt' and 'wt' refer to the salt concentration of interest and a reference salt concentration, respectively.

***Table 1.*** PB, sGBr$^6$, and Experimental Results (in kcal/mol) for $\Delta\Delta G_b$ of 17 Mutations on the Barnase-Barstar Complex[a]

| mutants | PB range | PB m ± sd | sGBr$^6$ m ± sd | expt |
|---|---|---|---|---|
| bnK27A | 3.97 to 6.77 | 5.28 ± 0.50 | 5.34 ± 0.43 | 5.4 |
| bnD54A | −2.93 to −1.36 | −2.05 ± 0.26 | −1.91 ± 0.17 | −0.9 |
| bnR59A | 3.11 to 6.52 | 4.82 ± 0.57 | 4.89 ± 0.53 | 5.2 |
| bnE60A | −1.40 to 0.65 | −0.52 ± 0.35 | −0.31 ± 0.36 | −0.3 |
| bnR83Q | 4.08 to 8.82 | 6.72 ± 0.61 | 6.69 ± 0.57 | 5.4 |
| bnR87A | 3.89 to 6.20 | 4.91 ± 0.43 | 4.84 ± 0.39 | 5.5 |
| bnH102A | 0.28 to 2.39 | 1.50 ± 0.34 | 1.55 ± 0.36 | 6.1 |
| bsD35A | 2.17 to 6.13 | 3.82 ± 0.72 | 3.92 ± 0.61 | 4.5 |
| bsD39A | 5.56 to 10.32 | 7.97 ± 0.83 | 8.05 ± 0.79 | 7.7 |
| bsE76A | 1.92 to 4.83 | 3.35 ± 0.54 | 3.37 ± 0.52 | 1.4 |
| bsE80A | −0.16 to 0.95 | 0.38 ± 0.18 | 0.39 ± 0.13 | 0.5 |
| bnK27A/bsD39A | 6.93 to 11.25 | 9.24 ± 0.83 | 9.25 ± 0.75 | 8.2 |
| bnR59A/bsD35A | 4.27 to 9.71 | 7.02 ± 0.89 | 7.03 ± 0.80 | 6.3 |
| bnR59A/bsE76A | 2.97 to 6.39 | 4.63 ± 0.58 | 4.68 ± 0.53 | 4.9 |
| bnR59A/bsE80A | 2.99 to 6.47 | 4.74 ± 0.60 | 4.88 ± 0.54 | 5.1 |
| bnR83Q/bsD39A | 6.93 to 11.68 | 9.36 ± 0.92 | 9.33 ± 0.91 | 6.4 |
| bnR87A/bsD39A | 5.88 to 10.52 | 8.17 ± 0.86 | 8.20 ± 0.80 | 7.1 |

[a] PB and sGBr$^6$ results were calculated on 548 conformations sampled from an MD simulation. Range, m, and sd refer to the range of variation, the mean, and the standard deviation, respectively, of $\Delta\Delta G_b$ among the conformations.

The scaling protocol for mutational effects on the folding free energy was similar. In this case, $\Delta G_f$, as defined in eq 4, was used for scaling. As noted in subsection 2.2, even $\Delta G_f$(wt) is mutant-specific; therefore, unlike in the case of binding free energy, it is not possible to have one set of five representative conformations that is applicable to all mutants. Instead, for each mutant, the five representative conformations were selected based on GBr$^6$ results for $\Delta G_f$(wt). The scaling factors for the wild-type protein and the mutant were then found by obtaining PB results for the solvation components of $\Delta G_f$(wt) and $\Delta G_f$(mt) on these five conformations.

Extensions of the scaling protocol to studies of pH dependence of protein solubility and electrostatic rate enhancement were straightforward. For the former we just mention that $\Delta G_t$, the electrostatic component of the transfer free energy, was used for scaling. For the latter study, we mention that the interaction energy $\Delta G_{el}$* itself, as opposed to its changes by mutation or by salt, was of direct interest.

## IV. Results and Discussion

**4.1. Importance of Conformational Sampling.** The premise of our method for GB reparameterization is that results calculated from an extensive conformational ensemble are much more reliable than those from a single conformation. It can be argued, from both fundamental and practical points of view, that a conformational ensemble is superior to a single conformation for predicting binding and folding free energies. In mimicking experimental measurements, an ensemble of conformations is obviously more realistic than a single conformation; experimental results are averages over a conformational ensemble under certain solvent conditions. In a statistical sense, the errors of calculation results decrease with increasing size of the conformational ensemble.

We use PB results for the effects of 17 mutations on the binding free energy of the barnase-barstar complex to illustrate the comparison between ensemble and single-conformation predictions. For each mutation, we take the mean value of the PB results for $\Delta\Delta G_b$ over the 548 sampled conformations as the ensemble prediction. The 548 conformations were sampled over the course of 5.5 ns of MD simulation. The pairwise root-mean-square-deviations (rmsd) between $C_\alpha$ atoms of the 548 conformations were peaked around 1 Å, typical of conformational fluctuations of structured proteins on the ns time scale. For all the mutants, the ranges of variation of $\Delta\Delta G_b$ within the conformational ensemble were found to be comparable to the magnitudes of the mean $\Delta\Delta G_b$ values (Table 1). For example, for the barnase R59A mutation, PB results for $\Delta\Delta G_b$ among the 548 conformations varied from 3.1 to 6.5 kcal/mol, spanning a range of 3.4 kcal/mol. In comparison, the mean $\Delta\Delta G_b$ is 4.8 kcal/mol, and the standard deviation is 0.6 kcal/mol. The fact that, relative to the mean value, the range of variation of $\Delta\Delta G_b$ is large whereas the standard deviation is small provides direct evidence that the ensemble prediction is much more reliable than single-conformation predictions.

We further assess the ensemble and single-conformation predictions against experimental binding data for the 17 mutations (Table 1).[48−50] For each single conformation, we measure prediction error by the root-mean-square-deviation (rmsd) between PB results for $\Delta\Delta G_b$ and experimental data for the 17 mutations. The errors calculated over the 548 sampled conformations are presented in Figure 3 as a histogram. They range from 1.25 to 2.27 kcal/mol. In comparison, the ensemble prediction has an rmsd of 1.55 kcal/mol from experiment. The latter value places the ensemble prediction in the 31st percentile of the single-conformation predictions. In other words, the probability of a randomly chosen single conformation performing worse than the ensemble is 69%. We thus show that, relative to single-conformation predictions, the ensemble prediction is not only more precise (as measured by standard deviation of the ensemble) but also more accurate (as measured by rmsd from experiment).

**4.2. Performance of Raw GBr$^6$.** To establish a baseline for assessing the performance of scaled GBr$^6$, we first benchmark the original GBr$^6$ against PB. For each $\Delta\Delta G_b$, $\Delta\Delta G_f$, $\Delta G_t$, or $\Delta G_{el}$* value, we take the ensemble-averaged PB result as the target and measure error as the deviation of the ensemble-averaged GBr$^6$ result from the PB target. When multiple mutants on the same protein or protein complex

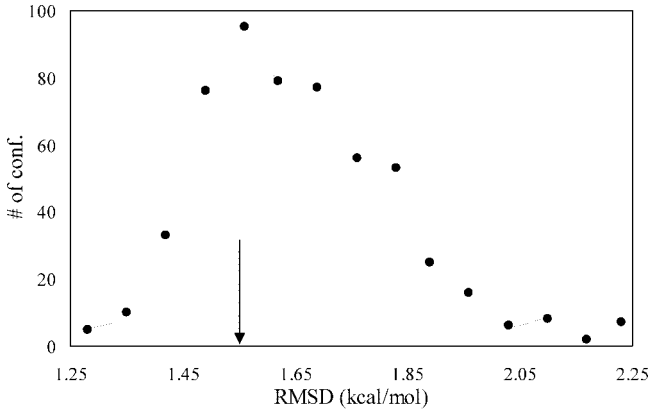



***Figure 3.*** Distribution of PB-experiment difference among 548 conformations of the barnase-barstar complex. On each conformation, PB results for $\Delta\Delta G_b$ of 17 mutations were obtained, and their rmsd from experiment was used as the measure of PB-experiment difference. A vertical arrow, at 1.55 kcal/mol, indicates the rmsd of the ensemble prediction.

are studied (for $\Delta\Delta G_b$ and $\Delta\Delta G_f$ results), or when one protein is studied at different pH (for $\Delta G_t$ results), we report a single error value, which is given by the rmsd from the PB targets among all the mutants or all the pH values.

The errors of the original GBr$^6$ for the various types of calculations are summarized in Table 2. For the barnase-barstar complex, $\Delta\Delta G_b$ errors calculated by GBr$^6$ for the 17 individual mutations ranged from 0.1 to 1.9 kcal/mol, resulting in an overall rmsd of 0.85 kcal/mol. Similar $\Delta\Delta G_b$ errors are also seen for mutations on the GTPase-WASp and U1A-U1SLII complexes.

It is interesting to note that the $\Delta\Delta G_b$ errors seen on these complexes are suppressed by error cancelation between calculations on a complex and calculations on the separate subunits. For example, for the barnase R59A mutation, GBr$^6$ overestimated the magnitude of the solvation energy of the complex by 2.22 kcal/mol and at the same time overestimated the magnitudes of the solvation energies of the two subunits by 2.61 and 0.05 kcal/mol, respectively. Taken together, the net error on $\Delta G_b$(mt) was only 0.44 kcal/mol. While this error cancellation was widely seen on the protein complexes studied here, there is no guarantee that it will always occur. It is possible that, for a given protein complex, GBr$^6$ overestimates the solvation energy of the complex but at the same time underestimates the solvation energies of the subunits. In that situation, the errors accumulate rather than cancel. Errors from calculations on a mutant and those on the wild-type complex may also either cancel or accumulate. For the barnase R59A mutation, GBr$^6$ errors on $\Delta G_b$(mt) and $\Delta G_b$(wt) occurred in opposite directions, resulting in an accumulative error of 0.64 kcal/mol on $\Delta\Delta G_b$. On the other hand, for the U1A-U1SLII complex, GBr$^6$ errors on $\Delta G_b$(mt) and $\Delta G_b$(wt) occurred in the same direction; otherwise the differences in $\Delta\Delta G_b$ between GBr$^6$ and PB, which was from solving the nonlinear PB equation for this particular complex, would have been much greater.

The $\Delta\Delta G_f$ errors calculated by GBr$^6$ for the 26 individual mutations on FKBP reached as much as 2.4 kcal/mol. Even though the overall rmsd of 1.1 kcal/mol of all the 26 mutations is only slightly more than what was found above for the mutational effects on the binding free energy of the barnase-barstar complex, the $\Delta\Delta G_f$ errors on FKBP are much more significant. According to PB calculations (see Figure 6 below), of the 26 mutations, 9 changed the folding free energy of FKBP by less than 0.5 kcal/mol, and another 8 by 0.5 to 1.0 kcal/mol; the largest effect of an individual mutation was 3.5 kcal/mol. In contrast, as shown in Table 1 under the "PB m ± sd" heading, only 2 of the 17 mutations on the barnase-barstar complex affected the binding free energy by less than 1.0 kcal/mol; the largest effect of an individual mutation was 9.4 kcal/mol.

The $\Delta G_t$ results calculated by GBr$^6$ for ribonuclease Sa at nine pH values deviated from their PB targeted by 0.2 to 0.9 kcal/mol, resulting in an overall rmsd of 0.46 kcal/mol. Deviations of such magnitudes are significant because the PB results at the nine pH values differed at most by only 1.3 kcal/mol, and it is these small differences that predict the pH dependence of the solubility.[17] For insulin, the deviations of GBr$^6$ results for $\Delta G_t$ results at eight pH values were quite large, ranging from 2.5 to 9.3 kcal/mol (with an rmsd of 6.8 kcal/mol). The larger deviations most likely were due to the large size of the insulin hexamer (with a total of 306 residues); our GBr$^6$ method was benchmarked on a set of proteins with up to 250 residues.[26]

In our previous study,[32] the PB results for $\Delta G_{el}$* calculated on the transient-complex ensembles of the barnase-barstar, IL4-IL4BP, E9-Im9, and AChE-fas protein pairs were −3.3, −4.3, −3.1, and −4.0 kcal/mol, respectively. The corresponding GBr$^6$ results obtained here were −0.45, −1.4, −0.73, and 0.57 kcal/mol, respectively. The raw GBr$^6$ results thus had large errors.

While the ensemble-averaged values of $\Delta\Delta G_b$, $\Delta\Delta G_f$, $\Delta G_t$, and $\Delta G_{el}$* obtained by GBr$^6$ had relatively large deviations from the PB targets, in each application the GBr$^6$ and PB sets of values calculated on the individual conformations did show strong correlations. In Figure 4 we illustrate the correlations on the E49K/E178K mutation of Cdc42 for binding and on the T21K mutation of FKBP for folding. From linear regression analyses with the requirement of a zero intercept, the correlation $R^2$ values were ∼0.99. The strong correlations between raw GBr$^6$ results and their PB counterparts are the foundation of our scaling-based GB reparameterization. The errors of raw GBr$^6$ are manifested by a nonunity slope of the GBr$^6$−PB correlation, which, as will be seen shortly, is rectified by the scaling.

**4.3. Performance of Scaled GBr$^6$.** Upon applying scaling factors calculated on a small subset of five conformations, the errors of GBr$^6$ on $\Delta\Delta G_b$, $\Delta\Delta G_f$, $\Delta G_t$, and $\Delta G_{el}$* all dropped to within 0.3 kcal/mol (Table 2). Just like in the previous subsection, errors refer to deviations from the PB targets. For the barnase-barstar complex, the errors of scaled GBr$^6$ for 13 of the 17 mutations were less than 0.1 kcal/mol, and the errors of the remaining 4 mutations were less than 0.2 kcal/mol; the overall rmsd was 0.09 kcal/mol. Notice that the latter value is much smaller than the fluctuations of the PB results for $\Delta\Delta G_b$ within the conformational ensemble (see Table 1). The final predictions of scaled GBr$^6$ for $\Delta\Delta G_b$ are given in Table 1 for comparison against experimental

***Table 2.*** Errors of Raw GBr$^6$ and sGBr$^6$ as Measured by Deviations from PB Results[a]

| system | no. of conf. | no. of mt or pH[b] | PB results | GBr$^6$ error | sGBr$^6$ error |
|---|---|---|---|---|---|
| | | | $\Delta\Delta G_b$ | | |
| Bn-Bs | 548 | 17 | −2.1 to 9.4 | 0.85 | 0.09 |
| GTPase-WASp #1 | 100 | 12 | −1.2 to7.9 | 1.12 | 0.14 |
| GTPase-WASp #5 | 100 | 12 | −1.9 to 7.2 | 1.28 | 0.19 |
| GTPase-WASp #9 | 100 | 12 | −1.8 to 8.6 | 0.95 | 0.20 |
| GTPase-WASp #13 | 100 | 12 | −1.5 to 8.3 | 1.30 | 0.32 |
| GTPase-WASp #17 | 100 | 12 | −3.1 to 6.7 | 1.30 | 0.32 |
| U1A-U1SLII | 100 | 7 | 0.1 to 3.1 | 0.69 | 0.05 |
| | | | $\Delta\Delta G_f$ | | |
| FKBP | 200 | 26 | −3.5 to 1.4 | 1.11 | 0.14 |
| | | | $\Delta G_t$ | | |
| RNase Sa | 100 | 9 | −13.8 to −12.5 | 0.46 | 0.07 |
| insulin | 100 | 8 | −63.2 to −60.9 | 6.83 | 0.34 |
| | | | $\Delta G_{el}$* | | |
| Bn-Bs | 100 | | −3.30 | 2.85 | 0.07 |
| IL4-IL4BP | 100 | | −3.05 | 2.94 | 0.22 |
| E9-Im9 | 100 | | −4.33 | 2.32 | 0.07 |
| AChE-fas | 100 | | −4.04 | 4.61 | 0.06 |

[a] All are in kcal/mol. [b] In $\Delta\Delta G_b$ and $\Delta\Delta G_f$ calculations, multiple mutations on a given protein or protein complex were studied. Error was measured by rmsd among the mutations. In $\Delta G_t$ calculations, a given protein was studied at multiple pH values; again rmsd was used as the measure of error.

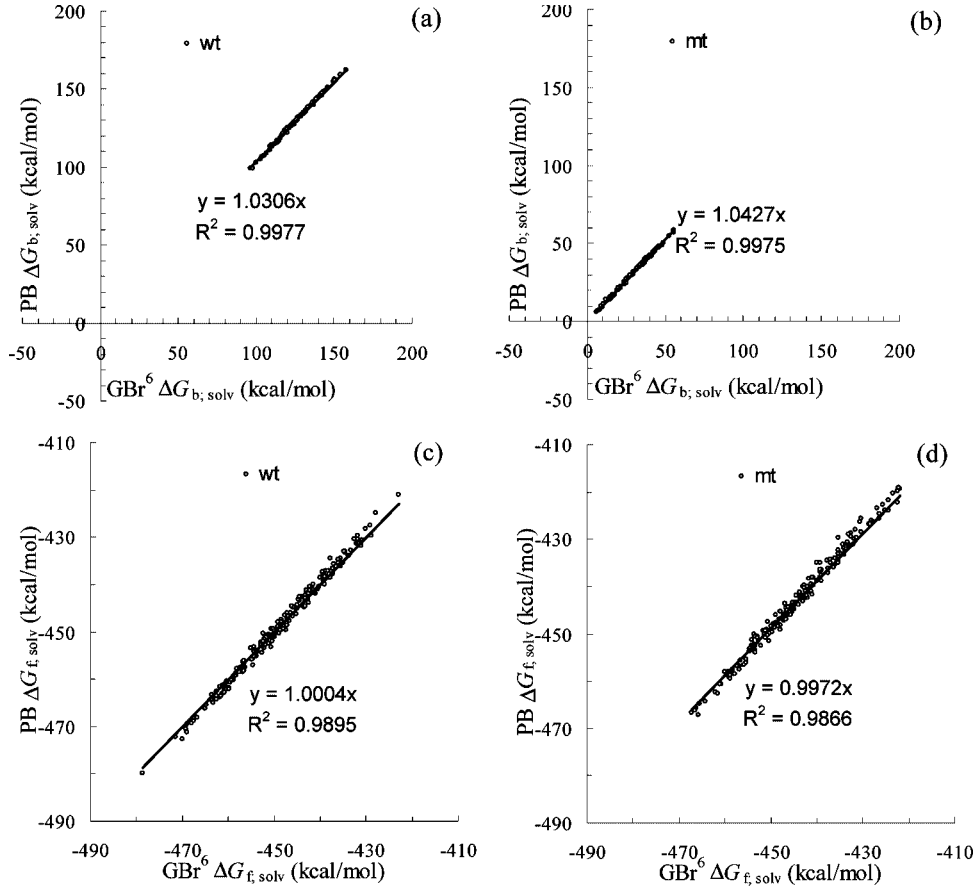



***Figure 4.*** Correlation of GBr$^6$ and PB results for $\Delta G_{b;solv}$ ($\Delta G_{f;solv}$), the solvation component of the electrostatic contribution to the binding (folding) free energy. Both sets of results were calculated on the same conformational ensemble. (a) and (b) $\Delta G_{b;solv}$ of wild-type Cdc42-WASp complex and the Cdc42 E49K/E178K mutant. (c) and (d) $\Delta G_{f;solv}$ of wild-type FKBP and the T21K mutant. In the linear regression analyses, the *y*-intercepts were always constrained at zero.

and PB results. The rmsd of scaled GBr$^6$ results from experiment was 1.53 kcal/mol, virtually the same as the 1.55-kcal/mol rmsd of PB results from experiment. We further tested scaled GBr$^6$ for predicting salt effects on the binding free energy of the barnase-barstar complex. According to PB calculations, increasing ionic strength from 25 mM to 1 M reduces the magnitude of the binding free energy by 2.2 kcal/mol. As Figure 5 shows, scaled GBr$^6$ remarkably well

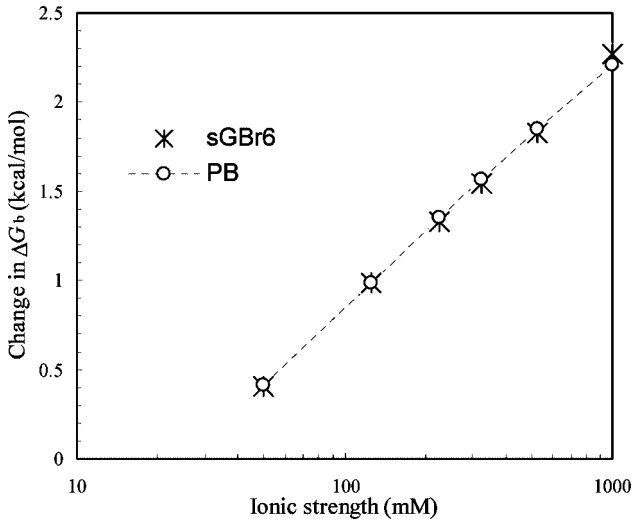



***Figure 5.*** Comparison of scaled GBr$^6$ and PB results for the salt dependence of the barnase-barstar binding free energy. Changes in $\Delta G_b$ from the respective results at an ionic strength of 25 mM are displayed.

reproduces the salt dependence of the binding free energy obtained by PB.

Very good agreement between scaled GBr$^6$ and PB was also observed for the GTPase-WASp complexes. As summarized in Table 2, the RMSDs of scaled GBr$^6$ from PB, calculated over 12 mutations on the two related complexes, ranged from 0.14 to 0.32 kcal/mol among conformational ensembles sampled using five different NMR models as starting structures. The average rmsd over the five conformational ensembles was 0.23 kcal/mol.

The highly charged protein-RNA complex between U1A and U1SLII did not present an obstacle for the scaling method. The rmsd between scaled GBr$^6$ and PB was just 0.05 kcal/mol among seven mutations. Note that for this system the PB results were obtained from solving the nonlinear PB equation, so the test on this system demonstrates that scaled GBr$^6$ can work well even for highly charged systems such as protein-nucleic acid complexes, for which the nonlinear PB equation would otherwise be required. Solving the full nonlinear PB equation takes much longer CPU time than solving the linearized version. For highly charged systems, scaled GBr$^6$ affords especially significant gain in computational speed and yet can still be highly accurate for calculating small electrostatic effects such as those caused by mutations on binding free energy.

The success of scaled GBr$^6$ with predicting binding free energy was repeated on predicting folding free energy. For 26 mutations on FKBP, the rmsd from PB targets was just 0.14 kcal/mol. A comparison between $\Delta\Delta G_f$ results obtained by scaled GBr$^6$ and PB for the 26 mutations is displayed in Figure 6. Note that scaled GBr$^6$ performed equally well for single mutations and for multiple mutations. Once again the deviations of GBr$^6$ from PB are much smaller than the fluctuations of the PB results for $\Delta\Delta G_f$ within the conformational ensemble.

Let us use a binding example and a folding example to illustrate how scaled GBr$^6$ achieved its accuracy. The binding example is the E49K/E178K mutation of Cdc42, and the folding example is the T21K mutation of FKBP. The correlations between raw GBr$^6$ and PB for these two examples have been shown in Figure 4. For binding, the slope of the linear regression between raw GBr$^6$ and PB was 1.0306 for the wild-type complex and 1.0427 for the mutant. The optimal scaling would be to use these slopes as scaling factors. However, obtaining these slopes would require PB results for the whole conformational ensemble, which defeats the purpose of designing GB methods. By using PB results for a small subset of five conformations, we obtained scaling factors of 1.0306 for the wild-type complex and 1.0403 for the mutant. These scaling factors are very close to the slopes from linear regression, hence explaining why scaled GBr$^6$ was so accurate. Similarly, for the folding examples, the slopes from linear regression were 1.0004 for wild-type FKBP and 0.9972 for the mutant. For the subset of five conformations, the scaling factors were found to be 1.0016 for the wild-type protein and 0.9982 for the mutant. The latter pairs of values again are very close to the respective slopes.

Raw GBr$^6$ already did a good job in reproducing PB for $\Delta G_t$ of ribonuclease Sa. Scaling was able to further reduce the rmsd, from 0.46 to 0.07 kcal/mol, among nine pH values. For insulin, scaled GBr$^6$ dramatically reduced the error on $\Delta G_t$, from 6.8 to 0.34 kcal/mol. For $\Delta G_{el}$* of the four protein pairs, scaling was able to essentially erase the large errors of raw GBr$^6$. Scaled GBr$^6$ results differed from the PB targets by 0.2 kcal/mol or less for all four protein pairs.

We carried out some experimentation with the number of representative conformations used for calculating scaling factors. Our conclusion is that five is the optimal compromise between accuracy (demanding more conformations) and computational cost (demanding less conformations). However, it seems that even with fewer numbers of conformations, the performance of scaled GBr$^6$ will not deteriorate significantly. In our test on the barnase-barstar complex, using only three representative conformations (the first, second, and third, or the first, third, and last of the original five), the rmsd between scaled GBr$^6$ and PB increased only minutely, from 0.09 to 0.14 kcal/mol.

The success of scaled GBr$^6$ in achieving very good agreement with PB results owes in large part to the quality of the original GBr$^6$ results. As shown in Figure 4, the original GBr$^6$ results have high linear correlations with PB results, and the *y*-intercepts of these correlations are nearly zero. In the early work of David et al.,[30] the GB method used produced significant *y*-intercepts, which were probably the main reason for the modest improvement achieved by reparameterization.

**4.4. Cross-Validations.** It is interesting to know how well the scaling scheme works when scaling factors obtained on five conformations selected from one subensemble are applied to another subensemble. As an example, an early portion of a MD trajectory may be used to generate the conformations for producing the scaling factors, which may then be applied to conformations sampled from the continuation of the MD simulation. To address the question posed above, we designed two types of cross-validation tests, depending on how the conformations were divided into subensembles. The barnase-barstar complex was chosen for

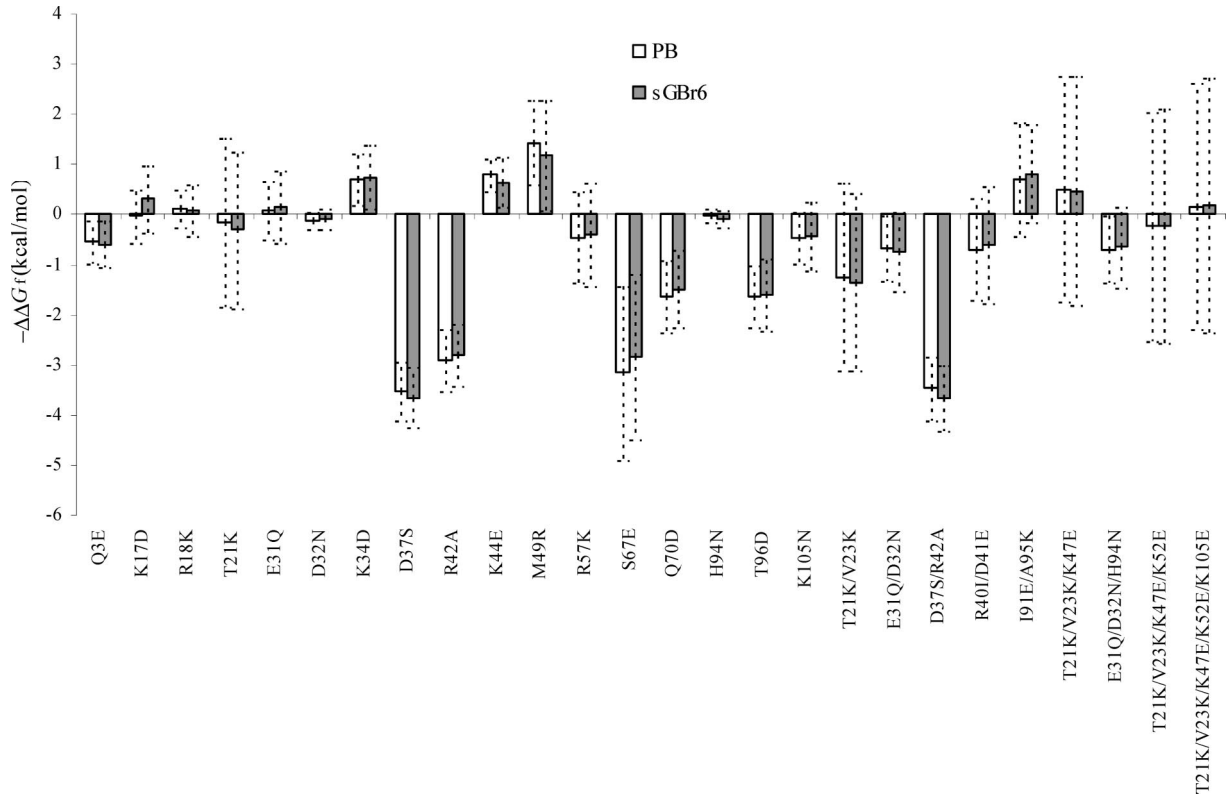



***Figure 6.*** Comparison of scaled GBr$^6$ and PB results for $\Delta\Delta G_f$ of 26 mutations on FKBP. Error bars indicate standard deviations within the conformational ensemble.

***Table 3.*** Cross-Validation of sGBr$^6$ on $\Delta\Delta G_b$ Results of 17 Mutations on the Barnase-Barstar Complex[a]

| training | type-1 test | type-2 test |
|---|---|---|
| subensemble 1 | 0.15 (0.08) | 0.06 (0.06) |
| subensemble 2 | 0.09 (0.09) | 0.15 (0.15) |
| subensemble 3 | 0.14 (0.16) | 0.14 (0.14) |
| subensemble 4 | 0.14 (0.12) | 0.12 (0.12) |
| average | **0.13 (0.11)** | **0.12 (0.12)** |

[a] Under each type of test, error in kcal/mol, as measured by rmsd from PB targets, is given. The number in parentheses represents the error when the scaling factors were applied to the training subensemble itself; the number outside represents the error obtained on the other three subensembles.

this purpose simply because it happened to have the most conformations (548) saved and PB results calculated for benchmarking. In the first cross-validation test, the 548 conformations were divided into four equal subensembles according to time sequence. In the second cross-validation test, every fourth conformation was collected into a subensemble, and each of the fist four conformations of the whole ensemble started a different subensemble.

We refer to the subensemble from which the scaling factors were obtained as the training subensemble and examine whether the scaling factors when applied to a different subensemble would lead to larger errors than when applied to the training subensemble itself. As the results in Table 3 show, for both types of cross-validation tests, applying the scaling factors to the training or a different subensemble produces very similar errors.

A caveat about the cross-validation tests is that the different subensembles must all be part of the conformational fluctuations within a single deep energy well. If the protein or protein complex undergoes a conformational transition and different subensembles are collected before and after the transition, cross validation clearly will fail. It is important that, in designing scaling schemes, such conformational transitions are recognized, and one set of scaling factors is used for each unique deep energy well.

## V. Conclusion

We have demonstrated that a scaled GB method accurately reproduces PB results for binding and folding free energies, transfer energies between crystalline and solution phases, and electrostatic interaction energies in transient complexes. The scaled GB method thus opens the door to incorporate conformational sampling in robust and accurate modeling of small electrostatic effects, which fall within the error range of current GB methods.

The deviations of scaled GBr$^6$ from PB are much smaller than the fluctuations of PB results within conformational ensembles. Our scaling scheme thus seems to have pushed GB methods to their accuracy limit. In our opinion, reparameterization without reference to any PB results will never be able to reach such a level of accuracy. Combining one of the most accurate raw GB methods with a PB-guided scaling method, scaled GBr$^6$ promises to be a prototype for a new generation of fast continuum solvation models for incorporating conformational sampling in binding and folding free energy and other related calculations. A major application of GB methods is in implicit-solvent molecular dynamics simulations. Applying scaled GBr$^6$ in such simulations is underway.

**Acknowledgment.** This work was supported in part by NIH grant GM058187.

## References

(1) Gilson, M. K.; Honig, B. Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. *Proteins* **1988**, *4*, 7–18.

(2) Madura, J. D.; Briggs, J. M.; Wade, R.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; McCammon, J. A. Electrostatic and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. *Comput. Phys. Commun.* **1995**, *91*, 57–95.

(3) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037–10041.

(4) Grant, J. A.; Pickup, B. T.; Nicholls, A. A smooth permittivity function for Poisson-Boltzmann solvation methods. *J. Comput. Chem.* **2001**, *22*, 608–640.

(5) Fogolari, F.; Brigo, A.; Molinari, H. The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J. Mol. Recognit.* **2002**, *15*, 377–392.

(6) Luo, R.; David, L.; Gilson, M. K. Accelerated Poisson-Boltzmann calculations for static and dynamic systems. *J. Comput. Chem.* **2002**, *23*, 1244–1253.

(7) Lu, Q.; Luo, R. A Poisson—Boltzmann dynamics method with nonperiodic boundary condition. *J. Chem. Phys.* **2003**, *119*, 11035–11047.

(8) Baker, N. A. Improving implicit solvent simulations: a Poisson-centric view. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137–143.

(9) Dong, F.; Vijayakumar, M.; Zhou, H.-X. Comparison of calculation and experiment implicates significant electrostatic contributions to the binding stability of barnase and barstar. *Biophys. J.* **2003**, *85*, 49–60.

(10) Dong, F.; Zhou, H.-X. Electrostatic contribution to the binding stability of protein-protein complexes. *Proteins* **2006**, *65*, 87–102.

(11) Qin, S.; Zhou, H.-X. Do electrostatic interactions destabilize protein-nucleic acid binding. *Biopolymers* **2007**, *86*, 112–118.

(12) Bertonati, C.; Honig, B.; Alexov, E. Poisson-Boltzmann calculations of nonspecific salt effects on protein-protein binding free energies. *Biophys. J.* **2007**, *92*, 1891–1899.

(13) Vijayakumar, M.; Zhou, H.-X. Salt bridges stabilize the folded structure of barnase. *J. Phys. Chem. B* **2001**, *105*, 7334–7340.

(14) Dong, F.; Zhou, H.-X. Electrostatic contributions to T4 lysozyme stability: solvent-exposed charges versus semi-buried salt bridges. *Biophys. J.* **2002**, *83*, 1341–1347.

(15) Tan, Y. H.; Luo, R. Protein stability prediction: a Poisson-Boltzmann approach. *J. Phys. Chem. B* **2008**, *112*, 1875–1883.

(16) Zhou, H.-X. Interactions of macromolecules with salt ions: an electrostatic theory for the Hofmeister effect. *Proteins* **2005**, *61*, 69–78.

(17) Tjong, H.; Zhou, H.-X. Prediction of protein solubility from calculation of transfer free energy. *Biophys. J.* **2008**, *95*, 2601–2609.

(18) 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*, 61276129.

(19) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.* **1996**, *100*, 19824–19839.

(20) Ghosh, A.; Rapp, C. S.; Friesner, R. A. Generalized Born model based on a surface integral formulation. *J. Phys. Chem. B* **1998**, *102*, 10983–10990.

(21) Bashford, D.; Case, D. A. Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.

(22) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III. New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.* **2003**, *24*, 1348–1356.

(23) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized Born model. *Proteins* **2004**, *55*, 383–394.

(24) Feig, M.; Brooks, C. L. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.

(25) Gallicchio, E.; Levy, R. M. AGBNP: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comput. Chem.* **2004**, *25*, 479–499.

(26) Tjong, H.; Zhou, H.-X. GBr$^6$: a parameterization-free, accurate, analytical generalized Born method. *J. Phys. Chem. B* **2007**, *111*, 3055–3061.

(27) Gohlke, H.; Kiel, C.; Case, D. A. Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J. Mol. Biol.* **2003**, *330*, 891–913.

(28) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Charles, L.; Brooks, I. Performance comparison of generalized Born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem.* **2004**, *25*, 265–284.

(29) Liu, H. Y.; Zou, X. Electrostatics of ligand binding: parametrization of the generalized Born model and comparison with the Poisson-Boltzmann approach. *J. Phys. Chem. B* **2006**, *110*, 9304–9313.

(30) David, L.; Luo, R.; Gilson, M. K. Comparison of generalized Born and Poisson methods: Energetics and dynamics of HIV protease. *J. Comput. Chem.* **2000**, *21*, 295–309.

(31) Alsallaq, R.; Zhou, H.-X. Energy landscape and transition state of protein-protein association. *Biophys. J.* **2007**, *92*, 1486–1502.

(32) Alsallaq, R.; Zhou, H.-X. Prediction of protein-protein association rates from a transition-state theory. *Structure* **2007**, *15*, 215–224.

(33) Alsallaq, R.; Zhou, H.-X. Electrostatic rate enhancement and transient complex of protein-protein association. *Proteins* **2008**, *71*, 320–335.

(34) Qin, S.; Zhou, H.-X. Prediction of salt and mutational effects on the association rate of U1A protein and U1 small nuclear RNA stem/loop II. *J. Phys. Chem. B* **2008**, *112*, 5955–5960.

(35) Zhou, H.-X. A Gaussian-chain model for treating residual charge-charge interactions in the unfolded state of proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 3569–3574.

(36) Hemsath, L.; Dvorsky, R.; Fiegen, D.; Carlier, M. F.; Ahmadian, M. R. An electrostatic steering mechanism of Cdc42 recognition by Wiskott-Aldrich syndrome proteins. *Mol. Cell* **2005**, *20*, 313–324.

(37) Buckle, A. M.; Schreiber, G.; Fersht, A. R. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-Å resolution. *Biochemistry* **1994**, *33*, 8878–8889.

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

(39) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65*, 712–725.

(40) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

(41) Abdul-Manan, N.; Aghazadeh, B.; Liu, G. A.; Majumdar, A.; Ouerfelli, O.; Siminovitch, K. A.; Rosen, M. K. Structure of Cdc42 in complex with the GTPase-binding domain of the 'Wiskott-Aldrich syndrome' protein. *Nature* **1999**, *399*, 379–383.

(42) Spencer, D. S.; Xu, K.; Logan, T. M.; Zhou, H.-X. Effects of pH, salt, and macromolecular crowding on the stability of FK506-binding protein: An integrated experimental and theoretical study. *J. Mol. Biol.* **2005**, *351*, 219232.

(43) Van Duyne, G. D.; Standaert, R. F.; Schreiber, S. L.; Clardy, J. Atomic structure of the rapamycin human immunophilin FKBP-12 complex. *J. Am. Chem. Soc.* **1991**, *113*, 7433–7434.

(44) Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem.* **2004**, *25*, 2038–2048.

(45) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(46) Bondi, A. van der Waals volumes and radii. *J. Phys. Chem.* **1964**, *68*, 441–451.

(47) Tjong, H.; Zhou, H.-X. GBr$^6$NL: a generalized Born method for accurately reproducing solvation energy of the nonlinear Poisson-Boltzmann equation. *J. Chem. Phys.* **2007**, *126*, 195102.

(48) Schreiber, G.; Fersht, A. R. Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering. *Biochemistry* **1993**, *32*, 5145–5150.

(49) Schreiber, G.; Fersht, A. R. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *J. Mol. Biol.* **1995**, *248*, 478–486.

(50) Frisch, C.; Schreiber, G.; Johnson, C. M.; Fersht, A. R. Thermodynamics of the interaction of barnase and barstar: changes in free energy versus changes in enthalpy on mutation. *J. Mol. Biol.* **1997**, *267*, 696–706.

CT8001656

JCTC Journal of Chemical Theory and Computation

# Anthrax Lethal Factor Investigated by Molecular Simulations

Rolando Hong, Alessandra Magistrato,* and Paolo Carloni

*International School for Advanced Studies (SISSA/ISAS), CNR-INFM-Democritos National Simulation Center, and Italian Institute of Technology (IIT), Trieste, Italy*

Received May 26, 2008

**Abstract:** The anthrax disease is caused by the lethal toxin secreted by the bacterium *Bacillus anthracis*. The toxin is a protein aggregate which contains a Zn-based hydrolase called anthrax Lethal Factor (LF). In this work, we investigate the structure of its Michaelis complex with an optimized MAPKK-like substrate using several computational methods including density functional theory, molecular dynamics, and coarse grained techniques. Our calculations suggest that (i) the presence of second-shell ligands is crucial for tuning the structure, energetics, and protonation state of the metal binding site, as found in other Zn-based enzymes; (ii) the nucleophilic agent is a Zn-bound water molecule; (iii) substrate binding to the active site groove is mainly stabilized by van der Waals interactions; (iv) the bonds most likely involved in the substrate hydrolysis are only mildly polarized by the protein scaffold; and (v) part of helix α19, which is present in one solid state structure of LF (PDB: 1JKY), assumes a coiled conformation.

## 1. Introduction

The anthrax infection caused by the bacterium *Bacillus anthracis* poses a significant threat in biological warfare and terrorism. If ingested or inhaled, the anthrax bacterial spores germinate, resulting in a toxaemia that is usually fatal to the host.[1–4]

Unfortunately, the only way to intervene against anthrax intoxication is to give a generic antibiotic treatment at an early stage of the disease.[5] Thus, there is presently a tremendous effort in investigating the molecular mechanisms responsible for anthrax infection to develop new therapeutic agents.

Most of anthrax's toxic effects are caused by the lethal toxin, a complex consisting of the Protective Antigen (PA) and Lethal Factor (LF) proteins.[6] PA is the membrane-translocating component of the complex; it binds a host cell-surface receptor and translocates LF into the cytosol.[7,10] LF is a cytoplasmatic zinc metalloprotease that cleaves the N-terminal region of selected members of the Mitogen-Activated-Protein-Kinase-Kinase (MAPKK) family;[11] MAPKKs govern the MAPK signaling pathway, controlling the genomic and physiological response of the cell to its environment.[12] LF alters different cell types, apparently, in an evolutionary conserved manner.[13−16]

X-ray crystallographic studies provided the structural details of LF in the free state (PDB: 1J7N),[17] with a segment of one of its substrates, MAPKK-2 (PDB: 1JKY),[17] with an optimized peptidic substrate (PDB: 1PWV, 1PWW)[18] and with synthetic inhibitors (PDB: 1PWP, 1ZXV).[19,20] [The optimized peptide substrate was built[18] using the consensus residues around the scissible bond based on a peptide library screen, flanked by residues of the actual substrate MAPKK-2.]

Domain **I** of LF binds to PA, while domains **II**−**IV** create a long groove that holds the MAPKK-2 N-term[18] (Figure 1). Domain **IV** performs the enzymatic catalysis. It features, in the active site, a zinc ion coordinated by two histidines (His686 and His690) and a glutamate (Glu735). The coordination is completed with a water molecule (or an hydroxide group), which is, most probably, the nucleophilic agent in the catalysis. The coordination is that of a distorted tetrahedron. Similar coordination spheres are found in related metalloproteases from the carboxypeptidase and thermolysin families.[21] Clearly the nature and protonation state of the

* Corresponding author phone: +39 040 3787 529; fax: +39 040 3787 528; e-mail: alema@sissa.it. Corresponding author address: SISSA, via Beirut 2-4, 34014 Trieste, Italy.

10.1021/ct8001877 CCC: $40.75 
Published on Web 09/09/2008

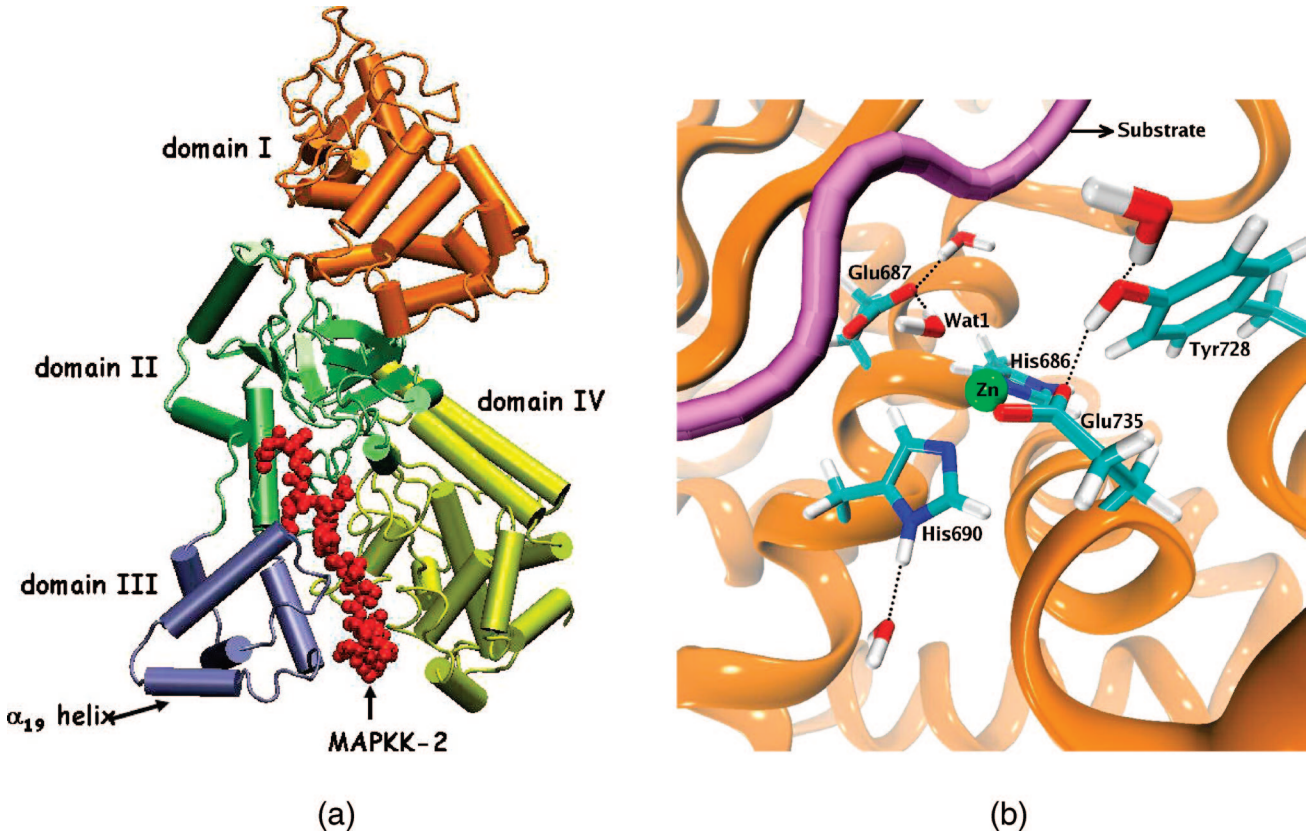



***Figure 1.*** (a) The structure of anthrax Lethal Factor (LF) in complex with its MAPKK-2 substrate (red), as obtained by X-ray crystallography (PDB: 1JKY),[17] includes the following domains: **I** (orange, residues 1−262) is the Protective Antigen (PA) binding domain; **II** (green, residues 263−297, 385−550) is called the Vegetative Insecticidal Protein 2 (VIP2)-like domain because of its similarity with the ADP-ribosyltransferase from *Bacillus cereus* toxin, **III** (blue, residues 303−383) is the helix bundle domain; and **IV** (yellow, residues 552−776) is the catalytic domain. (b) Snapshot from the all-atom MD trajectory featuring the active site (domain **IV**) with the residues invoked to be crucial for the catalytic activity[17,18,25] (shown in licorice representation) and the position of the optimized substrate[18] (shown in purple) used in the simulations.

nucleophilic agent and active site residues involved in the enzymatic process are crucial for the LF proteolytic reaction.

Similarly to what is found in other zinc enzymes,[22] the zinc ligands are stabilized by two outer shells groups,[17,23] including: (i) Glu687, which H-bonds the catalytic water and is believed to play a key role acting as a general base during the reaction.[17] This is consistent with the loss of activity of the LF mutant E687C.[24,25] (ii) Tyr728, which H-bonds Glu735 and whose conservative mutation to phenylalanine (Y728F) impairs the catalytic activity.[23] In addition, Glu739 forms an H-bond with His686 (see PDB: 1JKY).[2] Although no mutagenesis data are available for this residue, this interaction may still play a role because of its closeness to the Zn coordination sphere.

Structural information on the Michaelis complex may help to develop novel peptidomimetic inhibitors with therapeutical properties.

Here we provide a structural model of the complex between LF and an optimized substrate featuring the consensus sequence /VYPYPMEPT/ by using several computational tools. Density Functional Theory (DFT), all-atom Molecular Dynamics (MD) simulations, and mixed DFT-molecular mechanics (DFT/MM) approaches have been employed to study the structural and electronic properties of the Michaelis complex, while calculations based on Coarse-Grained (CG) models[26−28] and bioinformatics approaches[29−32] have been employed to verify the secondary structure stability of part of the model.

In all known Zn catalytic sites, a solvent species is always a Zn-ligand;[21] however, the nucleophilic agent (e.g., water or OH) can vary in concordance with the Zn environment which can even result in different reaction mechanisms.[33] In agreement with experimental and computational evidences obtained on related zinc metalloproteases,[33b,c] our calculations show that the nucleophilic agent for the LF catalyzed hydrolysis is a water molecule (not an OH group) and that an ionized Glu687 H-bonds the putative catalytic water. However, our study does not give a definitive answer to the LF catalytic mechanism. Substrate binding to the active site is mainly stabilized by van der Waals (vdW) interactions, with Tyr-P1′ and Tyr-P2 constituting the main anchors. The large scale motions of the enzyme do not affect the active site residues, and therefore it is unlikely that they play a mechanical role for the enzymatic reaction. Only a small polarization is exerted by the protein scaffold on the residues involved in the enzymatic reaction, as has been suggested for other proteases.[34] Finally, our results indicate that at least part of helix α19 in domain **III** (Figure 1a) actually assumes a coiled conformation; this is consistent with the unusually large temperature factors reported for this region.[17]

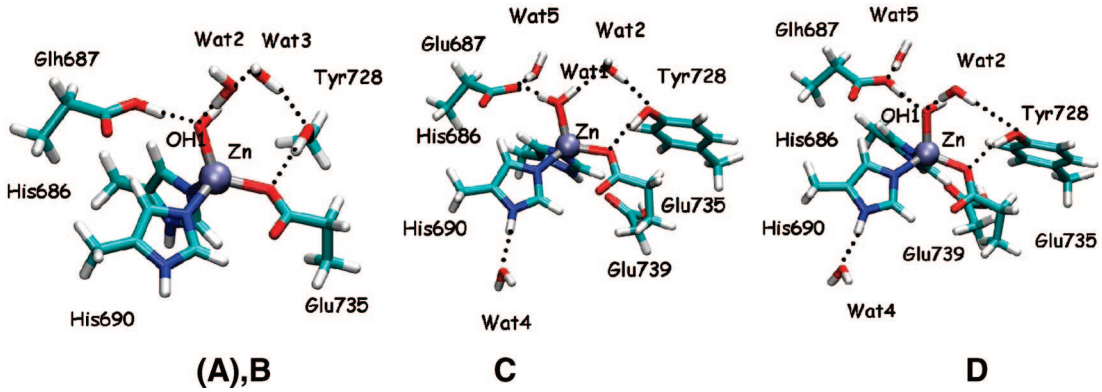



***Figure 2.*** Protomers of LF active site considered in this work: model **A** and **B** (60 atoms) and models **C** and **D** (84 atoms). Model **A** (not shown) was unstable; during geometry optimization it was rapidly transformed into model **B**, bearing a protonated Glu687 (labeled as Glh687 in the figure) and OH1 instead of Wat1. Models **C** and **D** turned out to have the same energetic stability within the accuracy of DFT.

## 2. Methods

**2.1. Construction of the LF Michaelis Complex.** The determination of the protonation state of LF active site residues is nontrivial. Here, using the same procedure as that of ref 22b, we perform DFT calculations on a series of models (**A-D**, Figure 2) based on the X-ray structure of the free enzyme (PDB: 1J7N).[17] Models **A** and **B** (60 atoms) included the following: (i) the zinc atom and its ligands: His686 and His690, Glu735 (all cut at C$\beta$ atom, and saturated with hydrogen atoms) and either a water molecule (Wat1, model **A**) or a hydroxide group (OH1, model **B**); (ii) Glu687, either in the ionized state (model **A**) or protonated in O$\epsilon$2 (model **B**); this residue, which forms an H-bond with the nucleophile, was also cut at C$\beta$; (iii) the water molecule Wat2, detected in the X-ray structure, which H-bonds to Wat1; (iv) the water molecule Wat3, which bridges Wat2 and Tyr728 (Figure 2); and (v) Tyr728 (modeled as methanol), which H-bonds to Glu735.

Models **C** and **D** (84 atoms) included the same groups as **A** and **B**, respectively, as well as (i) Glu739 (cut at the C$\gamma$), which H-bonds to His686; (ii) the crystallographic water molecules Wat4 and Wat5, which H-bond to His690 and Glu687, respectively; and (iii) the aromatic ring Tyr728 (cut at C$\beta$), which H-bonds to Wat2 and Glu735.

The DFT calculations were performed using the program CPMD[35] with a plane waves (PW) basis set up to an energy cutoff of 70 Ry. Core/valence interactions were described using norm conserving pseudopotentials of the Martins-Troullier type.[36] Integration of the nonlocal parts of the pseudopotential was obtained via the Kleinman-Bylander scheme[37] for all of the atoms except zinc, for which a Gauss-Hermite numerical integration scheme was used. The gradient corrected Becke exchange functional and the Lee−Yang−Parr correlation functional (BLYP) were used.[38,39] Periodic boundary conditions were applied, and we used orthorhombic cells with edges $a = 16.0$ Å, $b = 19.2$ Å, and $c = 12.8$ Å for models **A** and **B** and $a = b = 17.0$ Å, and $c = 14.5$ Å for models **C** and **D**. Isolated system conditions were applied.[40]

The Michaelis complex was built by reproducing the protonation state resulting from DFT calculations on the X-ray structure of the LF mutant E687C (which is unable to perform catalysis)[18] in complex with an optimized substrate featuring the MAPKK consensus sequence /VYPYPMEPT/ around the scissible bond (PDB: 1PWW).[18] The wild type enzyme was constructed by replacing Cys687 with a glutamic residue and by adding residues 346−367, missing in this X-ray structure, in the same conformation ($\alpha$-helix) as they were found in the only X-ray structure of LF which provides their positions (PDB: 1JKY). The histidines located outside the LF active site were protonated in N$\pi$, with the exception of His35, His91, His229, His277, His309, and His588, that were protonated in N$\tau$. [The IUPAC-IUB nomenclature was used.[41] N$\pi$ and N$\tau$ correspond to Nd and Ne, respectively, in the PDB atom naming system currently in use.] The two protomers were neutralized by adding 22 potassium counterions immersed in a box of 128.5, 81.3, and 94.0 Å, containing ~26,700 water molecules. The total size of the systems was ~92,600 atoms.

In the MAPKK consensus sequence /VYPYPMEPT/, the conserved reactive proline (i.e., the one placed between two tyrosines) was labeled as P1. Residues on the left side of P1 were labeled as P2, P3... to Pn, while residues on the right side were labeled as P1′, P2′... to Pn′.[18]

The structural characteristics of the Michaelis complex were studied using all-atom MD simulations, coarse-grained methods,[26−28] and bioinformatics tools;[29−32] special attention was given to the structural stability of the reconstructed region $\alpha$19 (residues 346−367). In addition, we investigated the electrostatic properties of the Michaelis complex using both the Poisson−Boltzmann approach and mixed DFT/MM calculations.

**2.2. All-Atom MD Simulations.** The AMBER parm98[42] force field was adopted for the substrate, the potassium counterions, and all the enzyme residues with the only exception of the zinc coordination sphere. For the parametrization of the latter we followed the procedure of ref 43 (see theSupporting Information, section 4).

The electrostatic interactions were evaluated using the Particle Mesh Ewald (PME) method.[44] A cutoff of 10 Å was used for the van der Waals interactions and the real part of the electrostatics interactions. The bonds involving hydrogen

atoms were kept fixed using the SHAKE algorithm.[45] A time step of 2 fs was used. The initial structures were relaxed by short minimization runs of 2000 steps using the conjugate gradient energy minimization algorithm. 100 ps of MD at constant volume were then performed during which the system was gradually heated to 300 K. Constant temperature (300 K) and pressure (1 atm) production runs were performed by coupling the systems to a Berendsen thermostat and barostat.[46] The NAMD simulation software was used.[47] A trajectory of 50 ns was computed, and the following properties were calculated:

*(a) RMSD/RMSF.* Root Mean Squared Deviations (RMSD) and Root Mean Squared Fluctuations (RMSF) of the Cα atoms were calculated from the all-atom MD trajectory. The structural stability of the complex during the simulations was monitored by using the RMSD; the RMSF were compared with the temperature factors from the X-ray structure (PDB: 1JKY).[2] The RMSF and the X-ray temperature factors were normalized to compare them (normalized B-values).

*(b) Principal Component Analysis (PCA).* Large scale motions were calculated as eigenvectors of the covariance matrix of Cα fluctuations, constructed from PCA. The Dynatraj program[48] was used to perform PCA on the last 15 ns of the all-atom MD simulation. For the first three principal components, rigid domains and hinges were identified using the scheme developed by Wriggers and Schulten.[49] Details of this calculation are reported in the Supporting Information, section 6.

*(c) Interaction Energies.* The interaction energies between LF and the optimized substrate were calculated over the last 15 ns of the all-atom MD simulation. The NAMD Energy plugin (v 1.0) from VMD (v 1.8.6)[50] was used to rerun NAMD[47] on the trajectory to calculate these energies. As the energy values extracted are ultimately dependent on the force field used (AMBER parm98[42]), these calculations are necessarily approximated, and our results are expressed as relative energies (i.e., normalized with respect to the highest interaction energy[51]).

**2.3. Hybrid Coarse-Grained/Molecular Mechanics (CG/MM) Simulations.** The structural instability of helix α19 (*vide infra*) imposed the employment of a variety of computational techniques to verify our findings obtained from the atomistic simulation.

In this respect in the CG/MM approach we treated helix α19 (residues 346−367) and protein or solvent atoms within 12.5 Å from α19 with an all-atom force field (MM region); the rest of the protein was treated with the Go simplified potential[52] (CG region).[26] The effect of the solvent outside the MM region was considered as the sum of stochastic and frictional forces proportional to the mass parm98[42] and Gromos96 43a1[54] force fields were performed; each of them started from a snapshot taken at 3 ns of the all-atom MD trajectory. These simulations were preceded by 1000 steps of energy minimization (using the steepest descend algorithm) followed by a gentle heating of the systems from 0 K to 300 K in 500 ps. From these simulations, the normalized Cα RMSFs were estimated and compared with the normalized temperature factors from the X-ray structure (PDB: 1JKY).[17] To ensure the reproducibility of these results, additional CG/MM simulations, starting from different initial structures, were computed (see the Supporting Information, section 5). These structures were selected from the equilibrated (last 15 ns) part of the all-atom MD trajectory (*vide infra*).

**2.4. Normal Mode Analysis (NMA).** NMA was performed with the NOMAD-ref server[55] on the energy minimized structure taken from the last frame of the all-atom MD simulation. In this scheme,[27,28] the protein was represented by a network of beads connected by harmonic springs; only the interactions between beads separated by a distance ≤3 Å were considered.[28] Normalized Cα RMSF were estimated and compared with normalized temperature factors from the X-ray structure (PDB: 1JKY),[2] after normalization of both terms. To ensure the reproducibility of these results, additional NMA calculations, starting from different initial structures, were performed (see the Supporting Information, section 6).

**2.5. Bioinformatics.** We investigated the propensity for disorder of helix α19 by using several prediction programs: e.g., PredictProtein,[29] PSIPRED,[30] SPRITZ,[31] and HNN[32] (see the Supporting Information, section 7).

**2.6. Electrostatics.** *(a) Poisson−Boltzmann Calculations.* Electrostatic surface potentials for LF+substrate adduct were calculated by solving the Poisson−Boltzmann equation with the APBS[56] and PDB2PQR[57] programs; the results were visualized using a PYMOL interface. These calculations were made on the energy minimized structure taken from the last frame of the all-atom MD trajectory featuring the protonation state of model C in the active site (see "Construction of the LF Michaelis complex" in the Methods section).

*(b) Polarization of the Active Site.* The polarization of selected chemical bonds in the active site was investigated using the so-called Bond Ionicity (BI) indexes[58] that can be estimated from DFT/MM[59] calculations. The region treated at the DFT level comprises the Zn, Glu687, Glu735, His686, and His690 (all cut at Cβ), the putative catalytic water (Wat1), and the scissible region of the optimized substrate which is formed by the backbone atoms from Tyr-P2 and Tyr-P1′ (all cut at Cα) and all the atoms from Pro-P1 (see the Supporting Information, Figure S2). The rest of the system was treated with the AMBER parm98[42] force field. We considered active site residues under different environments: e.g., *in vacuo*, and with the influence of the solvent and/or LF electrostatic fields.[60] To construct the models we used 15 equally spaced frames from the last 15 ns of the all-atom MD trajectory.

$BI_{AB}$ of a bond between atoms A and B is defined as

$$BI_{AB} = \frac{d_A}{d_{AB}} \tag{1}$$

where $d_A$ is the distance between atom $A$ and the Boys orbitals[58] along the $AB$ bond, and $d_{AB}$ is the length of the bond between $A$ and $B$. A value of BI = 0.5 (the Boys orbital is in the middle of the bond) indicates absence of polarization; while values close to 1 or 0 indicate polarization.

## 3. Results and Discussion

The purpose of this work was the characterization of the structural and electrostatic properties of the Michaelis complex formed by LF and an optimized substrate that features the MAPKK consensus sequence /VYPYPMEPT/.[18] To the best of our knowledge this is the first complete structural model of a Zn-bound LF in complex with a scissible substrate. The first step to achieve our goal was determining the correct protonation state in LF active site residues using DFT calculations.

**3.1. Protonation State at the Active Site.** As mentioned in the Introduction, a critical issue in Zn-based hydrolases is the determination of the protonation states of residues in the active site.[22] In this work we addressed this issue by performing DFT calculations on increasingly complex models of the active site (Figure 2). The smallest models (**A** and **B**) included only the Zn site, while the largest models (**C** and **D**) included additional second-shell ligands of established (e.g., Tyr728)[23] or putative (e.g., Glu739)[2] relevance for the enzymatic reaction (see the Methods section for details). The most likely protomers were defined as those associated with the lowest potential energy and with the lowest RMSD relative to the reference X-ray structure (i.e., LF in the free state, PDB code: 1J7N).[17]

In the smallest models, **A** featured Glu687 in the ionized state and Wat1 as nucleophile, while **B** exhibited Glu687 in its neutral state and the nucleophile was a hydroxide group. During the geometry optimization model **A** was unstable, as Wat1 transferred a proton to O$\epsilon$2@Glu687, resulting in model **B** (see Figure 2). The latter was instead stable and featured a slightly distorted tetrahedral coordination geometry and establishing the H-bonds b(O$\epsilon$2@Glu687, O@OH1) and b(O$\eta$@Tyr728, O$\epsilon$2@Glu735), also putatively present in the X-ray structure (Table 1). The distances Zn-X (X=coordinating atom), decreased by $\sim$0.1−0.3 Å relative to the X-ray structure (Table 1); the RMSD between model **B** and the X-ray structure was sizable ($\sim$0.55 Å); both effects were possibly caused by the limited size of the model.

In the largest models, **C** featured Glu687 in the ionized state and Wat1 as nucleophile (like **A**), while **D** exhibited Glu687 in its neutral state and the nucleophile was a hydroxide group (like **B**). Both models (**C** and **D**) were stable. Some general trends in models **C** and **D** with respect to model **B** could be identified: (i) the bonds b(Zn, N$\tau$@His686) were shorter ($\Delta$d=−0.11 for **C** and $\Delta$d=−0.10 Å for **D**), (ii) one bond b(Zn, N$\tau$@His690) was longer in **D** ($\Delta$d=+0.03Å) and unaltered in **C**, (iii) the bonds b(Zn, O@[OH1,Wat1]) were larger ($\Delta$d=+0.08Å for **C** and $\Delta$d=+0.05 Å for **D**), and (iv) the hydrogen bonds between Glu687 and the nucleophile b(O$\epsilon$2@Glu687, O@[OH1,Wat1]) were longer ($\Delta$d=+0.09Å for **C** and $\Delta$d=+0.01Å for **D**). On the other hand, the hydrogen bond between Wat5 and Glu687 (not included in the smallest models) was different between **C** and **D** (Table 1).

The RMSD of **C** is smaller than **D**; however, due to the small relative energies of the two protomers ($\Delta E \sim 2$ kcal/mol), which were not significant with respect to the accuracy of DFT calculations, we could not establish with certainty which is the most likely protonation state of the active site.[61]

***Table 1.*** Comparison between Experimental (X-ray; PDB: 1J7N)[17] and Calculated (DFT) Structural Parameters (Bond Lengths (Å) and Angles (deg)) for the Models of the Free Enzyme Used in These Calculations (**B**-**D**, Figure *2*)[a]

| | X-ray | B | C | D |
|---|---|---|---|---|
| Bond Length (b($A_i$,$B_j$); in Å) | | | | |
| b(Zn, O@[OH1,Wat1]) | 2.1 | 1.96 | 2.04 | 2.01 |
| b(Zn, N$\tau$@His690) | 2.1 | 2.04 | 2.04 | 2.07 |
| b(Zn, N$\tau$@His686) | 2.3 | 2.20 | 2.09 | 2.10 |
| b(Zn, O$\epsilon$2@Glu735) | 2.3 | 2.00 | 2.01 | 2.03 |
| b(O$\epsilon$2@Glu687, O@[OH1,Wat1]) | 3.6 | 2.57 | 2.66 | 2.58 |
| b(O$\eta$@Tyr728, O$\epsilon$2@Glu735) | 2.7 | 2.70 | 2.94 | 2.75 |
| b(N$\pi$@His690, O@Wat4) | 2.9 | | 2.97 | 2.97 |
| b(N$\pi$@His686, O$\epsilon$1@Glu739) | 2.9 | | 2.93 | 2.88 |
| b(O$\epsilon$2@Glu687, O@Wat5) | 2.9 | | 2.85 | 2.98 |
| Angles ($\tau$($A_i$, $B_j$, $C_k$), in deg) | | | | |
| $\tau$(O$\epsilon$1@Glu735, Zn, O@[Wat1, OH1]) | 96 | 121 | 116 | 113 |
| $\tau$(N$\tau$@His686, Zn, O@[Wat1,OH1]) | 107 | 102 | 105 | 102 |
| $\tau$(N$\tau$@His690, Zn, O@[Wat1,OH1]) | 128 | 110 | 105 | 105 |
| RMSD (Å) | | 0.55 | 0.36 | 0.58 |

[a] To facilitate the comparison with X-ray data, only the measures involving heavy atoms were used. The RMSD (in Å) with respect to the X-ray structure is also given.

***Table 2.*** Distances between Selected Pairs of Atoms in the LF Active Site, including (i) Zn-Coordination Bonds (*e.g.*, b(Zn, O$\epsilon$2@Glu735) and (ii) Hydrogen Bonds (*e.g.*, b(H@Wat1, O$\epsilon$2@Glu687) and Key Geometrical Features (*e.g.*, d(C$\delta$@Glu735, C$\delta$@Glu687) of the Active Site[a]

| pairs of atoms | distance (Å) |
|---|---|
| b(Zn, O$\epsilon$2@Glu735) | 2.22 (0.07) |
| d(Zn, O$\epsilon$2@Glu687) | 4.8 (0.2) |
| b(Zn, N$\tau$@His686) | 2.05 (0.05) |
| b(Zn, N$\tau$@His690) | 2.15 (0.05) |
| b(Zn, O@Wat1) | 1.98 (0.06) |
| d(C$\delta$@Glu735, C$\delta$@Glu687) | 6.9 (0.2) |
| d(N$\tau$@His686, N$\tau$@His690) | 2.9 (0.1) |
| b(H@Wat1, O$\epsilon$2@Glu687) | 1.9 (0.2) |
| b(O′@Tyr-P2, H@Wat1) | 1.8 (0.2) |
| b(H$\eta$@Tyr728, O′@Pro-P1) | 1.8 (0.2) |
| b(H$\pi$@His686, O$\epsilon$1@Glu739) | 1.97(0.2) |

[a] The distances were calculated during the last 15 ns of the all-atom MD simulation. Standard deviations are given in parentheses.

**3.2. Molecular Dynamics of LF Michaelis Complexes.** Next, we built two Michaelis complexes using protonation states **C** and **D** for the active site (see the Methods section), and we performed all-atom MD simulations on both Michaelis complexes.

*3.2.1. MD of LF Michaelis Complex in the C Protonation State.* This complex was stable during the entire simulation (50 ns). In particular, the bond lengths in the coordination sphere had small fluctuations around their average positions during the dynamics (Table 2), and the substrate remained in its binding site for the entire simulation time. The structure of the rest of the protein was also

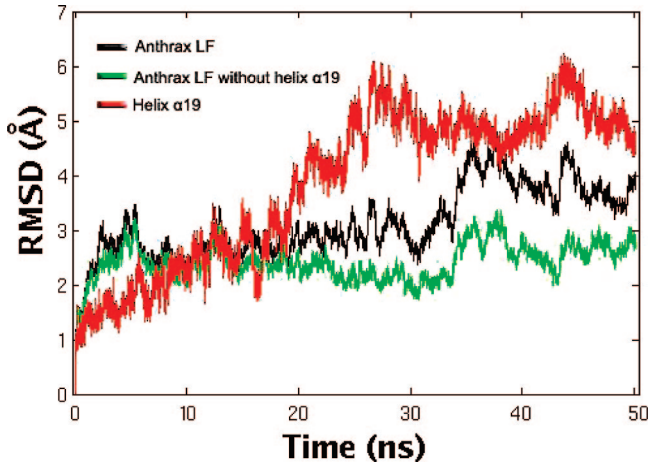



***Figure 3.*** RMSD of LF Cα atoms during the 50 ns all-atom MD simulation. Note the increase in RMSD for helix α19 during the first 25 ns; after this time α19 gradually achieved structural stability.

maintained except for helix α19 (residues 346−367), which became partially unfolded within the first 25 ns of the MD trajectory (Figure 3). This unfolding was localized in the second part of α19 (residues 361−367) and caused a sudden increase in the RMSD of the Cα atoms (Figure 3); the coil-like conformation of the second part of α19 was then maintained until the end of the simulation. [The structure of the complex (LF+substrate) after 50 ns of all-atom MD is available at http://people.sissa.it/~hong/projects__f iles/50ns_last_fr.pdb.]

To gain further insight into the instability of helix α19, it was necessary to carry out calculations with two types of coarse-grained methods and with disorder prediction servers:

*(a) Hybrid Coarse-Grained/Molecular-Mechanics (CG/MM) Simulations*. Here, α19 and nearby atoms were treated at the MM level, while the rest of the system was treated at the CG level (see the Methods section). Two 60 ns CG/MM simulations using AMBER parm98[42] and Gromos96 43a1[54] force fields for the MM part were performed. This computationally efficient method allows for the verification of the findings of the all atom MD simulation and for checking the dependence of these results on the employed force field. In both CG/MM simulations, α19 partially unfolds within the first ~25 ns (see the Supporting Information, Figure S5). The RMSF of the Cα atoms from the two 60 ns trajectories (Figure 4) were larger for α19 than for the rest of the protein. Since the all-atom MD trajectory reaches a fairly constant RMSD only after 25 ns, we have checked the reproducibility of these results by performing CG/MM simulations starting from 15 equally spaced frames taken from the last 15 ns of the all-atom MD simulation (see the Supporting Information, section 5). The results obtained by these simulations, along with those obtained using different force fields, confirm that α19 unfolds in the first 25 ns of simulations independently of the computational technique and force field employed.

*(b) Normal Mode Analysis (NMA).* Here a CG elastic network of Cα atoms was built based on the energy minimized structure taken from the last frame of the all-atom MD trajectory. As obtained before for CG/MM MD simulations, the calculated RMSF of Cα atoms estimated

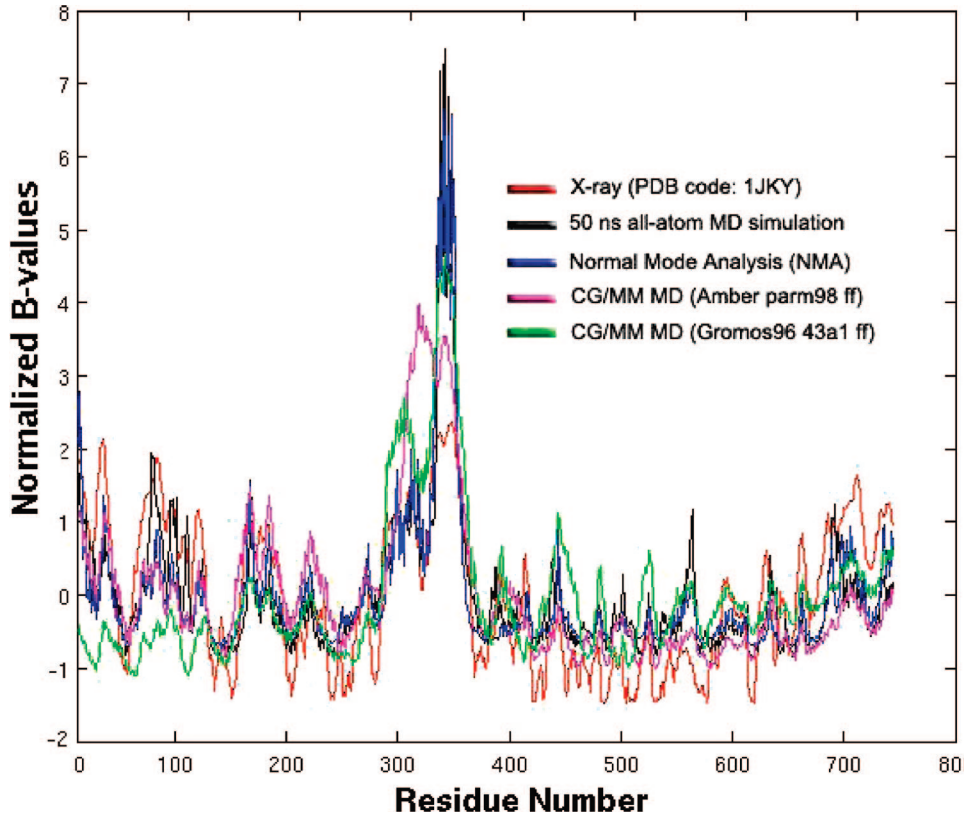



***Figure 4.*** Calculated (RMSF) and experimental (X-ray; (PDB: 1JKY)[17]) normalized B-values for LF. For the all-atom MD and CG/MM simulations, only the last 15 ns were considered.

from NMA for the α19 region were larger than those of the rest of the protein (Figure 4). To ensure the reproducibility of these results, NMA calculations were also performed using 15 equally spaced frames taken from the last 15 ns of the all-atom MD simulation (see the Supporting Information, section 6).

The normalized B-values calculated from all-atom MD, NMA, and CG/MM simulations (Figure 4) agree with those reported in the X-ray structure (PDB: 1JKY),[17] except for helix α19: part of the latter (residues 361−367) assumed a coiled conformation in aqueous solution, and its calculated normalized B-values were larger than those of the X-ray structure (Figure 4). This is consistent with the apparent difficulty to determine the solid state structure of α19, which, in fact, has only been resolved in the X-ray structure used here as the starting model for the α19 tract (PDB: 1JKY).[17]

*(c) Disorder Prediction.* To check the reliability of all-atom MD simulations showing the unfolding of a19 we also performed disorder prediction analysis using bioinformatics tools. The structural predictors PredictProtein,[29] PSIPRED,[30] SPRITZ,[31] and HNN[32] pointed out part of α19 as a disordered region. Particularly, the last segment of this region (residues 361−367) is more likely to be a loop than an α-helix (see Figure S9 in the Supporting Information, section 7).

Finally, we describe the structural and electrostatic features of the active site of the equilibrated Michaelis complex. Only the last 15 ns of all-atom MD simulation were used for this analysis since the RMSD of α19 became stable only at this simulation time.

At the active site, Glu735 acts as a monodentate ligand of Zn and H-bonds to the solvent, interacting on average with ∼0.7 water molecules, as obtained by integrating the radial distribution function of Oε2@Glu687 vs O@water (see the Supporting Information, Figure S4a). The Zn is bonded with two histidine residues [b(Zn, Nτ@His686)=2.05 ± 0.05 Å and b(Zn, Nτ@His690)=2.15 ± 0.05 Å].

The catalytic water molecule, Wat1, H-bonds to Glu687, which is deprotonated; Glu687 is believed to act as a general base during catalysis (i.e., accepting an hydrogen ion from Wat1).[17,62] Besides the H-bond to Wat1, Glu687 also interacts, on average, with ∼1 water molecule from the solvent. On the other hand, Wat1 also H-bonds to Tyr-P2 (Table 2); this interaction could help to orient Wat1 in a proper position to perform the hydrolysis of the substrate.

Tyr728 forms an H-bond with the reactive carbonyl group of proline substrate (O′@Pro-P1, Table 2) for ∼35% of the simulation time, orienting the substrate for the nucleophilic attack. This H-bond interaction might play a role for the catalysis, providing a plausible, yet speculative, explanation on why the Y728F mutant is not catalytically active.[23] Also Tyr728 and O′@Pro-P1 are exposed towards the solvent, and they H-bond, on average, to ∼2 and 1 water molecules, respectively.

The motions of Tyr728 (of the enzyme) and Tyr-P1′ (of the substrate) are correlated (see the Supporting Information, Figure S1b). The aromatic rings of these two residues lay at a distance smaller than 5 Å for approximately 53% of the simulation time, being that Tyr-P1′ is accommodated in the hydrophobic pocket (S1′) of the enzyme which is partially formed by Tyr728 (Figure 5b). The hydroxylic moiety of Tyr-P1′ also forms an H-bond with the carbonyl group in the Val675 backbone (Hη@Tyr-P1′...O′@Val675=1.8 ± 0.1 Å), and this interaction is maintained for ∼80% of the simulation time. The remaining 20% of the simulation time, Tyr-P1′ H-bonds with Glu739 (Hη@Tyr-P1′... Oε2@Glu739=1.9 ± 0.2 Å) in S1′. At the entrance of S1′ the backbone of Tyr-P1′ also forms two stable (∼90% of the simulation time) H-bonds with backbone atoms in LF; the first one, with Lys656 (O′@Tyr-P1′... H′@Lys656=1.9 ± 0.1 Å) and the second with Gly657 (H′@Tyr-P1′... O′@Gly657=2.0 ± 0.2 Å).

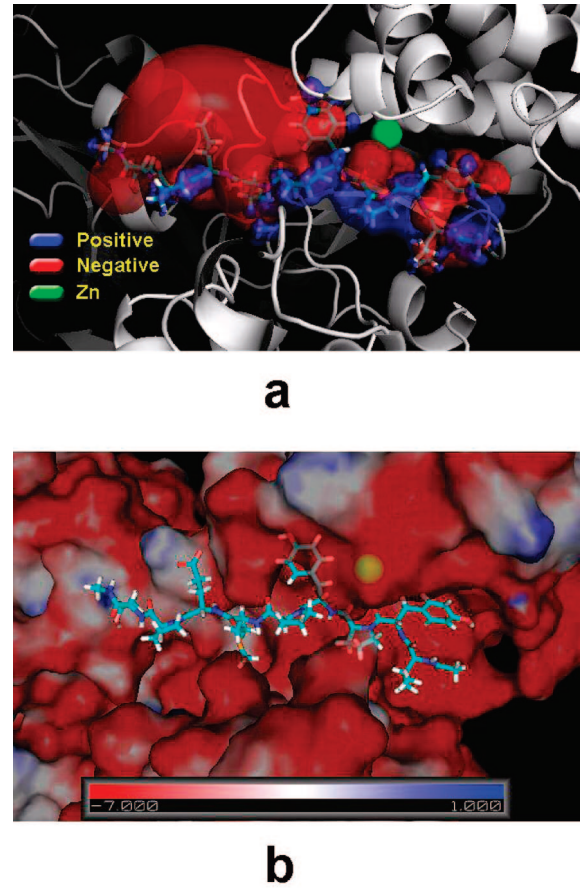



***Figure 5.*** Electrostatic isosurfaces on the Michaelis complex. (**a**) Optimized MAPKK2-like substrate. Note the complementarity of the bilobular (created by O′@Pro-P1 and O′@Tyr-P1′) around the Zn ion, while the larger negative patch (form by Glu-P4′) is outside the negative groove of LF. (**b**) LF active site groove. Note the tyrosine on the left side of the Zn ion (Tyr-P1′) inserted in the hydrophobic pocket (S1′) of the enzyme.

Poisson−Boltzmann calculations suggest that the substrate fits in the groove of the enzyme forming complementary electrostatic interactions (Figure 5). The Zn ion interacts between a bilobular negative patch on the substrate (formed by O′@Pro-P1 and O′@Tyr-P1′).

To dissect the single contributions of the substrate residues in binding to LF, we performed a calculation of substrate/LF interaction energies based on the AMBER parm98[42] force field. Such calculations are necessarily approximate, and they are used here only for qualitative comparisons; therefore,

***Table 3.*** LF Positional Selectivity[18] and Relative Interaction Energies (Electrostatics, van der Waals (vdW), and Total) between Individual Residues of the Substrate and LF, Calculated for the Last 15 ns of the All-Atom MD Simulation (Model **C**)[a]

| | Val | Tyr | Pro | Tyr | Pro | Met | Glu |
|---|---|---|---|---|---|---|---|
| | P3 | P2 | P1 | P1′ | P2′ | P3′ | P4′ |
| selectivity | 1.5 | **3.1** | - | **3.0** | 1.9 | 1.3 | 1.6 |
| total energy | (−0.26 ± 0.05) | **(−0.45 ± 0.09)** | (−0.1 ± 0.1) | **(−1.0 ± 0.1)** | (−0.48 ± 0.05) | (−0.43 ± 0.07) | (0 ± 0.5) |
| electrostatic energy | (−0.02 ± 0.02) | (−0.07 ± 0.07) | (0.1 ± 0.1) | **(−0.5 ± 0.1)** | (−0.24 ± 0.05) | (−0.09 ± 0.05) | (0.2 ± 0.6) |
| vdW energy | (−0.21 ± 0.05) | **(−0.36 ± 0.02)** | (−0.19 ± 0.05) | (−0.43 ± 0.05) | (−0.21 ± 0.02) | (−0.33 ± 0.05) | (−0.16 ± 0.05) |

[a] Because these values are based on the AMBER parm98[42] force field, they are used here only for qualitative comparisons. The energy values were normalized to the largest absolute value as in ref 51.

***Table 4.*** Comparison between Bond Ionicity Indexes (BIs)[58,59] in the Reactive Bonds of LF under Different Electrostatic Conditions Showing the Effect of the Protein and Solvent Electrostatic Properties on the Polarization of the Active Site[a]

| | charged | no charge in protein | no charge in solvent | no charge |
|---|---|---|---|---|
| $d(N_{pep}\text{-}BO^1)/d(N_{pep}\text{-}C_{pep})$ | 0.37(0.02) | 0.38(0.02) | 0.37(0.02) | 0.37(0.02) |
| $d(N_{pep}\text{-}BO^2)/d(N_{pep}\text{-}C_{pep})$ | 0.35(0.02) | 0.36(0.02) | 0.35(0.02) | 0.36(0.02) |
| $d(O_{pep}\text{-}BO^1{}_{lone})$ | 0.32(0.01) | 0.31(0.01) | 0.32(0.01) | 0.32(0.01) |
| $d(O_{pep}\text{-}BO^2{}_{lone})$ | 0.31(0.01) | 0.31(0.01) | 0.30(0.01) | 0.30(0.01) |
| $d(O_{pep}\text{-}BO^1{}_{C=O})/d(O_{pep}\text{-}C_{pep})$ | 0.39(0.01) | 0.40(0.01) | 0.40(0.01) | 0.40(0.01) |
| $d(O_{pep}\text{-}BO^2{}_{C=O})/d(O_{pep}\text{-}C_{pep})$ | 0.39(0.01) | 0.39(0.03) | 0.40(0.01) | 0.40(0.01) |
| $d(O_{Wat}\text{-}BO^1{}_{lone})$ | 0.28(0.04) | 0.28(0.05) | 0.28(0.02) | 0.27(0.02) |
| $d(O_{Wat}\text{-}BO^2{}_{lone})$ | 0.43(0.06) | 0.47(0.05) | 0.43(0.04) | 0.45(0.04) |
| $d(O_{Wat}\text{-}BO_{O-H1})/d(O_{Wat}\text{-}H_{Wat})$ | 0.50(0.01) | 0.49(0.01) | 0.50(0.01) | 0.49(0.01) |
| $d(O_{Wat}\text{-}BO_{O-H2})/d(O_{Wat}\text{-}H_{Wat})$ | 0.51(0.01) | 0.51(0.01) | 0.51(0.01) | 0.51(0.01) |

[a] Four conditions were evaluated: (i) "**charged**", calculation assigning charges to all atoms in the system; (ii) "**no charge in protein**", calculation assigning charges equal to zero in all atoms of the protein; (iii) "**no charge in solvent**", calculation assigning charges equal to zero to the atoms of all the water molecules; and (iv) "**no charge**", calculation assigning charges equal to zero to all atoms in the system. Atomic charges of the MM atoms were assigned using the AMBER parm98 force field,[42] while active site atoms were treated at the DFT level.[59] DFT/MM calculations were performed on 15 equally spaced frames taken from the last (equilibrated) 15 ns of the all-atom MD trajectory. Standard deviations are given in parentheses.

only normalized values[51] of the interaction energies are reported (Table 3).

The substrate's residues with the largest LF positional-dependent selectivity,[18] namely Tyr-P2, Tyr-P1′, and Pro-P2′, featured the strongest interaction energies with LF. Indeed, Tyr-P1′ is the main anchor for substrate binding forming significant vdW interactions (mainly provided by Try-P1′ aromatic ring and Tyr728, His686, and Leu677 side chains) as well as H-bonding and electrostatic interactions, mostly with Val675, Glu739, Lys656, and Gly657. Tyr-P2 plays also a significant role for the binding, forming vdW interactions with His690, Tyr659, Leu658, Pro661, and Ala734 (Table 3).

Although in the present work we do not perform any reactivity study, the polarization of the active site residues induced by the electrostatic properties of the protein environment may provide some information on the Michaelis complex, which is an important species in the catalytic cycle of the enzyme. Consequently, we decided to consider the polarization effect of different electrostatic environments (e.g., *in vacuo*, and with the influence of the solvent and/or LF electrostatic fields) on the putative reactive bonds of the Michaelis complex.[63] Our results (Table 4) show that the environment does not have a marked role in polarizing the active site residues, with the only exception of the nucleophilic water, which trivially, upon coordination to Zn, undergoes a polarization of the lone pairs. The polarization of the bonds involved in the LF-mediated substrate cleavage can be quantitatively compared with homologous reactive bonds of two other proteases, for which calculations have been carried out with an identical setup. These are the aspartyl protease from human immunodeficiency virus of type 1 (HIV-1 PR),[63,64] which is believed to use a water molecule for the hydrolysis,[64] and the serine protease furin,[63] which uses the hydroxylic group from a serine residue as a

***Table 5.*** Comparison between Bond Ionicity Indexes (BIs)[58,59] in the Reactive Bonds of Three Prototypical Proteases: Furin, HIV-1 PR, and LF[a]

| | furin | HIV-1 PR | anthrax LF |
|---|---|---|---|
| $d(N_{pep}\text{-}BO^1)/d(N_{pep}\text{-}C_{pep})$ | 0.33(0.02) | 0.31(0.03) | 0.37(0.02) |
| $d(N_{pep}\text{-}BO^2)/d(N_{pep}\text{-}C_{pep})$ | 0.38(0.01) | 0.38(0.02) | 0.35(0.02) |
| $d(O_{pep}\text{-}BO^1{}_{lone})$ | 0.33(0.01) | 0.34(0.01) | 0.32(0.01) |
| $d(O_{pep}\text{-}BO^2{}_{lone})$ | 0.31(0.01) | 0.32(0.01) | 0.31(0.01) |
| $d(O_{pep}\text{-}BO^1{}_{C=O})/d(O_{pep}\text{-}C_{pep})$ | 0.37(0.01) | 0.38(0.01) | 0.39(0.01) |
| $d(O_{pep}\text{-}BO^2{}_{C=O})/d(O_{pep}\text{-}C_{pep})$ | 0.38(0.01) | 0.39(0.01) | 0.39(0.01) |
| $d(O_{Hyd}\text{-}BO^1{}_{lone})$ | 0.31(0.01) | | |
| $d(O_{Hyd}\text{-}BO^2{}_{lone})$ | 0.31(0.01) | | |
| $d(O_{Hyd}\text{-}BO_{O-H})/d(O_{Hyd}\text{-}H_{Hyd})$ | 0.50(0.01) | | |
| $d(O_{Hyd}\text{-}BO_{C-O})/d(O_{Hyd}\text{-}C)$ | 0.39(0.01) | | |
| $d(O_{Wat}\text{-}BO^1{}_{lone})$ | | 0.32(0.31) | 0.28(0.04) |
| $d(O_{Wat}\text{-}BO^2{}_{lone})$ | | 0.33(0.01) | 0.43(0.06) |
| $d(O_{Wat}\text{-}BO_{O-H1})/d(O_{Wat}\text{-}H_{Wat})$ | | 0.47(0.02) | 0.50(0.01) |
| $d(O_{Wat}\text{-}BO_{O-H2})/d(O_{Wat}\text{-}H_{Wat})$ | | 0.53(0.02) | 0.51(0.01) |

[a] The location of the Boys Orbitals (BOs) for LF is given in Figures S2 and S3 of the Supporting Information. The calculations were performed on 15 equally spaced frames taken from the last (equilibrated) 15 ns of the all-atom MD trajectory. Standard deviations are given in parentheses. $X_{pep}$ refers to the atom type X from the peptidic bond which undergoes nucleophilic attack. $X_{Wat}$ refers to the atom type X from the nucleophilic water molecule. $X_{Hyd}$ refers to the atom type X from the nucleophilic hydroxilic group (belonging to the side chain of Ser).

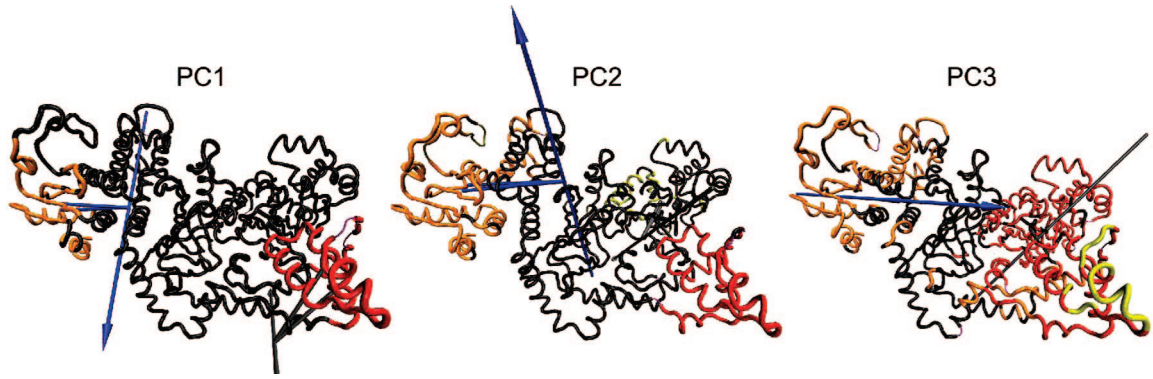



***Figure 6.*** The first three Principal Components ($PC_s$) as calculated with Dynatraj[48] from the last 15 ns of the all-atom MD simulation. Rigid domains, as calculated with the procedure of Wriggers and Schulten,[49] are depicted in different colors; details of these calculations can be found in the Supporting Information, section 6. The arrows indicate the effective rotation axis between the two adjacent rigid domains. The direction of the arrow represents the sense (e.g., clockwise) used to find the rotations axis between two rigid domains, and it is ultimately dependent on the choice of the "reference" rigid domain (the one kept steady during the calculation of the rotational angle). Note that PC3 was able to capture a movement in the α19 region (PC3, yellow), pointing to the higher flexibility of this region (see Figure 4).

nucleophile. As a measure of polarization we used the Bond Ionicity indexes[58] (see the section "Electrostatics" in the Methods section).

Based on the values of the BIs of $C_{pep}$═$O_{pep}$ and $N_{pep}$-$C_{pep}$ bonds in the substrate (Table 5 and Figure S3), we conclude that substrate's reactive bonds in LF are less polarized than those in HIV-1 PR and furin. In addition, the water (Wat1) O−H bonds were also less polarized in LF than in HIV-1 PR: (i) in LF, $BI_{O-H1}$@Wat1(LF) = 0.50 ± 0.01, $BI_{O-H2}$@Wat1(LF) = 0.51 ± 0.01 and (ii) in HIV-1 PR, $BI_{O-H1}$@Wat1(HIV-1 PR) = 0.47 ± 0.02 and $BI_{O-H2}$@Wat1(HIV-1 PR) = 0.53 ± 0.02. However, the lone pairs on the oxygen atom of the catalytic water (represented as $d(O_{Wat}\text{-}BO^{1[2]}{}_{lone})$ in Table 5) were more asymmetric in LF than in HIV-1 PR simply due to the coordination of the water to Zn.

Thus, although our analysis provides no information on the polarization of the transition state or of other species of the catalytic cycle (which were not investigated in this study), we do suggest here that small polarization effects on the substrate are induced by the LF scaffold, while the presence of zinc has a more critical role for the polarization of the reactants (Tables 4 and 5).

The large scale motions may play a role for substrate recognition and/or for enzymatic catalysis.[34,63−66] We explore this issue by performing PCA[48] on the last (equilibrated) 15 ns of the all-atom MD trajectory. In the first three Principal Components (PCs), which account for ∼54% of the overall motion, we identified large motions involving domain **III** (i.e., the domain that includes helix α19, see Figure 1a). In particular, PC3 was able to capture a relatively independent movement of α19 with respect to domain **III** (Figure 6). The observed large scale motions of LF did not affect the active site (see Table 2), similar to what has been found in another protease studied with the same computational setup, the serine protease furin.[63] Therefore, we conclude that it is unlikely that the large scale motions of LF could play a mechanical role for the enzymatic activity, although we cannot exclude that they may indirectly affect catalysis by long-range electrostatics.

*3.2.2. MD of LF Michaelis Complex in D Protonation State.* This protonation state turned out to be already unstable in the first 0.5 ns of simulation: the H-bond network was disrupted because of a rotation of Glu687 along the Cγ-Cδ bond, allowing the entrance of additional water in the active site and the departure of the substrate from the active site. This complex was therefore discarded.

## 4. Conclusions

We characterized the structural properties of the Michaelis complex formed by anthrax LF and an optimized substrate using a variety of computational tools. Our findings can be summarized as follows:

(i) The second shell ligands affect the structural properties of the Zn active site as has been observed in other Zn-based enzymes.[22,24,67] Our calculations confirm that second shell ligands have an influence on the protonation state of the active site geometry, stabilizing the hydrogen bond network within the active site residues.

(ii) The nucleophilic agent is a Zn-bound water molecule (not an OH group) H-bonded to Glu687.

(iii) The calculated substrate per residue interaction energies with LF correlate with the experimentally derived LF positional selectivity.[18] In particular, Tyr-P1′ and Tyr-P2 constitute the main substrate anchors to the LF active site. In fact, Tyr-P1′ has the largest interaction energy with LF, with similar contributions from electrostatics (featured mainly by a H-bond established between Hη@Tyr-P1′ and the backbone of the enzyme (O′@Val675)) and vdW interactions (with Tyr-P1′ interacting with the side chains of Tyr728, His686, and Leu677). In contrast, Tyr-P2 exhibits large vdW interactions principally with the Zn-bound His690. These results may provide a rationale to explain the selectivity of LF for substrates with tyrosines in the vicinity of the reactive proline.[18]

(iv) The LF scaffold induces small polarization effects on the substrate. A larger polarization is observed for the lone pairs of the nucleophilic agent Wat1 (see (ii)); the obvious cause for this effect is the Wat1 coordination to Zn.

(v) Large-scale motions do not affect the structure of the LF active site; it is therefore unlikely that these motions could play a mechanical role during the first step of the catalysis.

(vi) Part of helix α19 (residues 361−367) assumes a coil-like conformation in aqueous solution. This behavior was qualitatively predicted by several bioinformatics tools and confirmed by several computational approaches.

**Acknowledgment.** We thank Dr. Marilisa Neri for help in the setup of the CG/MM simulations and Dr. Vincenzo Carnevale for useful discussions. We also thank the CINECA-INFM for grants for the computational resources used in this work.

**Supporting Information Available:** Calculations on LF correlated of motions; parametrization of the active site residues, results polarization; and hydration of the active site, additional simulations on hybrid Coarse-Grained/Molecular Mechanics (CG/MM) simulations, Principal Component Analysis (PCA), Normal Mode Analysis (NMA), and additional structural bioinformatics predictions. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Mock, M.; Mignot, T. Anthrax toxins and the host: a story of intimacy. *Cell. Microbiol.* **2003**, *5*, 15–23.

(2) Guidi-Rotani, C. The alveolar macrophage: The Trojan horse of *Bacillus anthracis*. *Trends. Microbiol.* **2002**, *10*, 405–409.

(3) Dixon, T. C.; Fadl, A. A.; Koheler, T. M.; Swanson, J. A.; Hanna, P. C. Early *Bacillus anthracis*-macrophage interactions: intracellular survival and escape. *Cell. Microbiol.* **2000**, *2*, 453–463.

(4) Friedlander, A. M. Anthrax: clinical features, pathogenesis, and potential biological warfare threat. *Curr. Clin. Top. Infect. Dis.* **2000**, *20*, 335–349.

(5) Shoop, W. L.; Xiong, Y.; Woods, A.; Guo, J.; Pivnichny, J. V.; Felcetto, T.; Michael, B. F.; Bansal, A.; Cummings, R. T.; Cinnungam, B. R.; Friedlander, A. M.; Douglas, C. M.; Patel, S. B.; Wisniewski, D.; Scapin, G.; Spaolowe, S. P.; Zaller, D. M.; Chapman, K. T.; Scolnick, E. M.; Schmatz, D. M.; Bartizal, K.; MacCoss, M.; Hermes, J. D. Anthrax lethal factor inhibition. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7958–7963.

(6) Pezard, C.; Berche, P.; Mock, M. Contribution of individual toxin components to virulence of *Bacillus anthracis*. *Infect. Immun.* **1991**, *59*, 3472–3477.

(7) Petosa, C.; Collier, R. J.; Klimpel, K. R.; Leppla, S. H.; Liddington, R. C. Crystal structure of the anthrax toxin protective antigen. *Nature* **1997**, *385*, 833–838.

(8) Montecucco, C.; Tonello, F.; Zanotti, G. Stop the killer: how to inhibit the anthrax lethal factor metalloprotease. *Trends. Biochem. Sci.* **2004**, *29*, 282–285.

(9) Abrami, L.; Reig, N.; van der Goot, F. G. Anthrax toxin: the long and winding road that leads to the kill. *Trends. Microbiol.* **2005**, *13*, 72–78.

(10) Bann, J. C.; Hultgreen, S. J. Structural Biology: Anthrax hijacks host receptor. *Nature* **2004**, *430*, 843–844.

(11) Duesbery, N. S.; Webb, C. P.; Leppla, S. H.; Gordon, V. M.; Klimpel, K. R.; Copeland, T. D.; Ahn, N. G.; Oskarsonn, M. K.; Fukasawa, K.; Paul, K. D.; Vande Woude, G. F. Proteolytic inactivation of MAP-Kinase-Kinase by anthrax lethal factor. *Science* **1998**, *280*, 734–737.

(12) Weston, C. R.; Lambright, D. G.; Davis, R. J. MAP Kinase signalling specificity. *Science* **2002**, *296*, 2345–2347.

(13) Friedlander, A. M. Macrophages are sensitive to anthrax lethal toxin through an acid-dependent process. *J. Biol. Chem.* **1986**, *261* (16), 7123–7126.

(14) Kirby, J. E. Anthrax lethal toxin induces human endothelial cell apoptosis. *Infect. Immun.* **2004**, *72* (1), 430–439.

(15) Milne, J. C.; Furlong, D.; Hanna, P. C.; Well, J. S.; Collier, R. J. Anthrax protective antigen forms oligomers during intoxication of mammalian cells. *J. Biol. Chem.* **1994**, *269*, 20607–20612.

(16) Guichard, A.; Park, J. M.; Cruz-Moreno, B.; Karin, M.; Bier, E. Anthrax lethal factor and edema factor act on conserved targets in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 3244–3249.

(17) Pannifer, A. D.; Wong, T. Y.; Schwarzenbacher, R.; Renatus, M.; Petosa, C.; Collier, R. J.; Bienkowska, J.; Lacy, D. B.; Park, S.; Leppla, S. H.; Hanna, P.; Liddington, R. C. Crystal structure of the anthrax lethal factor. *Nature* **2001**, *414*, 229–233.

(18) Turk, B. J.; Wong, T. Y.; Schwarzenbacher, R.; Jarrell, E. T.; Leppla, S. H.; Collier, R. J.; Liddington, R. C.; Cantley, L. C. The structural basis for substrate and inhibitor selectivity of the anthrax lethal factor. *Nat. Struct. Mol. Biol.* **2004**, *11*, 60–66.

(19) Panchal, R. G.; Hermone, A. R.; Nguyen, T. L.; Wong, T. Y.; Schwarzenbacher, R.; Schmidt, J.; Lane, D.; McGrath, C.; Turk, B. E.; Burnett, J.; Aman, M. J.; Little, S.; Sausville, E. A.; Zaharevitz, D. W.; Cantley, L. C.; Liddington, R. C.; Gussio, R.; Bavari, S. Identification of small molecular inhibitors of anthrax lethal factor. *Nat. Struct. Mol. Biol.* **2004**, *11*, 67–72.

(20) Forino, M.; Johnson, S.; Wong, T. Y.; Rozanov, D.; Savinov, A. Y.; Li, W.; Fattorusso, R.; Becattini, B.; Orry, A. J.; Abagyan, R. A.; Smith, J. W.; Alibek, K.; Liddington, R. C.; Strongin, A. Y.; Pellecchia, M. Efficient synthetic inhibitors of anthrax lethal factor. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (27), 9499–9504.

(21) Auld, D. S. Zinc coordination sphere in biochemical zinc sites. *Biometals* **2001**, *14*, 271–313.

(22) (a) Magistrato, A.; DeGrado, W. F.; Laio, A.; Rothlisberger, U.; VandeVondele, J.; Klein, M. L. Characterization of the dizinc analogue of the synthetic diiron protein DF1 using *ab initio* and hybrid quantum/classical molecular dynamics simulations. *J. Phys. Chem. B* **2003**, *107*, 4182–4188. (b) Dal Peraro, M.; Vila, A. J.; Carloni, P. Structural determinants and hydrogen-bond network of the mononuclear zinc(II)-beta-lactamase active site. *J. Biol. Inorg. Chem.* **2002**, *7*, 704–712. (c) Gervasio, F. L.; Schettino, V.; Mangani, S.; Krack, M.; Carloni, P.; Parrinello, M. Influence of outer-shell metal ligands on the structural and electronic properties of horse liver alcohol dehydrogenase zinc active site. *J. Phys. Chem. B* **2003**, *107*, 6886–6892.

(23) Tonello, F.; Naletto, L.; Romanello, V.; Dal Molin, F.; Montecucco, C. Tyrosine-728 and glutamic acid 735 are essential for the metalloproteolitic activity of the lethal factor of *Bacillus anthracis*. *Biochem. Biophys. Res. Commun.* **2004**, *131*, 496–502.

(24) Klimpel, K. R.; Arora, N.; Lepla, S. H. Anthrax toxin lethal factor contains a zinc metalloprotease consensus sequence which is required for lethal toxin activity. *Mol. Microbiol.* **1994**, *13* (6), 1093–1100.

(25) Hammond, S. E.; Hanna, P. C. Lethal factor active-site mutations affect catalytic activity in vitro. *Infect. Immun.* **1998**, *66*, 2374–2378.

(26) Neri, M.; Anselmi, C.; Cascella, M.; Maritan, A.; Carloni, P. Coarse-Grained model of proteins incorporating atomistic detail of the active site. *Phys. Rev. Lett.* **2005**, *95*, 218102.

(27) Tozzini, V. Coarse grained models for proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.

(28) Tirion, M. M. Large amplitude elastic motions in proteins from a single-parameter atomic analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905–1908.

(29) Rost, B.; Yachdav, G.; Liu, J. The PredictProtein server. *Nucleic Acids Res.* **2004**, *32* (Web Server Issue), W321−W326.

(30) Bryson, K.; McGuffin, L. J.; Marsden, R. L.; Ward, J. J.; Sodhi, J. S.; Jones, D. T. Protein structure prediction servers at University College London. *Nucleic Acids Res.* **2005**, *33*, W36−38.

(31) Vullo, A.; Bortolami, O.; Pollastri, G.; Tosatto, S. C. E. Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res.* **2006**, *34*, W164−168.

(32) Guermeur, Y.; Geourjon, C.; Gallinari, P.; Deleage, G. Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics* **1999**, *15* (5), 413–421.

(33) (a) Makinen, M. W.; Kuo, L. C.; Dymowski, J. J.; Jaffer, S. Catalytic role of the metal ion of carboxypeptidase A in ester hydrolysis. *J. Biol. Chem.* **1978**, *254* (5), 356–366. (b) Christianson, D. W.; Lipscomb, W. N. Carboxypeptidase A. *Acc. Chem. Res.* **1989**, *22*, 62–69. (c) Pelmenschikov, V.; Blomberg, M. R. A.; Siegbahn, P. E. M. A theoretical study of the mechanism for peptide hydrolysis by thermolysin. *J. Biol. Inorg. Chem.* **2002**, *7*, 284–298. (d) Cross, J. B.; Vreven, T.; Meroueh, S. O.; Mobashery, S.; Schlegel, H. B. Computational investigation of irreversible inactivation of the zinc-dependent protease carboxypeptidase A. *J. Phys. Chem. B* **2005**, *109*, 4761–4769.

(34) Piana, S.; Carloni, P.; Parrinello, M. Role of conformational fluctuations in the enzymatic reaction of hiv-1 protease. *J. Mol. Biol.* **2002**, *319*, 567–583.

(35) *CPMD*; Copyright IBM Corp 1990−2006, Copyright MPI für Festkörperforschung Stuttgart 1997−2001.

(36) Trouiller, N.; Martins, J. L. Efficient pseudopotentials for plane-wave calculation. *Phys. Rev. B* **1991**, *43*, 1993–2006.

(37) Keinman, L.; Bylander, D. M. Efficacious Form for Model Pseudopotentials. *Phys. Rev. Lett.* **1982**, *48*, 1425–1428.

(38) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behaviour. *Phys. Rev. A* **1998**, *38*, 3098–3100.

(39) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789.

(40) Barnett, R. N.; Landman, U. Born-Oppenheimer molecular-dynamics simulations of finite systems: Structure and dynamics of (H2O)2. *Phys. Rev. B* **1993**, *48*, 2081–2097.

(41) (a) IUPAC-IUB. Abbreviations and symbols for description of conformation of polypeptide chains. Tentative rules Biochemistry**1970**, *9* (18), 3471−3479. (b) IUPAC-IUB. Nomenclature and symbolism for amino acid and peptides. Recommendations 1983 Pure Appl. Chem. **1984**, *56* (5), 595−624.

(42) (a) Cheatham, T. E., III; Cieplak, P.; Kollman, P. A. A modified version of the Cornell *et al*. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845–862. (b) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E., III; DeBolt, S.; Ferguson, D.; Seibel, G. L.; Kollman, P. A. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics an1d free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **1995**, *91*, 1–41.

(43) (a) Suárez, D.; Díaz, N.; Merz, K. M., Jr. Molecular dynamics simulations of the mononuclear zinc $\beta$-lactamase from *Bacillus cereus* complexed with Benzylpenicillin and a quantum chemical study of the reaction mechanism. *J. Am. Chem. Soc.* **2001**, *123*, 9867–9879. (b) Suarez, D.; Diaz, N.; Merz, K. M., Jr. Molecular dynamics simulation of the dinuclear zinc-$\beta$-lactamase from *Bacteroides fragilis* complexed with imipenem. *J. Comput. Chem.* **2002**, *28*, 1587–1600. (c) Dal Peraro, M.; Villa, A. J.; Carloni, P. Substrate binding to Mononuclear metallo-$\beta$-lactamase from *Bacillus cereus*. *Proteins* **2004**, *54*, 412–423.

(44) Cheatham, T. E., III; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A. Molecular dynamics simulations on solvated biomolecular systems: the Particle Mesh Ewald method leads to stable trajectories of DNA, RNA, and proteins. *J. Am. Chem. Soc.* **1995**, *117*, 4193–4194.

(45) 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.

(46) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(47) (a) Laxmikant, K.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. NAMD2: greater scalability for parallel molecular dynamics. *J. Comp. Phys.* **1999**, *151*, 283–312. (b) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781−1802.

(48) Barret, C. P.; Hall, B. A.; Noble, E. M. Dynamite: a simple way to gain insight into protein motions. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *D60*, 2280–2287.

(49) Wriggers, W.; Schulten, K. Protein domain movements: Detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. *Proteins* **1997**, *29*, 1–14.

(50) Humphrey, W.; Dalke, A.; Schulten, K. VMD-Visual Molecular Dynamics. *J Mol. Graphics* **1996**, *14*, 33–38.

(51) Guidoni, L.; Torre, V.; Carloni, P. Water and potassium dynamics inside the KcsA K+ channel. *FEBS Lett.* **2000**, *477*, 37–42.

(52) Noguti, T.; Go, N. Collective variable description of small-amplitude conformational fluctuations in a globular protein. *Nature* **1982**, *296*, 776–278.

(53) Doi, M. *Introduction to Polymer Physics*; Oxford University Press: Oxford, 1996.

(54) van Gunsteren, W. F.; Daura, X.; Mark, A. E. GROMOS force field In *Encyclopedia of Computational Chemistry*; John Wiley & Sons: New York, 1998; 1211p.

(55) Lindahl, E.; Azuara, C.; Koehl, P.; Delarue, M. NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom Normal Mode Analysis. *Nucleic Acids. Res.* **2006**, *34*, W52−56.

(56) Bank, R.; Holst, M. A new paradigm for parallel adaptive meshing. *SIAM J. Sci. Comput.* **2000**, *22*, 1411–1443.

(57) Dolinsky, T.; Nielsen, J.; McCammon, A.; Baker, N. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatic calculations. *Nucleic Acid. Res.* **2004**, *32*, W665−W667.

(58) (a) Silvestrelli, P. L.; Marzari, N.; Vanderbilt, D.; Parrinello, M. Maximally-localized Wannier functions for disordered systems: application to amorphous silicon. *Solid State Commun.* **1998**, *107*, 7–11. (b) Berghold, G.; Mundy, C. J.; Romero, A. H.; Hutter, J.; Parrinello, M. General and efficient algorithms for obtaining maximally localized Wannier functions. *Phys. Rev. B* **2000**, *61*, 10040–10048. (c) Magistrato, A.; Robertazzi, A.; Carloni, P. Nitrogen fixation by a molybdenum catalyst mimicking the function of the nitrogenase enzyme: A critical evaluation of DFT and solvent effects. *J. Chem. Theory Comput.* **2007**, *3*, 1708–1720.

(59) Laio, A.; VandeVondele, J.; Rothlisberger, U. A Hamiltonian electrostatic coupling scheme for hybrid Car-Parrinello molecular dynamics simulations. *J. Chem. Phys.* **2002**, *116*, 6941–6947.

(60) (a) Vargiu, A. V.; Ruggerone, P.; Magistrato, A.; Carloni, P. Anthramycin-DNA binding explored by molecular simulations. *J. Phys. Chem. B* **2006**, *110*, 24687–24695. (b) Spiegel, K.; Magistrato, A. Modeling anticancer drug-DNA interactions via mixed QM/MM molecular dynamics simulations. *Org. Biomol. Chem.* **2006**, *4*, 2507–2517.

(61) Simona, F.; Magistrato, A.; Vera, D. M. A.; Garau, G.; Vila, A. J.; Carloni, P. Protonation state and substrate binding to B2 metallo-beta-lactamase CphA from. *Aeromonas hydrofila. Proteins* **2007**, *69*, 595–605.

(62) Jedrzejas, M. The structure and function of novel proteins of *Bacillus anthracis* and other spore forming bacteria: Development of novel prophylactic and therapeutic agents. *Crit. Rev. Biochem. Mol. Biol.* **2002**, *37*, 339–373.

(63) Carnevale, V.; Raugei, S.; Micheletti, C.; Carloni, P. Large-scale motions and electrostatic properties of furin and HIV-1 protease. *J. Phys. Chem. A* **2007**, *111*, 12327–12332.

(64) Piana, S.; Bucher, D.; Carloni, P.; Rothlisberger, U. Reaction Mechanism of HIV protease by hybrid Car-Parrinello/Classical MD simulations. *J. Phys. Chem. B* **2004**, *108*, 11139–11149.

(65) Henzler-Wildman, K. A.; Thai, V.; Lei, M.; Ott, M.; Wolf-Watz, M.; Fenn, T.; Pozharski, E.; Wilson, M. A.; Petsko, G. A.; Karplus, M.; Hübner, C. G.; Kern, D. Intrinsic motions along an enzymatic reaction trajectory. *Nature* **2007**, *450*, 838–844.

(66) Perryman, A. L.; Lin, J. H.; McCammon, J. A. HIV-1 Protease Molecular Dynamics of a Wild-Type and of the V82F/I84V Mutant: Possible Contributions to Drug Resistance and a Potential New Target Site for Drugs. *Protein Sci.* **2004**, *13*, 1108–1123.

(67) Christianson, D. W.; Cox, D. Catalysis by metal-activated hydroxide in zinc and manganese metalloenzymes. *Annu. Rev. Biochem.* **1999**, *68*, 33–57.

CT8001877

JCTC Journal of Chemical Theory and Computation

# Effects of Protein Subunits Removal on the Computed Motions of Partial 30S Structures of the Ribosome

Aimin Yan,[†] Yongmei Wang,[‡] Andrzej Kloczkowski,[†] and Robert L. Jernigan*,[†]

*Laurence H. Baker Center for Bioinformatics and Biological Statistics and Department of Biochemistry, Biophysics and Molecular Biology Iowa State University, Ames, Iowa 50011, and Department of Chemistry, University of Memphis, Memphis, Tennessee 38152*

Received June 16, 2008

**Abstract:** The Anisotropic Network Model (ANM) is used to study motions of the 30S small ribosomal subunit. The effect of the absence of certain subunits on the motions of the remaining partial structures was investigated by removing one protein, pairs of proteins, and selected sets of proteins at a time. Our results show that the removal of some proteins does not change the large-scale dynamics of the partial structures, but the removal of certain subunits does cause significant changes in motion of the remaining structure, and these changes can be reverted by the removal of other subunits, which indicates interdependence between motions of various parts of the 30S ribosomal structure. We further found that the subunits showing such interdependence have strong positive correlation of their motions, which indicates that these subunits function as a unit block in the 30S small ribosomal subunit. Dynamically interdependent subunit pairs identified in this paper are consistent with previous experimental observations that suggested dimerization of those subunits.

## Introduction

Many biological functions of proteins are related to their large-scale domain motions. By treating a protein as a 3D collection of masses connected by springs and using a theoretical framework developed earlier for rubberlike polymer networks,[1,2] Elastic Network Models of proteins have been proposed. The Gaussian Network Model (GNM) developed by Bahar et al.[3] assumes that fluctuations of residues (or atoms) in proteins around their mean equilibrium positions are spherically symmetric, while the Anisotropic Network Model (ANM) proposed by Atilgan et al.[4] takes into account the directionality of fluctuations and their anisotropy, measured by semiaxes of ellipsoids representing the fluctuations. Theoretical predictions of the extent of fluctuations of atoms and residues in biological structures agree surprisingly well with experimental temperature factors (B-factors) deposited in the Protein Data Base (PDB). Elastic Network Models have been used extensively to study large-scale functional motions of proteins and protein−protein complexes as well as for oligonucleotides and protein-DNA/RNA complexes. Bahar and Jernigan applied GNM to calculate atomic fluctuations for tRNA in the isolated form and the bound to Gln synthetase form, and their results reproduce closely experimental B-factors.[5] Another evidence of the successfulness of the elastic network approach to protein-nucleic acid systems comes from the application of the GNM to HIV-1 reverse transcriptase (RT). The RTs in different forms (bound to DNA or to inhibitors) were analyzed to infer the mechanism of function, and the predicted results were highly consistent with available experimental data.[6−8] Ramaswamy et al. studied motions of the nucleosome core particles using the elastic network model and revealed higher mobility of nucleosomes with variant histones, in accord with existing experimental observations.[9] Yang et al. studied recently 64 oligonucleotides and oligonucleotide-protein complexes, represented each nucleotide by three GNM nodes with uniform interaction cut-offs defining contacts for all components of the complexes, and achieved a very good agreement between the values of computed fluctuations and the experimental B-factors.[10] Tama et al. performed elastic network model computations

* Corresponding author e-mail: jernigan@iastate.edu.
[†] Iowa State University.
[‡] University of Memphis.

10.1021/ct800223g CCC: $40.75 
Published on Web 09/20/2008

for the ribosome and found the ratchetlike motion rearrangements of the 70S ribosome and a hingelike motion in the 30S ribosomal subunit, and such dynamical behavior of the ribosome was indeed observed in cryo-electron microscopy experiments.[11] Wang et al. applied GNM and ANM to study global motions in the ribosome and independently observed a similar ratchetlike motion in the 70S ribosome in agreement with experimental data.[12] All these results clearly show an enormous usefulness of GNM and ANM methods to study dynamics and function of DNA, RNA, and protein-DNA/RNA complexes, especially for large structures where traditional molecular dynamics simulations, requiring enormous computational resources, usually fail.

Ribosomes are large protein/RNA complexes that perform protein biosynthesis in all forms of life. Bacteria ribosomes are composed of a small (30S) and a large (50S) subunit that associate to form the intact 70S ribosome. The 30S subunit of the ribosome consists of 16S rRNA and 20 proteins. In 1970, it was found that the 30S subunit can reassemble from the 16S rRNA and a mixture of the 30S ribosomal proteins, and such a spontaneous reassembly process can produce a biologically active 30S structure.[13] Later it was shown that purified individual 30S subunit proteins and the naked 16S rRNA could also be reconstituted into active 30S particles *in vitro*. This shows that all the necessary information required for *in vitro* reassembly is contained within these molecular components. The 30S ribosomal subunit has been frequently used as a model system for studying ribosomal assembly. The reason for choosing the 30S ribosomal subunit as a model is due to its simplicity and the possibility of experimental control and manipulation of the assembly process *in vitro*.[14−16] By using sequential and combinatorial addition of proteins, Normura and co-workers[13] determined the assembly map of the 30S subunit. They divided the proteins into three categories, the primary binding proteins (S4, S7, S8, S16, S17, and S20) which bind 16S rRNA directly independent of other proteins, the secondary binding proteins (S5, S6, S11, S12, S13, S6, S18, and S19) which require at least one of the primary proteins to be bound to the 16S rRNA prior to binding, and the tertiary binding proteins (S2, S3, S10, S14, and S21) which require at least one protein from both of the previous sets be bound to the developing RNP core. Besides this assembly map determined by Normura and co-workers, another map based on the kinetics of assembly was obtained by Powers et al.[17] In this kinetic map, proteins are divided into early, mid, midlate, and late binding groups. The apparent agreement between the two maps is that the tertiary binding proteins are consistently found to be late binding proteins. These earlier experimental studies were performed before the structure of the 30S subunit was known. The availability of the X-ray crystal structure of the 30S subunit from *Thermus thermophilus*[18] now allows the researchers to examine the assembly process in light of the final end product. It is believed that the folding of 16S rRNA showed a 5′ to 3′ polarity (i.e., the 5′ domains fold before the 3′ domains). The early stage of the assembly involves the folding of individual domains of 16S rRNA, perhaps initiated by one of the primary binding proteins, and the late stage of the assembly involves the alignment of domain orientations, assisted by the binding of tertiary or late binding proteins.[19] Despite many years of investigation, understanding of the 30S assembly still remains elusive.

The availability of the X-ray crystal structure of the 30S subunit allows for computational investigation of the 30S assembly. Stagg et al. used coarse-grained Monte Carlo simulations to study the fluctuation changes upon the binding of proteins in the 3′ domain assembly for the 30S ribosomal subunit from *Thermus thermophilus* (1FJG) and examined the contributions of individual proteins to the formation of binding sites for the sequential proteins in the S7 pathway.[20] Hamacher et al. studied the dependencies of protein binding to 16S rRNA for the *Thermus thermophilus* 30S small ribosomal subunit by removing one protein or a pair of proteins at a time from the intact 30S small subunit using the self-consistent pair contact probability approximation method and produced a similar dependency map of proteins as that in *Escherichia coli* established earlier by the experimental methods.[21] The challenge in applying computational tools to study the assembly process is how to account for chemical/structural specificity of such large macromolecular complexes, knowing that chemical specificity matters. Both Stagg et al. and Hamacher et al. did not include or at most accounted very minimally for chemical specificity (Cα atoms versus P atoms). Yet both studies have been able to reveal useful information of the assembly pathway and even in some cases obtain agreement with the experimental results. The success of these approaches could be ascribed, we believe, to the fact that individual molecular contacts in biological assembly all have interaction strength on the order of thermal motions. The profound example supporting this statement can be found in the free energy change per base pair formation during the duplex formation.[22] Therefore, by accounting for the contact pairs observed in the final assembled particle, in principle one may trace useful information about the assembly pathway.

Elastic network models (ANM or GNM) are based on a similar principle as those used by Stagg et al. and Hamacher et al., namely, that contacts within protein structures or their macromolecular assemblies determine their dynamics. However, Stagg et al. performed Monte Carlo simulations of the partial structures distorted far away from the equilibrium. Similarly, Hamacher et al. study also allowed examining partial structures far from the equilibrium. In this regard, simple elastic network model calculations will not reveal as much information as the two previous studies. Nevertheless, it is worth applying simple elastic network model calculations to partial structures of the 30S assembly to examine the extent of information embedded in the partial structures that may reveal mechanisms of the assembly pathway.

Our ANM calculations were based on coarse-grained models of partial ribosome structures. A typical coarse-graining level is to use *C*α (for proteins) and P atoms (for DNA or RNA) as the nodes of the network and neglect side chains and other atoms. Doruker et al.[23] have shown that for large biological structures the coarse-graining level can be significantly reduced by using only a small fraction of nodes along the backbone (*n*/10, *n*/20, or *n*/40 nodes where

*n* is total number of residues) almost without any loss (over 95% correlation) of information on the large-scale dynamics. Our earlier results from comparison of different coarse-graining level models have shown that a removal of a part of the structure does not cause significant changes in large-scale motions if the global shape of the molecule is maintained. In the present paper we extend this study to the small 30S ribosomal subunit and examine whether the removal of some protein subunits in the 30S changes large-scale functional motions of the remaining part of the 30S structure.

## Materials and Methods

**Structure Used in This Study.** A 3.05 Å resolution crystal structure from *T.thermophilu*s is used in this study. Its PDB entry code is 1J5E. This structure is a native form of the 30S subunit and does not bind with any ligand. This large biomolecule consists of 21 chains, each chain being a separate subunit. Among these subunits, chain A is the nucleic acid 16S rRNA, and all other chains are proteins. The organizations and interactions of these subunits and binding rate of the protein subunits to 16S rRNA can be found in Figure 2(b),(c) in ref 24.

**Overlap Matrix Calculation.** The similarities between two sets of vectors are measured by the overlap matrix. Each element in the overlap matrix is calculated by the following equation

$$\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}||\mathbf{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{|\mathbf{x}||\mathbf{y}|} \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are two vectors, $|\mathbf{x}|$ and $|\mathbf{y}|$ are their lengths, $x_i$ and $y_i$ denote their $i$-th components, and $\theta$ is the angle between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. If two vectors are exactly collinear, then their absolute overlap is 1. If they are orthogonal to each other, then the overlap is 0.

**Correlation of Substructure Motions in the Complete 30S Subunit.** In order to understand the relative motions of each structural subunits in the 30S subunit, the orientation correlation between the center of mass of the displacement of the structure subunits are examined by using the following equation

$$C_{I,J}(k) = \frac{\Delta\mathbf{R}_I^{cm}(k)\cdot\Delta\mathbf{R}_J^{cm}(k)}{|\Delta\mathbf{R}_I^{cm}(k)||\Delta\mathbf{R}_J^{cm}(k)|} \tag{2}$$

where $I$ and $J$ denote indices of the subunits (16S rRNA, S2, S3,...S20, THX), $k$ denotes the mode, and $\Delta\mathbf{R}_I^{cm}(k)$ is the center of the mass displacement for the $I$-th subunit in the $k$-th mode, which is computed by summing up displacement vectors of all nodes for the $I$-th subunit in the $k$-th mode.

**Calculation of the Deformation Energy.** The deformation energy of each residue is calculated by using Wang et al. method.[12] This method is described by eq 3

$$D_i(k) = \frac{\sum_{j=1}^{n_{ci}} \frac{1}{2}\gamma(|\mathbf{R}_{ij}^0 + \Delta\mathbf{R}_j(k) - \Delta\mathbf{R}_i(k)| - |\mathbf{R}_{ij}^0|)^2}{N\lambda(k)} \tag{3}$$

where $n_{ci}$ is the number of nodes connected to the $i$-th node (number of contacts based on the assumed value of the cutoff distance $R_c$), and $N$ is the number of nodes; $\lambda(k)$ is the eigenvalue of the $k$-th normal mode, which is used as a weighting factor. $|\mathbf{R}_{ij}^0|$ is the average value of the distance between residues $i$ and $j$, and $\Delta\mathbf{R}_j(k)$ is the displacement vector of the $j$-th residue in the $k$-th mode; while $D_i(k)$ indicates the deformation energy for the $i$-th residue in the $k$-th mode.

**Protein Removal Method.** As we have already mentioned, the 30S ribosomal structure used in this study contains 21 independent subunits. Among these subunits, chain A is the 16S rRNA and is always included in all partial structures generated in our protein removal experiments. The numbers of all possible partial structures that are obtained by the removal of $m$ out of 20 subunits can be computed from the formula

$$N_m = \binom{20}{m} = \frac{20!}{m!\,(20-m)!} \tag{4}$$

and are listed in Table 7. If we try to remove all different combinations of proteins, the numbers of possible cases are too large, to be computationally treatable. Therefore we focus only on the simplest cases of removing all protein subunits, one protein, pairs of proteins, and the selected sets of proteins at a time. After the partial structures are obtained, the following procedures are performed:

- We use ANM to calculate the mean-square fluctuations for the partial structure.
- We use ANM to calculate the mean-square fluctuations for the complete structure.
- We compare the difference in the mean-square fluctuation profiles between the partial structure and the corresponding part in the whole structure by calculating the root-mean-square error (RMSE) between them, which is shown in the following equation

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (MSFP_i - MSFC_i)^2}{N}} \tag{5}$$

where $\mathrm{MSFP}_i$ is the mean-square fluctuation for residue $i$ of the partial structure, $\mathrm{MSFC}_i$ is the mean-square fluctuation for residue $i$ of the corresponding part in the whole structure, and $N$ is the number of residue in the partial structure.

To compare the deformation energy difference between the partial structures and the corresponding parts in the complete structure, the same procedures are used.

**Computation Cost.** In the ANM calculations, we used the positions of the P and the *O*4* atoms as nodes representing each nucleotide and the *C*α to represent each residue. The spring constant is set to 1. The same model was used in our previous studies,[12] and our attempts to use different node representations or different values of the spring constant did not significantly affect the computed modes. For the complete 30S ribosomal structure, there are total 5422 nodes. Using this coarse-grained model, a 16266 × 16266 Hessian matrix is constructed. The full spectral decomposition of such large matrices is computationally very expensive.

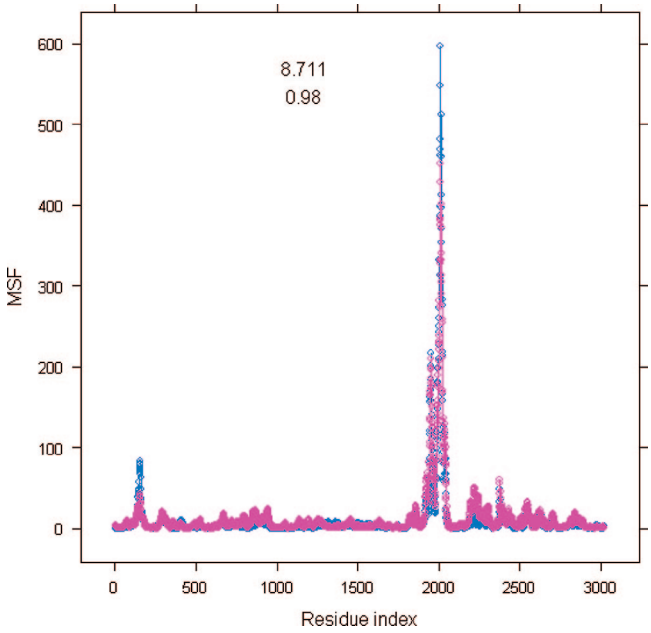



***Figure 1.*** Comparison in the mean-square fluctuations between the 16S rRNA part in the complete structure and using the single 16S rRNA molecule only. The X-axis is the residue index. The Y-axis is mean-square fluctuations for 16S rRNA alone (labeled by pink color) and the 16S rRNA subunit in the whole structure (labeled by blue color).

Since we are only interested in the slowest modes, we take the first 100 eigenvalue-eigenvector pairs into consideration only. In order to calculate the first 100 eigenvalue-eigenvector pairs, we can use the Matlab function (*eigs*), but it takes about 2 h to complete the computations. Therefore we have used BLZPACK to perform the spectral decomposition. BLZPACK stands for the Block LancZos PACKage and is a standard Fortran 77 implementation of the block Lanczos algorithm for the solution of the eigenvalue problem.[25] It takes 1 min for a complete computation on a Linux machine with 2.8 GHz CPU.

## Results

**The Influence of Removing All Protein Subunits on the Computed Motions of 16S rRNA.** Wang et al. pointed out that the shape of the 30S subunit is determined mainly by the 16S rRNA.[12] The dominant role of the molecular shape on the large-scale functional motions computed with the elastic network models was suggested by Ma.[26,27] In this section, we discuss the importance of the shape of the 16S rRNA on functional motions of the 30S subunit of the ribosome. In order to estimate this effect, we removed all protein subunits from the 30S structure and compared the mean-square fluctuations of the 16S rRNA subunit in the complete 30S structure with the 16S rRNA alone. Figure 1 shows this comparison in the slowest mode: we computed mean-square fluctuations profiles with corresponding root-mean-square errors and the correlation coefficient between them. The root-mean-square error and the correlation coefficient between two types of the mean-square fluctuations are 8.711 and 0.98, respectively, which indicates that the computed large-scale motions for 16S rRNA subunit using

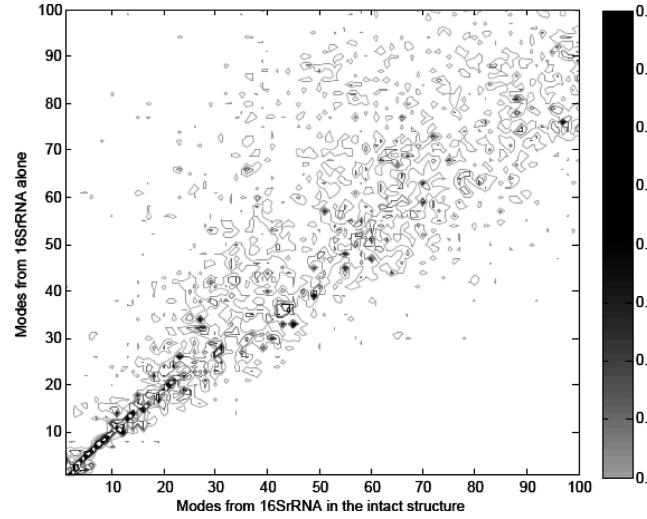



***Figure 2.*** Overlap between modes unbinding 16S rRNA (Y-axis) and modes from binding 16S rRNA (X-aixs). The black color spectrum stands for the higher overlap value.

the complete 30S structure are very close to that based on the 16S rRNA structure alone. Furthermore we have checked for possible differences in the directionality of the normal modes of 16S rRNA between the complete 30S structure and the structure of a single 16S rRNA. The overlap matrix between the two sets of normal modes was calculated and is shown in Figure 2. From Figure 2, we can see that the first several slowest modes are highly overlapped, which indicates that the motion of the 16S rRNA subunit is mainly determined by the structure of 16S rRNA alone, and that intermolecular interactions between the 16S rRNA subunit and protein subunits have little effect on the dynamics of 16S rRNA in the 30S ribosomal structure.

In addition, we were interested in the role of different protein subunits in the global dynamics of the 30S subunit. To answer this question, we performed single protein subunit removal experiments described in the next section.

**Influence of the Removal of Single Proteins from the 30s complex.** We have removed each one of the 20 constituent proteins at a time from the whole 30S complex and then computed the mean-square fluctuations of the remaining part of the 30S structure and the corresponding root-mean-square errors (RMSE). Figure 3 shows the root-mean-square error due to the removal of each of protein subunits. From Figure 3, it is clear that the removal of S18 will have the largest impact on the difference between mean-square fluctuations in the complete 30S structure and the incomplete structure with the removed chain. Removal of subunits S2, S8, S14, and S17 will also cause some noticeable changes in the global dynamics. The detailed values for these changes are listed in Table 1. In addition to the computations of changes in the mean-square fluctuations due to a protein chain removal we have also calculated corresponding changes in the deformation energy. (See the Methods section for computational details.) Figure 4 shows these changes in the deformation energy. Figure 4 shows that the changes in the deformation energy are the largest for removing subunits S2, S8, S14, S17, and S18 which agrees with the results of the corresponding changes in the

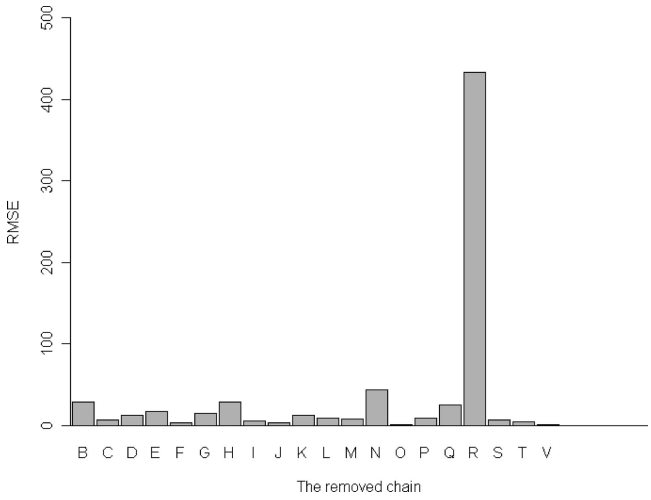



***Figure 3.*** One protein removal experiment. X-axis: the removed protein subunit, Y-axis: RMSE for mean-square fluctuations profile.

***Table 1.*** Root-Mean-Square Error for the Mean Square Fluctuations between the Partial Structure and the Corresponding Parts in the Whole Complex Structure in the Slowest Mode

| removed chain | corresponding subunit | RMSE for MSF |
|---|---|---|
| B | S2 | 28.63* |
| C | S3 | 6.41 |
| D | S4 | 12.90 |
| E | S5 | 17.18 |
| F | S6 | 2.88 |
| G | S7 | 14.83 |
| H | S8 | 28.56* |
| I | S9 | 5.82 |
| J | S10 | 2.93 |
| K | S11 | 12.51 |
| L | S12 | 9.01 |
| M | S13 | 7.87 |
| N | S14 | 43.61* |
| O | S15 | 0.60 |
| P | S16 | 8.43 |
| Q | S17 | 25.30* |
| R | S18 | 433.60* |
| S | S19 | 6.91 |
| T | S20 | 3.79 |
| V | THX | 0.89 |

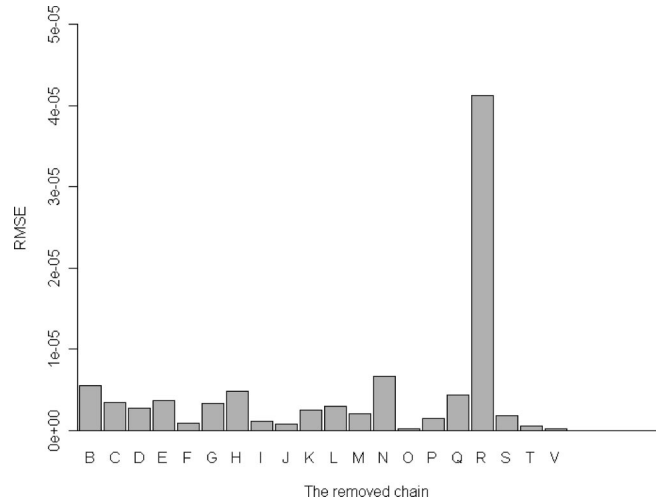



***Figure 4.*** One protein removal experiment. X-axis: the removed protein subunit, Y-axis: RMSE for the deformation energy profile.

***Table 2.*** Root-Mean-Square Error for Deformation Energies between the Partial Structure and the Corresponding Parts in the Whole Complex Structure in the Slowest Mode

| removed chain | corresponding subunit | RMSE for deformation energy |
|---|---|---|
| B | S2 | 5.48e-06* |
| C | S3 | 3.46e-06 |
| D | S4 | 2.78e-06 |
| E | S5 | 3.65e-06 |
| F | S6 | 8.36e-07 |
| G | S7 | 3.26e-06 |
| H | S8 | 4.82e-06* |
| I | S9 | 1.180e-06 |
| J | S10 | 7.85e-07 |
| K | S11 | 2.50e-06 |
| L | S12 | 3.00e-06 |
| M | S13 | 2.07e-06 |
| N | S14 | 6.66e-06 |
| O | S15 | 1.82e-07 |
| P | S16 | 1.46e-06 |
| Q | S17 | 4.41e-06* |
| R | S18 | 4.13e-05* |
| S | S19 | 1.79e-06 |
| T | S20 | 5.70e-07 |
| V | THX | 2.36e-07 |

mean-square fluctuations. Table 2 lists all these changes in the deformation energy for removal of each of the 20 proteins from the 30S structure. From Table 2, we see that the change caused by removing subunit S18 is the largest one, which is consistent with the results in Table 1 for the difference in the mean-square fluctuations.

We generated a movie showing motions corresponding to the slowest mode of partial structure after removing S18. From this movie, we have clearly observed that the removal of the subunit S18 induces large amplitude motions of the terminal residues in subunit S6. Therefore we have removed the two terminal residues in S6 and then repeated the single-protein removal experiment. Similarly as before, we have calculated the root-mean-square error for the mean-square fluctuations and deformation energies between the two structures. Tables 3 and 4 show these results for the slowest mode.

From the results of the computed mean-square fluctuations and deformation energies, it is obvious that by removing two terminal residues in S6 we cancel the effects caused by the removal of S18. These results indicate possible interdependence between different subunits in the global dynamics manifested in the single protein removal experiments. In order to better understand this interdependence we have extended our computational protein removal experiment to pairs of proteins at a time, and the results are shown in the following section.

**Influence of the Removal of Pairs of Proteins from the 30S Complex.** We have performed the removal of a pair of proteins at a time and then computed the root-mean-square error of the mean-square fluctuations of residues in the partially remaining structure of the 30S ribosomal subunit and the complete 30S structure in the slowest mode. These

***Table 3.*** Root-Mean-Square Error for Mean Square Fluctuations between the Partial Structure and the Corresponding Parts in the Structure after Removing Two Terminal Residues in the Subunit S6 in the Slowest Mode

| removed chain | corresponding subunit | RMSE for MSF |
|---|---|---|
| B | S2 | 28.65* |
| C | S3 | 6.44 |
| D | S4 | 12.87 |
| E | S5 | 17.26 |
| F | S6 | 2.86 |
| G | S7 | 14.89 |
| H | S8 | 28.59* |
| I | S9 | 5.77 |
| J | S10 | 2.76 |
| K | S11 | 12.54 |
| L | S12 | 9.05 |
| M | S13 | 7.78 |
| N | S14 | 43.46* |
| O | S15 | 0.63 |
| P | S16 | 8.44 |
| Q | S17 | 25.39* |
| R | S18 | 2.97** |
| S | S19 | 6.77 |
| T | S20 | 3.80 |
| V | THX | 0.74 |

***Table 4.*** Root-Mean-Square Error for Deformation Energies between the Partial Structure and the Corresponding Parts in the Structure after Removing Two Terminal Residues in the Subunit S6 in the Slowest Mode

| removed chain | corresponding subunit | RMSE for deformation energy |
|---|---|---|
| B | S2 | 5.48e-6* |
| C | S3 | 3.46e-6 |
| D | S4 | 2.78e-6 |
| E | S5 | 3.66e-6 |
| F | S6 | 8.28e-7 |
| G | S7 | 3.27e-6 |
| H | S8 | 4.83e-6* |
| I | S9 | 1.18e-6 |
| J | S10 | 7.86e-7 |
| K | S11 | 2.50e-6 |
| L | S12 | 3.00e-6 |
| M | S13 | 2.06e-6 |
| N | S14 | 6.66e-6* |
| O | S15 | 1.85e-7 |
| P | S16 | 1.47e-6 |
| Q | S17 | 4.43e-6* |
| R | S18 | 4.25e-7** |
| S | S19 | 1.79e-6 |
| T | S20 | 5.73e-7 |
| V | THX | 2.360e-7 |

results are shown in Figure 5. Figure 5 shows that out of the total of 190 investigated pairs there are 17 pairs of proteins that have significantly larger root-mean-square errors than the rest. If we inspect all these 17 pairs of proteins, we see that subunit S18 is involved in all these pairs. We were also interested in the two remaining pairs of proteins containing the S18 subunit. In Table 5, we have listed all pairs of proteins that include S18 and the corresponding RMSE in the slowest mode. From Table 5, it is clear that the root-mean-square error of the mean-square fluctuations of residues for the partial 30S structure and for the intact 30S structure are significantly reduced, respectively to all other cases, if we remove subunits S3 and S18 together (RMSE = 43.8) or remove subunits S6 and S18 together

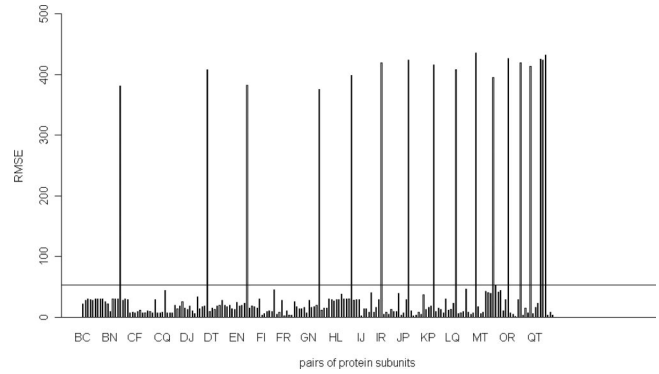



***Figure 5.*** The two proteins removal experiment. X-axis: the removed protein pairs (190 pairs), Y-axis: root-mean-square error (RMSE) for the mean-square fluctuations profile. The horizontal line indicates the position that RMSE is 53.1786.

***Table 5.*** Root-Mean-Square Error for Mean-Square Fluctuations between the Partial Structure and the Corresponding Parts in the Complete Structure after Removing S18 Involved Pairs of Proteins in the Slowest Mode

| removed chains | corresponding subunits | RMSE for MSF |
|---|---|---|
| BR | S2-S18 | 381.34 |
| CR | S3-S18 | 43.77** |
| DR | S4-S18 | 407.31 |
| ER | S5-S18 | 382.32 |
| FR | S6-S18 | 2.15** |
| GR | S7-S18 | 374.90 |
| HR | S8-S18 | 399.16 |
| IR | S9-S18 | 418.98 |
| JR | S10-S18 | 423.63 |
| KR | S11-S18 | 416.50 |
| LR | S12-S18 | 407.46 |
| MR | S13-S18 | 435.55 |
| NR | S14-S18 | 394.65 |
| OR | S15-S18 | 426.25 |
| PR | S16-S18 | 419.73 |
| QR | S17-S18 | 413.63 |
| RS | S18-S19 | 425.61 |
| RT | S18-S20 | 423.88 |
| RV | S18-THX | 431.56 |

(RMSE = 2.15). We have also computed the root-mean-square error in the deformation energies for the remaining 30S structure and the intact 30S structure in the slowest mode. These results for all pairs or proteins containing the S18 subunit are plotted in Figure 6 and listed in Table 6. From the computations of the deformation energy in Table 6, we see that the removal of the S6 and the S18 subunits together will cause almost no changes in the deformation energy, which agrees with the computations of the difference of the mean-square fluctuations for the same pair of proteins in Table 5.

In a single-protein removal experiments, removing of the subunit S18 significantly changes the large-scale dynamics of the remaining 30S structure relative to the intact 30S structure. We have also observed that the two terminal residues of the subunit S6 exhibit large fluctuations after the removal of the subunit S18. However, if we remove the subunit S18 and these two terminal residues of the subunit S6 together, the root-mean-square error of fluctuations of residues in the partial 30S structure and in the intact 30S

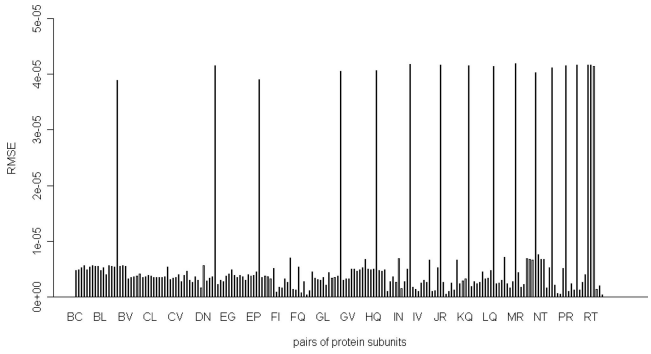



***Figure 6.*** The two proteins removal experiment. X-axis: the removed protein pairs (190 pairs), Y-axis: root-mean-square error (RMSE) for deformation energy profile.

***Table 6.*** Root-Mean-Square Error for Deformation Energies between the Partial Structure and the Corresponding Parts in the Complete Structure after Removing S18 Involved Pairs of Proteins in the Slowest Mode

| removed chains | corresponding subunits | RMSE for deformation energy |
|---|---|---|
| BR | S2-S18 | 3.89e-05 |
| CR | S3-S18 | 5.31e-06** |
| DR | S4-S18 | 4.16e-05 |
| ER | S5-S18 | 3.90e-05 |
| FR | S6-S18 | 7.70e-07** |
| GR | S7-S18 | 4.06e-05 |
| HR | S8-S18 | 4.07e-05 |
| IR | S9-S18 | 4.18e-05 |
| JR | S10-S18 | 4.17e-05 |
| KR | S11-S18 | 4.15e-05 |
| LR | S12-S18 | 4.15e-05 |
| MR | S13-S18 | 4.19e-05 |
| NR | S14-S18 | 4.03e-05 |
| OR | S15-S18 | 4.12e-05 |
| PR | S16-S18 | 4.16e-05 |
| QR | S17-S18 | 4.16e-05 |
| RS | S18-S19 | 4.17e-05 |
| RT | S18-S20 | 4.17e-05 |
| RV | S18-THX | 4.14e-05 |

structure lowers to 2.97 (see Table 3), which is similar to the RMSE value of 2.15 when we remove the whole subunits S6 and S18 together (see Table 5). These results indicate that the subunit S18 serves a role of a spatial constraint that prevents the two terminal residues of the subunit S6 to have large fluctuations. After the subunit S18 is removed, this constraint is missing, and we observe large fluctuations of the two terminal residues of the subunit S6.

**Remove the Sets of Protein Subunits Based on Their Binding Order.** Earlier work by Nomura et al. showed that 20 protein subunits bind with 16S rRNA in a specific order. Using this order, these proteins are classified into the primary, secondary, and tertiary binding proteins. The primary binding proteins are chains S17, S4, S20, S8, S15, and S7. The secondary binding proteins include S12, S16, S18, S6, S9, S19, S13, S5, and S11. The tertiary binding proteins contain chains S14, S10, THX, S3, and S2. We use Bp, Bs, and Bt to indicate the different sets of protein subunits:

Bp = {S4 S7 S8 S15 S17 S20}

Bs = {S5 S6 S9 S11 S12 S13 S16 S18 S19}

Bt = {S2 S3 S10 S14 THX}

***Table 7.*** Number of Partial Structures in the Protein Removal Experiments

| number of proteins removed | number of partial structures |
|---|---|
| 1 | 20 |
| 2 | 190 |
| 3 | 1140 |
| 4 | 4845 |
| 5 | 15504 |
| 6 | 38760 |
| 7 | 77520 |
| 8 | 125970 |
| 9 | 167960 |
| 10 | 184756 |
| 11 | 167960 |
| 12 | 125970 |
| 13 | 77520 |
| 14 | 38760 |
| 15 | 15504 |
| 16 | 4845 |
| 17 | 1140 |
| 18 | 190 |
| 19 | 20 |
| 20 | 1 |

The combinations of Bp, Bs, and Bt are listed below:

BpBs = {S4 S5 S6 S7 S8 S9 S11 S12 S13 S15 S16 S17 S18 S19 S20}

BpBt = {S2 S3 S4 S7 S8 S10 S14 S15 S17 S20 THX}

BsBt = {S2 S3 S5 S6 S9 S10 S11 S12 S13 S14 S16 S18 S19 THX}

BpBsBt = all protein subunits

Here we attempt to find how the removal of these sets of protein subunits affects motions of partial 30S structures. In order to answer this question, we perform the protein removal simulations by removing the groups of Bp, Bs, Bt, BpBs, BpBt, and BsBt subunits separately and calculating the mean deviations per residue of the mean-square fluctuations and deformation energies between the partial structure after the removal and the corresponding part in the intact structure. For the slowest mode, the mean deviations per residue are 26.24, 12.65, 12.44, 33.40, 6.53, and 11.10 for removal of Bp, Bs, Bt, BpBs, BpBt, and BsBt, respectively. It is clear that removing the primary and secondary binding proteins together causes the largest mean deviation, while removing both the primary and the tertiary binding proteins together leads to the smallest mean deviation.

The similar change patterns are also reflected from the mean deviation per residue for the deformation energy. The mean deviations per residue are 4.79e-06, 3.37e-06, 4.32e-06, 5.91e-06, 3.97e-06, and 4.60e-06 for removal of Bp, Bs, Bt, BpBs, BpBt, and BsBt, respectively, which shows that removing the primary and secondary binding proteins together causes the largest mean deviation in the computed deformation energy.

**Effects of the Removal of Protein Subunits on Motions of Partial Structures Depend on Contacts between the Protein Subunits and the 16SrRNA Subunit in the 30S Structure.** In the protein removal experiments, we observe that the removal of some proteins, pairs of proteins, and subsets of proteins causes larger changes in the mean-square fluctuations for partial structures, while the removal of other proteins have smaller effects on the corresponding mean

square fluctuations. We are especially interested whether the effects from removing the different protein subunits are related to the contacts between the protein subunits and the 16S rRNA subunit. To answer this question, we calculate the contact numbers between subunits and construct a contact map (see Figure 10). We assume that two nodes (one from the 16S rRNA subunit and another from the protein subunit) are being in contact if the distance between them is less than or equal to 15 Å. From Figure 10, we calculate the average contact number for the primary, secondary, and tertiary binding proteins. The average contact numbers for the primary, secondarym and tertiary binding proteins are calculated by dividing the total contact numbers for each category by the number of subunits in this category. The average contact numbers for these three sets of proteins are 1669, 1414, and 1088, respectively, which indicates that the earlier the proteins bind with 16S rRNA, the more contact the proteins have with 16S rRNA. In addition, the relationship between the contact numbers and the effects of the removal of the protein subunits on the motion of the partial 30S structure are also studied. We calculate the Pearson and Spearman rank correlation between the contact number and the RMSE in Table 1, and these values between the two measures are −0.34 and 0.009, respectively, which indicates that there is no obvious linear relationship between them. However we did observe that S6 has the smallest contact number, and the removal of S6 causes the smaller effect on the motion of the partial structure (RMSE for S6 in Table 1 is 2.88 only). For the removal of pairs of proteins always including S18, the average contact number for the pairs including S6−S18 is the smallest, and the RMSE in Table 5 is 2.15, which is the smallest RMSE for all pairs including S18. In addition, we are interested whether there is a linear relationship between contact ratios and the effects of the removal of the protein subunits on the motion of the partial 30S structure. The contact ratio between protein subunit and 16SrRNA is calculated by dividing the total contact number between two subunits by the product of numbers of residues of two subunits. We calculate the Pearson and Spearman rank correlation between the contact ratios and the RMSEs in Table 1, and these values between the two measures are −0.18 and −0.31, respectively, which again indicates that there is no obvious linear relationship between contact ratios and the effects of the removal of the protein subunits on the motion of the partial 30S structure. We also note that the Pearson and Spearman rank correlation between the contact numbers and the contact ratios are 0.27 and 0.46, respectively.

**The Correlation of Motions of Different Subunits.** Since the ribosome is a biological machine for protein synthesis, we expect that motions of its different subunits are highly correlated to process the synthesis of proteins smoothly. Wang et al.[12] studied the correlations of motions of different subunits in the whole 70S ribosome structure. However, they did not examine correlations of motions of various subunits in the 30S ribosomal structure. Our results of the removal of pairs of proteins at a time suggest that the subunit S18 constrains motions of the subunit S6. We are especially interested how motions of S18 are correlated with motions of S6 and with other protein chains. We have computed

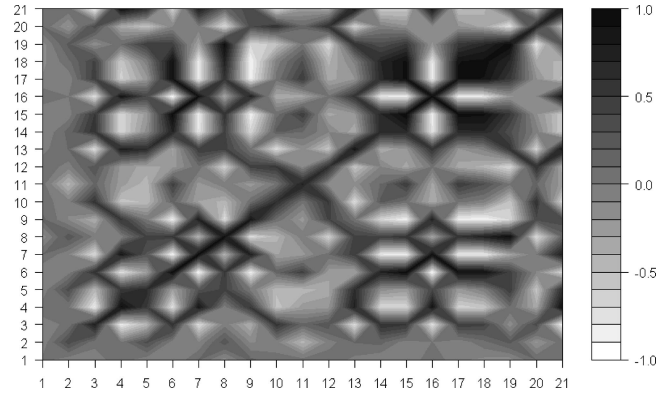



***Figure 7.*** The motion correlation between subunits in the 30S structure. X-axis and Y-axis indicate 21 subunits. The black color spectrum stands for the higher correlation value, and the correlations are calculated based on only the slowest mode.

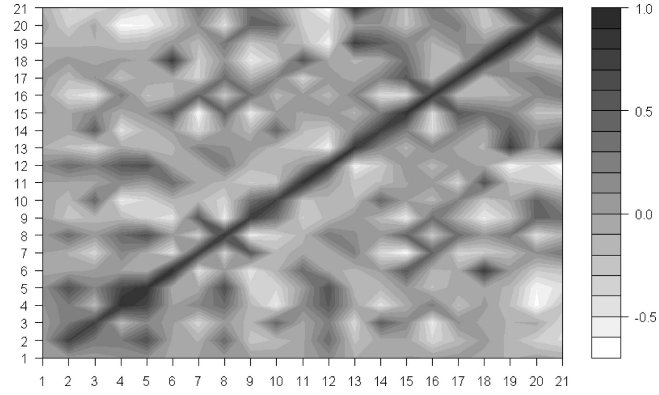



***Figure 8.*** The motion correlation between subunits in the 30S structure. X-axis and Y-axis indicate 21 subunits. The black color spectrum stands for the higher correlation value, and the correlations are calculated based on the first 10 slowest modes.

correlations of motion among 21 subunits of the 30S structure. Figure 7 shows these results corresponding to the slowest mode. From Figure 7, it is obvious that there are strong positive correlations of motion between subunits S6 and S18. Furthermore, we have investigated the effect of higher modes by computing these correlations for the first 10 slowest and the first 100 slowest modes. Figures 8 and 9 show the results of our studies.

It is clear that there is a positive correlation between the motions of subunits S6 and S18 even if we take other slowest modes into consideration. The correlation coefficients between the motions of subunits S6 and S18 in the first slowest mode, the first 10 slowest modes, and the first 100 slowest modes are 0.91, 0.77, and 0.59, respectively. By combining this information on correlation coefficients with the results of the protein removal experiments, we are lead to the conclusion that subunits S6 and S18 function together as a single block in the whole 30S ribosomal structure. The removal of the subunit S18 alone will significantly change the dynamics of the remaining structure; however, the removal of both subunits S6 and S18 at once will eliminate the changes caused by the removal of the single chain S18.

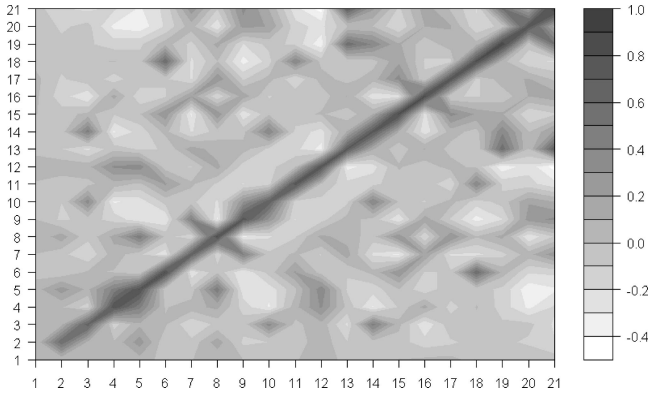


***Figure 9.*** The motion correlation between subunits in the 30S structure. X-axis and Y-axis indicate 21 subunits. The black color spectrum stands for the higher correlation value, and the correlations are calculated based on the first 100 slowest modes.

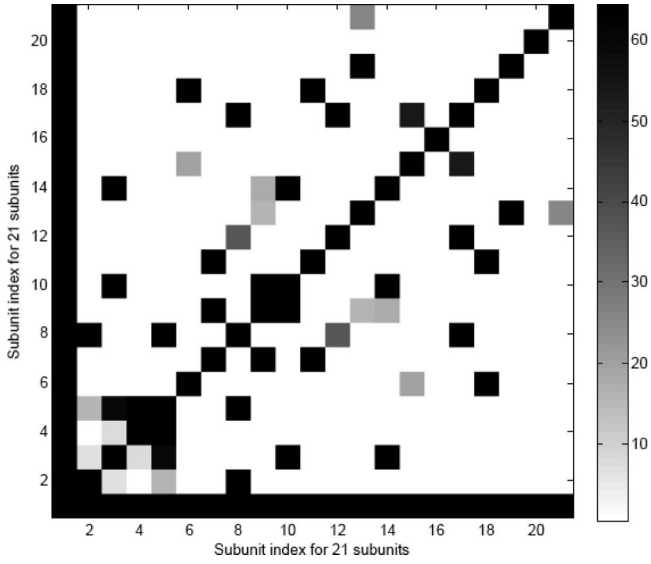



***Figure 10.*** Contact map between subunits in the 30S ribosome structure, using 15 Å as the cut off distance for defining contact. X-axis and Y-axis indicate 21 subunits. The black color spectrum stands for the larger contact number.

**Subunits That Have More Contacts Have Stronger Correlated Motions Computed from ENM.** In this section, we study a relationship between contacts among subunits and the correlated motions of these subunits. We calculated the contact ratio of pairs of proteins and compared them with the correlated motions of these pairs. These results are shown in Figure 11. The contact ratio for a pair of proteins is calculated by dividing the total contact number between the two subunits by the product of numbers of residues of two subunits. Since there are 21 subunits, there are 210 different pairs of subunits. Among these 210 pairs, there are 44 pairs of subunits that have a nonzero contact ratio. Therefore we only include these 44 pairs of subunits in Figure 11. The Pearson and Spearman correlation coefficient between the contact ratio of pairs of subunits and the motion correlation coefficient between subunits in Figure 12 are 0.42 and 0.52, respectively, which indicate the relationships between the contacts among subunits and their correlated motions. Particularly, subunit F and subunit R have the larger contact ratio, and the motion correlation between them is also stronger (labeled by an arrow in Figure 11).

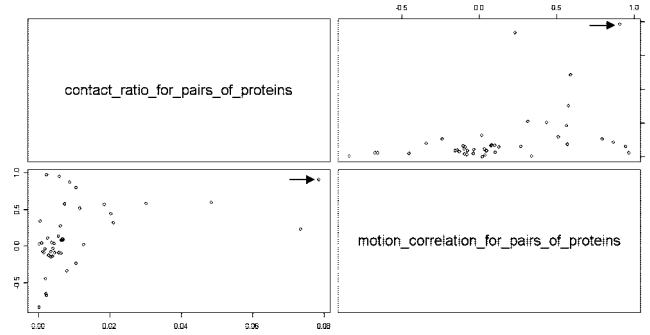



***Figure 11.*** The relationship between the contact ratio for pairs of proteins and the motion correlation of pairs of proteins. The arrow indicates the pair of subunits S6 and S18.

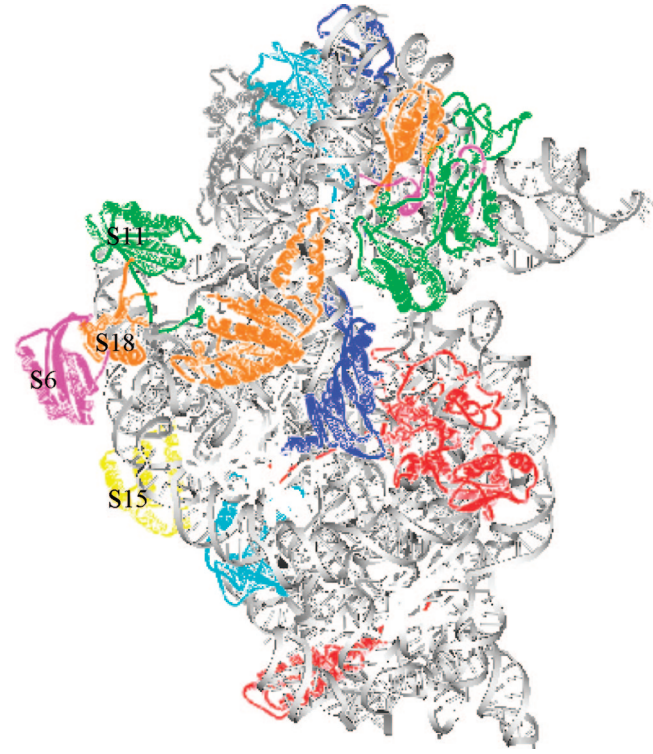



***Figure 12.*** Structure of the 30S ribosomal subunit (viewed using ViewerPro 4.2). Gray color indicates 16S rRNA. Protein subunits are represented by the different colors. Some proteins involved in the S15 binding pathway are labeled (S6, S11, S15, S18).

## Conclusions

In this study, we find that the slowest modes of 16S rRNA after removing all protein subunits are very similar to the slowest mode of the 16S rRNA part in the whole 30S ribosome subunit. However, the slowest mode of partial structures obtained after removing S18 or some pairs of protein subunits containing S18 are very different from the slowest modes of the corresponding parts in the whole 30S ribosome. This should not be considered to be contradictory to what we indicated that the dynamics of the 16S rRNA alone is affected little by the interactions with the protein subunits. The partial structures obtained after removal of S18

or pairs of subunits containing chain S18 still contain other protein subunits, and the slowest modes that are computed using ANM are strongly influenced by some of these remaining subunits. We have computed the extent of change of the mean-square fluctuations due to the remaining chains after the removal of chain S18. Our calculations show that 99.52% of changes in the mean-square fluctuations profile is related to chain S6, and only 0.47% is due to 16S rRNA.

Hamacher et al. studied the dependency map of proteins in the 30S small ribosomal subunit assembly by calculating the difference in the binding free energy performing a single protein removal and two proteins removal experiments.[21] Their studies have shown that subunits S6 and S18 influence each other. Some early experimental studies indicated that chains S6 and S18 bind to each other forming a dimer.[14,17] Other experiments using the 30S ribosomal subunit from hyperthermophilic bacteria *Aquifex aeolicus* also suggested a possible dimerization of subunits S6 and S18.[28,29] The previous studies showed that S6 and S18 are located in the central domain of 16S rRNA, and the crystal structure of this domain is already solved by Williamson.[30] The principal interface protrusion of the 50S subunit penetrates deeply into this domain and remains virtually unchanged in the 70S complex.[31] Some studies also showed that this domain includes the S15 protein binding pathway.[29,30,32] The cooperative binding of S6 and S18 follows the binding of S15 to 16S rRNA and is required for the binding of S11 and S21.[27] The location of these protein subunits can be seen in Figure 12. Thermodynamics and kinetic experiments of S6 and S18 binding to an S15-RNA complex indicates that S6 and S18 bind, forming a stable heterodimer in solution, and this S6:S18 heterodimer binds to the S15-rRNA complex.[29] Our present results obtained from elastic network model computations also indicate that S6 and S18 possibly function as a block, and this result is consistent with the above experimental data.

The purpose of our research was to apply the Anisotropic Network Model (ANM) to study the functional dynamics of the assembly of the 30S subunit of the ribosome. The optimal way to study this problem would be to compare normal modes of the crystal partial structures with the modes of the intact structure and explore conformational transitions between the open forms (partial structures) and the closed forms (in the intact structure). Although the crystal intact structure of the 30S ribosome is available, the crystal partial structures are mostly unavailable in PDB. Therefore we computationally generated these partial structures from the intact structure by removing singe protein subunits, pairs of protein subunits, and selected larger sets of protein subunits. We compared the slowest normal modes computed from the partial structures with the slowest modes from the corresponding parts in the intact structure. As more crystal partial structures will become available in PDB in the future it might be worthwhile to use such PDB data to reinvestigate the problem. We also hope that our present results may motivate further experimental studies on 30S ribosome assembly.

In summary, the effects of various subunits on the large-scale dynamics of the partial 30S ribosomal structure have been studied by removing single proteins, pairs of proteins, or larger sets of proteins at a time. From these protein removal experiments, we have found that the S6 and S18 subunits behave as a single functional block in the 30S structure, exhibited by a strong correlation of motions of S6 and S18. The existing experimental data provide additional support for our finding derived from elastic network model computations.

**Acknowledgment.** It is a pleasure to acknowledge the financial support provided by NIH grants 1R01GM073095, 1R01GM072014, and 1R01GM081680.

## References

(1) Flory, P. J. *Proc. R. Soc. London, Ser. A* **1976**, *351*, 351–380.

(2) Kloczkowski, A.; Mark, J. E.; Erman, B. *Macromolecules* **1989**, *22*, 1423–1432.

(3) Bahar, I.; Atilgan, A. R.; Erman, B. *Fold. Des* **1997**, *2*, 173–181.

(4) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. *Biophys. J.* **2001**, *80*, 505–515.

(5) Bahar, I.; Jernigan, R. L. *J. Mol. Biol.* **1998**, *281*, 871–884.

(6) Bahar, I.; Erman, B.; Jernigan, R. L.; Atilgan, A. R.; Covell, D. G. *J. Mol. Biol.* **1999**, *285*, 1023–1037.

(7) Gregory, S. T.; Lieberman, K. R.; Dahlberg, A. E. *Nucleic Acids Res.* **1994**, *22*, 279–284.

(8) Jernigan, R. L.; Bahar, I.; Covell, D. G.; Atilgan, A. R.; Erman, B.; Flatow, D. T. *J. Biomol. Struct. Dyn.* **2000**, (Sp. Iss. S1), 49–55.

(9) Ramaswamy, A.; Bahar, I.; Ioshikhes, I. *Proteins* **2005**, *58*, 683–696.

(10) Yang, L. W.; Rader, A. J.; Liu, X.; Jursa, C. J.; Chen, S. C.; Karimi, H. A.; Bahar, I. *Nucleic Acids Res.* **2006**, *34*, W24−W31.

(11) Tama, F.; Valle, M.; Frank, J.; Brooks, C. L., III *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9319–9323.

(12) Wang, Y.; Rader, A. J.; Bahar, I.; Jernigan, R. L. *J. Struct. Biol.* **2004**, *147*, 302–314.

(13) Mizushima, S.; Nomura, M. *Nature* **1970**, *226*, 1214.

(14) Held, W. A.; Ballou, B.; Mizushima, S.; Nomura, M. *J. Biol. Chem.* **1974**, *249*, 3103–3111.

(15) Held, W. A.; Mizushima, S.; Nomura, M. *J. Biol. Chem.* **1973**, *248*, 5720–5730.

(16) Culver, G. M.; Noller, H. F. *RNA* **1999**, *5*, 832–843.

(17) Powers, T.; Daubresse, G.; Noller, H. F. *J. Mol. Biol.* **1993**, *232*, 362–374.

(18) Schluenzen, F.; Tocilj, A.; Zarivach, R.; Harms, J.; Gluehmann, M.; Janell, D.; Bashan, A.; Bartels, H.; Agmon, I.; Franceschi, F.; Yonath, A. *Cell* **2000**, *102*, 615–623.

(19) Culver, G. M. *Biopolymers* **2003**, *68*, 234–249.

(20) Stagg, S. M.; Mears, J. A.; Harvey, S. C. *J. Mol. Biol.* **2003**, *328*, 49–61.

(21) Hamacher, K.; Trylska, J.; McCammon, J. A. *PLoS. Comput. Biol.* **2006**, *2*, e10.

(22) SantaLucia, J., Jr. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 1460–1465.

(23) Doruker, P.; Jernigan, R. L.; Bahar, I. *J. Comput. Chem.* **2002**, *23*, 119–127.

(24) Talkington, M. W.; Siuzdak, G.; Williamson, J. R. *Nature* **2005**, *438*, 628–632.

(25) Marques, O. A. *BLZPACK: Description and User's Guide 1995*; http://crd.lbl.gov/∼osni/#Software (accessed July 24, 2008).

(26) Ming, D.; Kong, Y.; Wakil, S. J.; Brink, J.; Ma, J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7895–7899.

(27) Lu, M. Y.; Ma, J. P. *Biophys. J.* **2005**, *89*, 2395–2401.

(28) Recht, M. I.; Williamson, J. R. *J. Mol. Biol.* **2004**, *344*, 395–407.

(29) Recht, M. I.; Williamson, J. R. *J. Mol. Biol.* **2001**, *313*, 35–48.

(30) Agalarov, S. C.; Prasad, G. S.; Funke, P. M.; Stout, C. D.; Williamson, J. R. *Science* **2000**, *288*, 107–112.

(31) Matadeen, R.; Patwardhan, A.; Gowen, B.; Orlova, E. V.; Pape, T.; Cuff, M.; Mueller, F.; Brimacombe, R.; van Heel, M. *Structure* **1999**, *7*, 1575–1583.

(32) Williamson, J. R. *Curr. Opin. Struct. Biol.* **2008**, *18*, 299–304.

CT800223G

JCTC Journal of Chemical Theory and Computation

# Remarkably Strong T-Shaped Interactions between Aromatic Amino Acids and Adenine: Their Increase upon Nucleobase Methylation and a Comparison to Stacking

Lesley R. Rutledge and Stacey D. Wetmore*

*Department of Chemistry and Biochemistry, University of Lethbridge, 4401 University Drive, Lethbridge, Alberta, Canada T1K 3M4*

Received June 19, 2008

**Abstract:** T-shaped geometries and interaction energies between select DNA nucleobases (adenine or 3-methyladenine) and all aromatic amino acids (histidine, phenylalanine, tyrosine, or tryptophan) were examined using BSSE-corrected MP2/6−31G*(0.25) potential energy surface scans, which determined the preferred nucleobase (face)–amino acid (edge) and nucleobase (edge)–amino acid (face) interactions. The energies of dimers with the strongest interactions were further studied at the CCSD(T)/CBS level of theory, which suggests that the T-shaped interactions in adenine dimers are very strong (up to −35 kJ mol$^{-1}$). Nucleobase methylation to form a cationic damaged base (3-methyladenine) plays a large role in the relative monomer orientations and magnitude of the interactions, which increase by 17−125%. Most importantly, this study is the first to compare the stacking and T-shaped interactions between all aromatic amino acids and select (natural and damaged) DNA nucleobases where the differences between stacking and T-shaped interactions at the CCSD(T)/CBS level are small. Therefore, our results indicate that T-shaped interactions cannot be ignored when studying biological processes, and this manuscript discusses the importance of these interactions in the context of DNA repair.

## 1. Introduction

Noncovalent interactions play a major role in determining structures and properties of molecular assemblies in biology, chemistry, and materials science.[1] These interactions, which include hydrogen bonding, $\pi$-$\pi$, cation-$\pi$, and X–H$\cdots\pi$ (where X=N, O, C) among others, control the design of molecular devices and govern the self-assembly of natural and artificial systems.[2,3] In biology, the dynamic interactions between phospholipid bilayers and proteins, the double helical structure of DNA, and the three-dimensional structure of proteins are all dependent on noncovalent interactions.[4,5] Many pharmaceutical ligand–protein interactions are also noncovalent. For example, some drugs, including anticancer agents, use $\pi$-$\pi$ interactions to intercalate into DNA.[6]

Noncovalent interactions between protein and DNA building blocks also play vital roles in the development of pharmaceuticals and biochemical techniques as well as in nature. For example, these interactions are essential for DNA replication,[7] transcription,[7] and DNA repair.[8,9] We are specifically interested in the role of these interactions in DNA base excision repair (BER). BER is perhaps the most important natural repair mechanism, which utilizes multiple enzymes.[8,9] Specifically, the first step in the BER process, where damaged bases are removed by enzymes (DNA glycosylases) that cleave the (glycosidic) bond connecting the base to the deoxyribose sugar,[10−13] likely relies on several noncovalent DNA−protein interactions for substrate identification and removal.

There are many classes of DNA glycosylases, which each act on damaged bases formed through different pathways (such as deamination, oxidation, or alkylation) and each use different catalytic mechanisms. For example, uracil DNA glycosylase (UDG) and formamidopyrimidine [fapy]-DNA glycosylase (FPG) use hydrogen bonds to bind and remove (neutral) damaged nucleobases formed via deamination and oxidation, respectively.[10,12] Alternatively, DNA glycosylases that repair alkylation damage, which include 3-methyladenine

* Corresponding author fax: (403)329-2057; e-mail: stacey.wetmore@uleth.ca.

10.1021/ct8002332 CCC: $40.75 
Published on Web 09/03/2008

DNA glycosylase II (AlkA, *E. coli*) and human alkyladenine DNA glycosylase (AAG, human), possess no obvious polar groups that can form strong hydrogen bonds to the cationic damaged base in the active site pockets.[12] Instead, crystal structures[14,15] show stacking (face-to-face) and T-shaped (edge-to-face) interactions between the aromatic amino acids and nucleobases bound at the active site. Therefore, it has been proposed that stacking and T-shaped interactions are responsible for substrate recognition and stabilization.[12] Specifically, these enzymes may use aromatic $\pi$-cation interactions to attract cationic damaged bases into the active site and to differentiate between damaged and natural (undamaged) bases.[16]

Although crystal structures are highly informative about biomolecular geometries, they leave many unanswered questions about the nature or magnitude of discrete interactions between substrates and enzymes. Furthermore, in the case of DNA glycosylases that repair alkylation damage, no structure of a cationic nucleotide bound to the active sites exists. This is due to difficulties with experimental synthesis of stable cationic substrates,[12] which undergo spontaneous hydrolysis more rapidly than their neutral counterparts.[12,17,18] Therefore, unanswered questions related to DNA alkylation repair enzymes include the following: What is the nature of the attractive forces between active site residues and the substrates? What is the magnitude of these interactions? How does structure affect the magnitude of these interactions? What orientations between amino acids and bases are preferred? How do the interactions differ between neutral and cationic nucleobases? These lead to more general questions such as the following: Do these interactions affect the catalytic power of the enzyme?

Computational chemistry can address some of these unanswered questions regarding the nature of active site interactions. Previously, we used computational chemistry to systematically examine the MP2/6−31G*(0.25) stacking interactions between all aromatic amino acids (histidine (HIS), phenylalanine (PHE), tyrosine (TYR), tryptophan (TRP)) and adenine (A) or 3-methyladenine (3MeA, Figure 1),[19] which is the second most common alkylation product and has been shown to stop DNA replication.[20−23] This selection of molecules allowed us to separately characterize $\pi$-$\pi$ and $\pi$-cation interactions and thereby determine the effects of nucleobase alkylation as well as the amino acid on the magnitude of stacking interactions.

Although our previous calculations revealed important information about biologically relevant stacking (face-to-face) interactions, crystal structures of AlkA and AAG show that the amino acid and substrate molecular planes are not always above/below each other in a perfectly parallel alignment. Instead, many T-shaped (edge-to-face) interactions also exist. Furthermore, these interactions have not been well characterized, and the influence of (cationic) charge on these interactions is even less understood. To fully understand the implications of these T-shaped interactions, we must also investigate the natural bases. Therefore, the T-shaped interactions between all aromatic amino acids and adenine or 3-methyladenine are systematically investigated in the present study. We consider both amino acid (edge)–nucleobase (face) and nucleobase (edge)–amino acid (face) interactions. High-level calculations (CCSD(T)) and extrapolation to the complete basis set limit were performed on dimer orientations that yield the strongest T-shaped interactions as well as the strongest stacking interactions previously reported.[24,25]

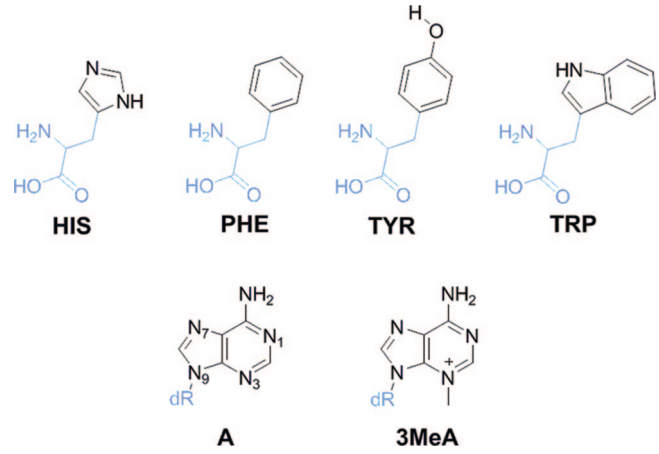



***Figure 1.*** The structure of the amino acids (histidine (HIS), phenylalanine (PHE), tyrosine (TYR) and tryptophan (TRP)) and nucleobases (adenine (A) and 3-methyladenine (3MeA)) considered in this study, where blue fragments were replaced with a hydrogen atom in our models.

In addition to revealing important information (geometries and magnitudes) about biologically relevant T-shaped interactions, this work represents the most accurate comparison in the literature of the stacking and T-shaped interactions between natural or damaged nucleobases and the aromatic amino acids. Indeed, our work will reveal the magnitude of, and differences between, stacking and T-shaped interactions that occur within the active sites of DNA repair enzymes. This study also has more general implications due to the use, and significance, of noncovalent protein–DNA interactions in a variety of biological processes. Since information about the strengths of these interactions is difficult to extract from experiment alone, it is important to study these interactions using the highest levels of theory possible.

## 2. Computational Methods

Computational chemistry has been widely used to examine noncovalent interactions. Although an abundance of studies on hydrogen bonding exist, recent emphasis has been placed on anticipated weaker stacking (face-to-face) interactions between two aromatic rings. For example, there is a wealth of literature that examines the stacking interactions of benzene or substituted benzenes[26] as well as the stacking interactions between the natural nucleobases.[27−31] Stacking interactions in amino acid dimers have also been considered.[32] However, a comparatively limited number of studies have examined nucleobase–amino acid dimers. Rooman et al. investigated $\pi$-$\pi$ or cation-$\pi$ nucleobase–amino acid (Arg, Lys, Asn, and Gln) interactions[33] and more recently investigated the stacking interactions between (neutral or protonated) histidine and adenine or phenylalanine.[34] In addition, the stacking interactions between the four aromatic amino acids and the natural nucleobases[24,35] as well as the 10 most

common (cationic) methylated nucleobases[25] have been investigated. T-shaped (X–H···$\pi$, where X=N, O, C) interactions between small molecules or molecular fragments and various aromatic rings[36−42] or biomolecular (DNA or protein) $\pi$-systems have also been examined.[43−46] However, to our knowledge, very few studies have been completed on T-shaped (edge-to-face) interactions between two aromatic rings.[26,34,47] Indeed, these studies have been primarily limited to the T-shaped dimers between benzene and substituted benzenes.[26] In terms of biological systems, Cauët et al. investigated select T-shaped orientations between adenine and phenylalanine or histidine (neutral and protonated).[34] Despite their importance, a full systematic study of interactions between the aromatic amino acids and the nucleobases has not yet been performed.

The vast literature discussed above has revealed that stacking and T-shaped interactions are sensitive to the level of theory and basis set implemented, where appropriate levels of theory to study these interactions in biological systems have been identified. We use an approach similar to that used by Hobza and Šponer to study stacking interactions between DNA nucleobases[27] and our group to study stacking interactions between aromatic amino acids and nucleobases.[19,24,25] Specifically, monomers (Figure 1) were optimized in fixed planar geometries using MP2/6−31G(d). The nucleotide monomers were modeled by replacing the sugar−phosphate backbone with a hydrogen atom, while the protein backbone and $\beta$-carbon of the amino acids were replaced by a hydrogen atom. Therefore, HIS was modeled as imidazole, PHE as benzene, TRP as indole, and TYR as phenol. The gas-phase potential energy surfaces between monomers were scanned using basis set superposition error (BSSE)-corrected MP2 single-point calculations with the 6−31G*(0.25) basis set, which replaces the standard d-exponent for second-row atoms (0.8) with 0.25.[27] MP2/6−31G*(0.25) has been previously justified for use to study 'weak' interactions and has been shown to produce the same trends as, and recover approximately 80% of, the CCSD(T) stacking energies for natural nucleobase dimers calculated at the complete basis set (CBS) limit.[29] This manuscript will show an even better agreement between MP2 and CCSD(T) for nucleobase–amino acid stacking and T-shaped binding energies. During the potential energy surface scans, the relative orientations of the monomers were varied as a function of different variables as outlined in the following sections.

**2.1. Stacking Interactions.** Previously, we scanned the BSSE-corrected MP2/6−31G*(0.25) gas-phase potential energy surface of nucleobase–amino acid dimers by stacking monomers (face-to-face) with respect to their centers of mass.[19,24,25] Two relative orientations of the molecular planes were considered, where the first is defined by stacking the amino acid and nucleobase in the orientation shown in Figure 1 and the second, denoted as flipped using the subscript f, is obtained by flipping the amino acid relative to Figure 1 prior to stacking with the nucleobase. Three variables were investigated that define the relative orientation between the nucleobases and amino acids (Figure 2a): vertical separation ($R_1$), angle of rotation ($\alpha$), and horizontal displacement ($R_2$). Our methodology for considering these variables is discussed

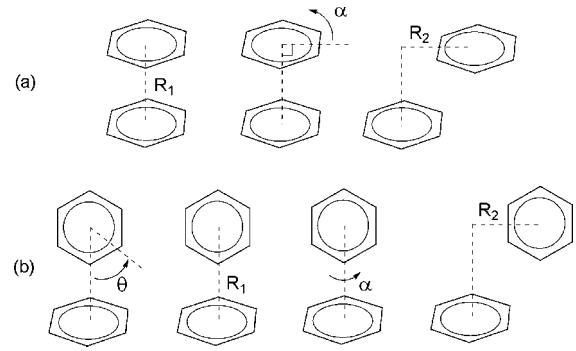



***Figure 2.*** The definition of the variables considered in (a) previous stacking potential energy surface scans (vertical separation ($R_1$), angle of rotation ($\alpha$), and horizontal displacement ($R_2$))[24,25] and (b) the present T-shaped potential energy surface scans (angle of 'edge' rotation ($\theta$), vertical separation ($R_1$), angle of rotation ($\alpha$), and horizontal 'edge' displacement ($R_2$)).

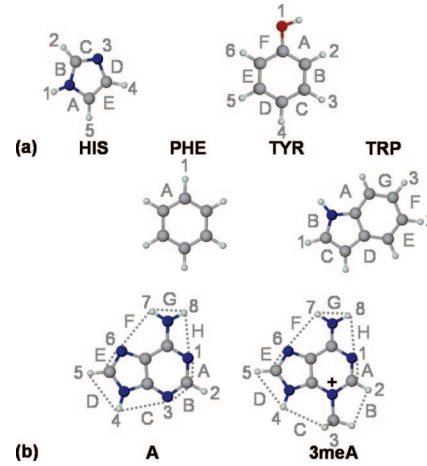



***Figure 3.*** The definition of $\theta$ for (a) amino acid edges and (b) nucleobase edges considered in potential energy surface scans.

in detail in our previous publications.[24,25] In the present study, the strongest (most negative) MP2/6−31G*(0.25) stacking energies are reported, and the corresponding geometries were used for higher-level calculations (discussed below).

**2.2. T-Shaped Interactions.** To characterize a different part of the same potential energy surface examined in our previous study on nucleobase–amino acid stacking interactions, we used a series of gas-phase BSSE-corrected MP2/6−31G*(0.25) single-point calculations to identify the strongest T-shaped interactions between each amino acid and A or 3MeA. Four variables were considered (Figure 2b). First, the angles of 'edge' rotation ($\theta$) were chosen, which define the ring edge (monomer edge) directed toward the center of mass of the $\pi$-system (monomer face). Figure 3 shows the amino acid and base edges considered. Our nomenclature uses numbers to indicate the atom directed toward the center of mass of the $\pi$-system and letters to indicate a bridged structure involving more than one atom directed toward the $\pi$-system. For example, in dimers involving a PHE edge, $\theta$=1 indicates a hydrogen from the PHE ring is directed at the center of mass of the nucleobase,

while $\theta$=A indicates a C−C bond of the benzene ring is parallel to the nucleobase molecular plane.[48] In total, 2 different edges were considered for PHE,[48] 10 for HIS and TRP, 12 for TYR, and 16 for A and 3MeA. Therefore, 196 different monomer orientations were considered (i.e., 34 amino acid edges directed toward each of two nucleobases and 32 nucleobase edges directed toward each of four amino acids). We emphasize that in addition to identifying the strongest T-shaped interactions between each amino acid and A or 3MeA, we have characterized interactions that may not correspond to the global minimum but may be important in biological systems due to natural dynamics or structural constraints of proteins or DNA.

The initial structures for dimers involving an amino acid edge were obtained by aligning the centers of mass of the amino acid and nucleobase and setting the molecular planes perpendicular. For dimers involving a nucleobase edge, the edge is sometimes located off the amino acid $\pi$-system when initial structures with aligned centers of mass are considered due to the smaller size of the amino acids. Therefore, when a nucleobase atom edge was directed toward the amino acid face ($\theta$ = number, Figure 3b), the atom was placed directly on top of the center of mass of the amino acid in the initial structure. Alternatively, when a nucleobase bond was set parallel to the amino acid face ($\theta$ = letter, Figure 3b), a dummy atom was placed at the midpoint between the atoms linked by the dotted lines in Figure 3b, and the dummy atom was aligned with the amino acid center of mass. In all T-shaped calculations, $\alpha$=0° was defined as the structure with the monomer face in the XY plane, where the model glycosidic (nucleobase face) or $\beta$-carbon and peptide backbone (amino acid face) bond is parallel to the Y-axis. The monomer edge is placed in the YZ plane with the molecular plane parallel to the glycosidic (nucleobase face) or $\beta$-carbon and peptide backbone (amino acid face) bond.

Once the edges ($\theta$) were chosen and $\alpha$=0° was defined, the vertical separation distance ($R_1$) was altered by 0.1 Å increments along the Z-axis. For dimers involving the amino acid edge, $R_1$ is the distance between the center of mass of the two monomers. For the nucleobase edge dimers, $R_1$ is the distance between the amino acid center of mass and the nucleobase atom ($\theta$ = number) or the dummy atom at the midpoint of the line connecting the two atoms that define the monomer edge ($\theta$ = letter). Once the preferred vertical separation was determined, $R_1$ was held fixed in the remaining calculations.

Next, the angle of rotation ($\alpha$) was altered by rotating the monomer edge in 30° increments in the right-hand sense. For dimers involving the amino acid edge, the rotation axis passes through the centers of mass of both monomers. For the nucleobase edge dimers, the rotational axis passes through the center of mass of the amino acid and the nucleobase atom ($\theta$ = number) or the dummy atom defining the midpoint of the line connecting the two relevant base atoms ($\theta$ = letter).

Finally, the horizontal 'edge' displacement ($R_2$) was considered. Due to the large number of calculations required to completely scan the monomer faces (81 calculations for HIS face and up to 225 calculations for 3MeA face), only the edges (up to 7) that lead to the strongest interactions after varying $R_1$ and $\alpha$ for each monomer pair were considered in $R_2$ scans. To perform the $R_2$ shift, the center of mass of the monomer face (in the XY plane) was defined as the origin (0,0). For all dimers, the Y-axis was defined to be parallel to the glycosidic (nucleobase face) or $\beta$-carbon and peptide backbone (amino acid face) bond, and this bond lies in quadrant III. The monomer edge was shifted by 0.5 Å along the X and Y axis, where single-point calculations were completed at each increment over the entire monomer face. Thus, despite the reduction in the number of edges considered for $R_2$ scans, dimers involving an amino acid edge still required approximately 225 (PHE edge) to 900 (TRP edge) calculations per nucleobase–amino acid pair, while dimers involving a nucleobase edge required approximately 425 (PHE face) to 700 (TRP face) calculations per pair.

**2.3. Higher-Level ab Initio Methods and Extrapolation Techniques.** To verify our computational approach, higher-level calculations were performed on all dimers yielding the strongest (most negative) interaction energies as identified from the MP2/6−31G*(0.25) potential energy surface scans. Since extrapolation techniques have been shown to accurately estimate the stacking interactions between the natural DNA nucleobases at the limit of large basis sets and high levels of correlation,[30] a similar approach was used in the present study to approximate CCSD(T)/CBS results. Specifically, the Helgaker basis set extrapolation technique was implemented,[49] which has been used previously for many different systems and has been specifically shown to work well for T-shaped interactions in general[38,39,45,47] as well as stacking interactions between the DNA nucleobases.[29e,f,30a,b,31c,32d] In our work, extrapolation from the aug-cc-pVDZ and aug-cc-pVTZ basis sets was used to estimate the MP2/CBS level. Previous results for hydrogen bonded and stacked DNA and RNA nucleobases showed that this extrapolation scheme was only improved by 2 kJ mol$^{-1}$ when increased to the aug-cc-pVTZ and aug-cc-pVQZ extrapolation,[32d] and calculations at the aug-cc-pVQZ were not feasible given our current computer resources for the size of the complexes examined in the present study. A $\Delta$(CCSD(T)–MP2) correlation correction factor was subsequently evaluated using the 6−31G*(0.25) basis set and added to the MP2/CBS binding strengths to yield estimated CCSD(T)/CBS results. Previous work supports this extrapolation approach for the correlation effects, where the 6−31G*(0.25) basis set was determined to yield a satisfactory $\Delta$(CCSD(T)–MP2) correction for the natural DNA nucleobases due to the basis set insensitivity of this correction.[29e,f,30a,b,31c,32d]

All energy calculations include basis set superposition error (BSSE) corrections.[50] All MP2 calculations were performed using Gaussian 03,[51] while CCSD(T) calculations were completed using MOLPRO.[52]

## 3. Results and Discussion

**3.1. T-Shaped Interactions.** As mentioned in the Introduction, one of the driving forces of the present study is to address unanswered questions regarding the nature of T-shaped interactions between the aromatic amino acids and

natural versus damaged nucleobases. Our approach for scanning the potential energy surface as a function of four variables allows us to understand the dependence of the interaction energy on the relative monomer orientations, and therefore we will begin by discussing these dependencies in the following section. Subsequently, the geometry of the preferred T-shaped complexes for adenine and 3-methyladenine will be discussed and compared, and important structural characteristics that optimize T-shaped interaction energies will be summarized. Next, the dependence of T-shaped interaction energies on the cationic charge introduced upon nucleobase methylation will be highlighted, where results from our MP2/6−31G*(0.25) scans will be validated using CCSD(T) calculations extrapolated to the complete basis set limit.

*3.1.1. Dependence of T-Shaped Interactions on Monomer Orientations.* Since a large number of data points (up to 900) were considered in the potential energy surface scans for each monomer pair, we focus our discussion on major conclusions and use select examples to illustrate our findings. Data tables in the Supporting Information summarize the optimal MP2/6−31G*(0.25) T-shaped interactions after consideration of each variable (Figure 2b), and the overall strongest (most negative) interaction energies for each dimer after consideration of all variables are presented in Table 1. For nucleobase edge dimers, two interaction energies are reported in Table 1: (1) the overall strongest interaction energy, which involves the model N−H glycosidic bond for all dimers, and (2) the strongest interaction that does not involve the model glycosidic bond (denoted as 'enzymatic'). The latter energies are also reported since the hydrogen atom in the glycosidic bond is replaced with a sugar in DNA, and part of our goal is to understand the magnitude of these interactions in DNA repair enzymes. For amino acid edge dimers, only one interaction energy is reported in Table 1 since the strongest interaction does not involve the bond where the $\beta$-carbon and protein backbone are attached.

A direct comparison of the $R_1$ distances (Supporting Information) for dimers involving each amino acid or nucleobase edge ($\theta$, Figure 3) is not meaningful since the definition of $R_1$ changes depending on the type of dimer examined. Instead, the most important conclusion from our $R_1$ scans is that the interaction energies are not largely dependent upon changes in the vertical separation, which was also previously reported for stacking interactions in the same nucleobase–amino acid systems.[19,24,25] Specifically, the interaction energy changes by less than 1.4 kJ mol$^{-1}$ when $R_1$ deviates from the optimum interaction orientation by 0.1 Å.

Using the optimal vertical separations, Figure 4 illustrates the dependence of T-shaped interactions on α for HIS complexes as a representative example. For complexes involving a HIS edge, the dependence on α was found to range between 2.6 kJ mol$^{-1}$ ($\theta$=1) and 8.4 kJ mol$^{-1}$ ($\theta$=B) for (neutral) A complexes (Figure 4a) and between 2.7 kJ mol$^{-1}$ ($\theta$=1) and 20.0 kJ mol$^{-1}$ ($\theta$=C) for (cationic) 3MeA complexes (Figure 4b). Similarly, when an A edge interacts with the HIS face, the largest effect of α is 20 kJ mol$^{-1}$ ($\theta$=8, Figure 4c), while the largest effect for a cationic 3MeA

***Table 1.*** Strongest MP2 and CCSD(T) Interaction Energies (kJ mol$^{-1}$) between Adenine or 3-Methyladenine and the Four Aromatic Amino Acids Calculated with a Variety of Basis Sets[a,b]

| | adenine | | | | | | 3-methyladenine | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MP2/6−31G*(0.25) | MP2/aug-cc-pVDZ | MP2/aug-cc-pVTZ | MP2/CBS | CCSD(T)/6−31G*(0.25) | CCSD(T)/CBS | MP2/6−31G*(0.25) | MP2/aug-cc-pVDZ | MP2/aug-cc-pVTZ | MP2/CBS | CCSD(T)/6−31G*(0.25) | CCSD(T)/CBS |
| HIS edge | −22.5 | −23.8 | −25.1 | −25.7 | −19.2 | −22.4 | −61.6 | −63.2 | −64.2 | −64.6 | −58.7 | −61.6 |
| PHE edge | −14.1 | −17.1 | −18.1 | −18.5 | −10.8 | −15.2 | −16.8 | −20.0 | −21.2 | −21.7 | −13.0 | −17.8 |
| TYR edge | −21.9 | −22.9 | −23.7 | −24.1 | −20.3 | −22.4 | −33.9 | −35.2 | −36.5 | −37.0 | −32.2 | −35.2 |
| TRP edge | −23.2 | −25.3 | −26.6 | −27.2 | −18.9 | −22.9 | −28.0 | −31.4 | −32.7 | −33.1 | −23.0 | −28.2 |
| HIS face[c] | −33.6 (−22.6) | −34.4 (−23.8) | −35.8 (−25.2) | −36.4 (−25.9) | −30.5 (−20.3) | −33.3 (−23.6) | −64.5 (−43.1) | −63.6 (−43.1) | −66.6 (−44.5) | −67.9 (−45.1) | −61.6 (−41.4) | −65.1(−43.5) |
| PHE face[c] | −25.6 (−16.0) | −27.5 (−17.8) | −30.3 (−19.3) | −31.5 (−20.0) | −20.5 (−13.5) | −26.4 (−17.4) | −46.4 (−37.2) | −47.1 (−36.0) | −49.3 (−38.2) | −50.3 (−39.2) | −41.7 (−33.3) | −45.6 (−35.4) |
| TYR face[c] | −27.8 (−18.4) | −29.8 (−20.5) | −31.9 (−21.8) | −32.8 (−22.4) | −23.3 (−16.0) | −28.3 (−20.0) | −54.1 (−39.2) | −54.6 (−40.8) | −57.2 (−42.9) | −58.3 (−43.8) | −48.7 (−36.2) | −52.9 (−40.8) |
| TRP face[c] | −34.8 (−23.1) | −37.5 (−25.3) | −39.8 (−26.6) | −40.8 (−27.1) | −28.8 (−20.0) | −34.8 (−24.0) | −72.0 (−52.6) | −72.8 (−54.6) | −75.8 (−57.0) | −77.2 (−58.1) | −64.8 (−48.4) | −70.1 (−53.9) |
| HIS stacked | −27.2[d] | −33.5 | −35.6 | −36.4 | −18.5 | −27.8 | −52.7[e] | −55.2 | −57.4 | −58.3 | −43.6 | −49.3 |
| $HIS_t$ stacked | −29.7[d] | −35.7 | −37.7 | −38.5 | −20.9 | −29.8 | −51.7[e] | −54.4 | −56.6 | −57.6 | −42.6 | −48.4 |
| PHE stacked | −24.3[d] | −31.5 | −33.7 | −34.6 | −13.0 | −23.3 | −48.6[e] | −52.8 | −56.1 | −57.5 | −35.3 | −44.2 |
| TYR stacked | −30.7[d] | −38.2 | −40.2 | −41.1 | −19.6 | −30.0 | −55.0[e] | −59.0 | −62.3 | −63.7 | −42.2 | −50.9 |
| $TYR_t$ stacked | −28.9[d] | −36.2 | −38.4 | −39.3 | −18.0 | −28.4 | −53.9[e] | −57.8 | −61.0 | −62.4 | −41.1 | −49.6 |
| TRP stacked | −35.0[d] | −42.4 | −44.1 | −44.8 | −23.6 | −33.5 | −69.7[e] | −73.6 | −76.9 | −78.3 | −55.7 | −64.2 |
| $TRP_t$ stacked | −32.0[d] | −39.9 | −41.7 | −42.4 | −20.2 | −30.6 | −71.5[e] | −75.5 | −78.6 | −79.9 | −57.1 | −65.5 |

[a] The strongest interactions for both T-shaped and stacked complexes were determined through MP2/6−31G*(0.25) potential energy surface scans (see Figure 2 for variables considered in scans). [b] Extrapolation to the complete basis set limit was completed by the Helgaker scheme (see Computational Methods). [c] The nucleobase edge interactions involving the model glycosidic bond with 'enzymatic' interactions not involving the model glycosidic bond, which are more relevant to biological processes involving nucleosides and nucleotides, provided in parentheses. [d] Reference 24. [e] Reference 25.

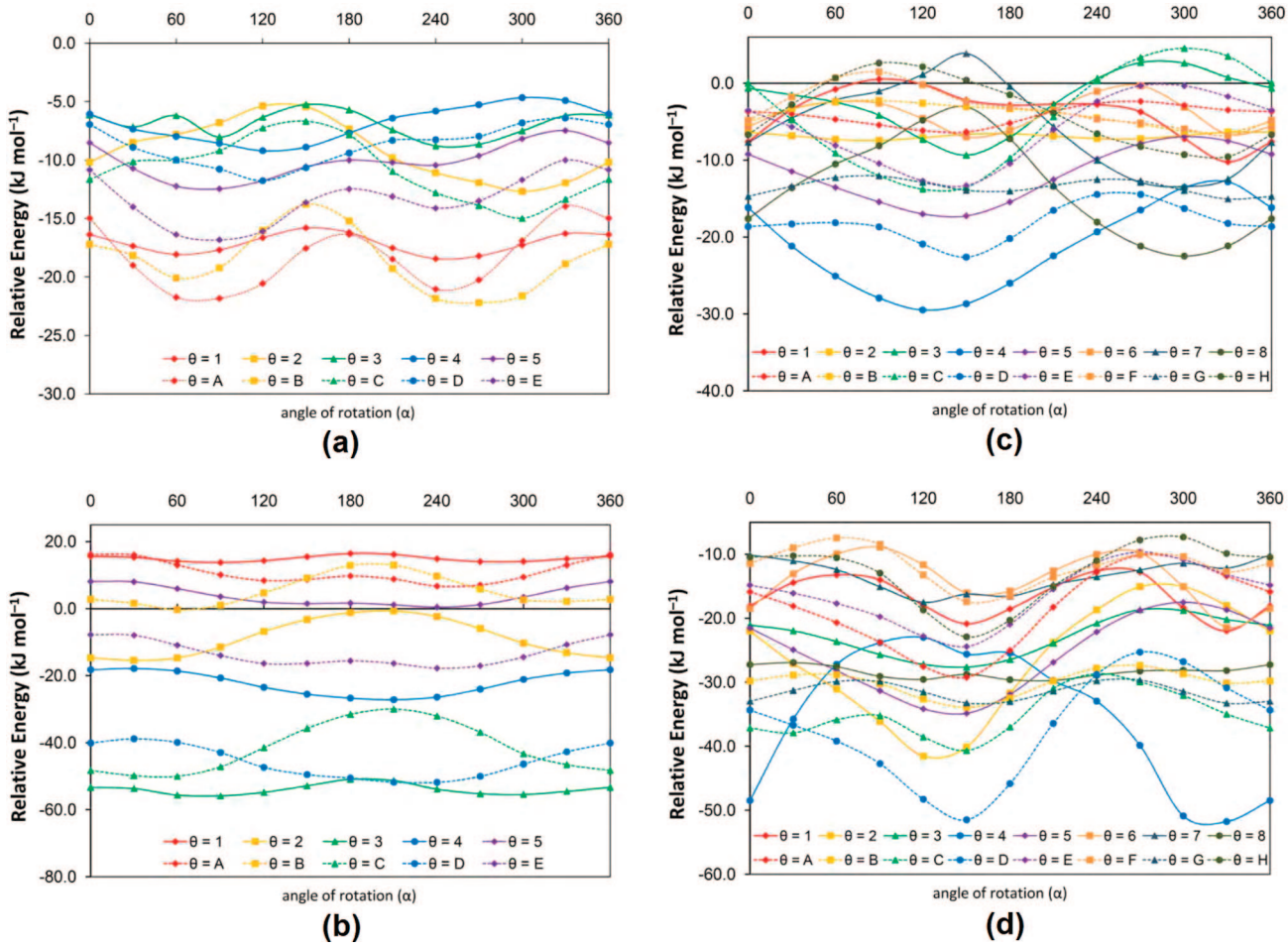



***Figure 4.*** Interaction energy between HIS and (a) adenine (face), (b) 3-methyladenine (face), (c) adenine (edge), and (d) 3-methyladenine (edge) as a function of the angle of rotation (α) for different edges ($\theta$) (see Figure 2b and Figure 3 for variable and edge definitions, respectively).

edge is 29 kJ mol$^{-1}$ ($\theta$=4, Figure 4d). Therefore, although the potential energy surface for rotation about α is shallow for many dimers, the interaction energy has a larger dependence on the angle of rotation (α) compared to the vertical separation.

For any given $\theta$, the dependence on the angle of rotation (α) is due to the strength of secondary intramolecular interactions, where the strongest interaction arises when the acid−base interactions between monomer edge and monomer face are maximized. To illustrate this point, Figure 5 shows the orientations with the strongest interactions (after considering $R_1$ and α) for HIS complexes, where the secondary intramolecular interactions that govern the optimal α alignment are highlighted with dotted lines. In HIS edge complexes (Figure 5a,b), the preferred orientation aligns HIS protons toward N1 and N7 of the nucleobase, which are the sites with the largest proton affinity.[53] Alternatively, in nucleobase (edge) complexes, the best α aligns the electron-rich N atoms and/or the acidic N−H of HIS with strong nucleobase proton donors and/or acceptors, respectively. In all cases, the strongest interaction also occurs for the α that best directs the monomer edge across the entire $\pi$-system of the monomer face and thereby maximizes monomer overlap.

Although the interaction energies change with the angle of rotation (α) by up to 29 kJ mol$^{-1}$, the interaction energy has a larger dependence on $\theta$. This can be seen in Figure 4 since the separation between lines (dependence on $\theta$) is larger

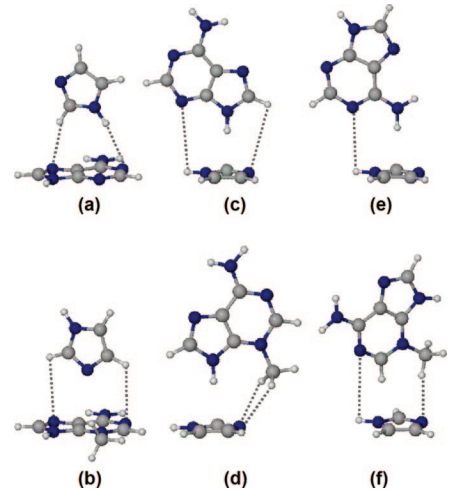



***Figure 5.*** HIS T-shaped dimers with (a) adenine (face), (b) 3-methyladenine (face), (c) adenine (edge involving the model glycosidic bond), (d) 3-methyladenine (edge involving the model glycosidic bond), (e) adenine ('enzymatic' edge, not involving model glycosidic bond), and (f) 3-methyladenine ('enzymatic' edge, not involving model glycosidic bond) for optimal $\theta$ after considering $R_1$ and α.

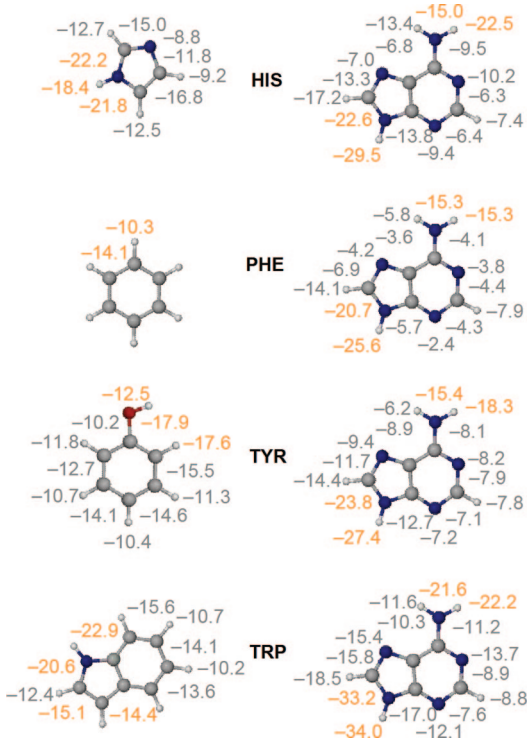



***Figure 6.*** The strongest interactions for each $\theta$ in adenine (face)–amino acid (edge) (left) and adenine (edge)–amino acid (face) (right) dimers after considering $R_1$ and $\alpha$. Edges highlighted in orange were considered for $R_2$ shifts.

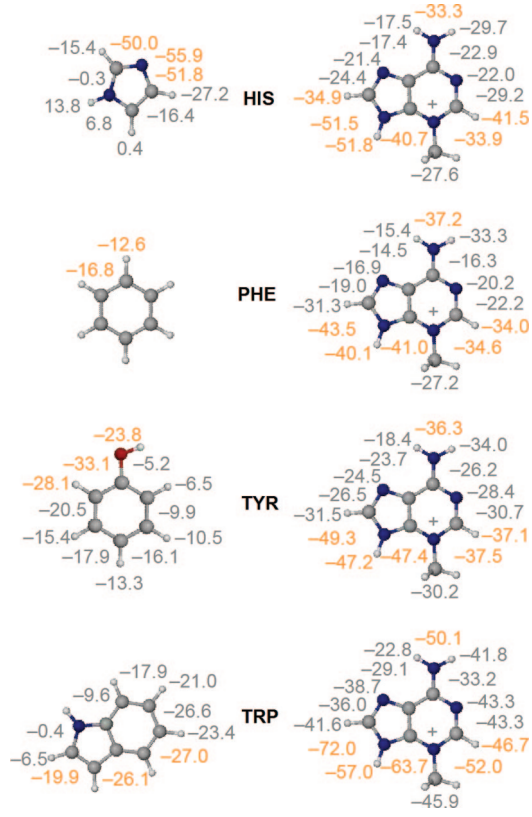



***Figure 7.*** The strongest interactions for each $\theta$ in 3-methyladenine (face)–amino acid (edge) (left) and 3-methyladenine (edge)–amino acid (face) (right) dimers after considering $R_1$ and $\alpha$. Edges highlighted in orange were considered for $R_2$ shifts.

than variations within a line (dependence on $\alpha$). Figures 6 and 7 report the strongest (most negative) interaction energy (after considering $R_1$ and $\alpha$) for each $\theta$ in adenine and 3-methyladenine complexes, respectively. The range in the interaction energy as a function of $\theta$ falls between 3.3 and 55.5 kJ mol$^{-1}$ for each monomer. This large effect is partially due to variations in the properties of the monomer edges. The dependence of the interaction energy on $\theta$ will be discussed in greater detail in the following section.

Since we are in part searching the potential energy surface to determine the strongest interaction, the edges with the optimal interactions after consideration of $R_1$, $\alpha$, and $\theta$ (highlighted in Figures 6 and 7 with orange) were subsequently examined by varying the $R_2$ horizontal displacement. We find that the horizontal displacement generally does not strengthen the T-shaped interactions by a considerable amount, and performing $R_2$ shifts does not change the preferred $\theta$ edge. These results justify our decision to consider $R_2$ effects on only select $\theta$ for each complex. More specifically, for the 64 complexes considered, most of the interaction energies increase (become more negative) by less than 1 kJ mol$^{-1}$ and only 22 increase by 2−13 kJ mol$^{-1}$. Furthermore, the $R_2$ shifts that yield the strongest interaction energies are very small, where only 13 structures involve an $R_2$ shift greater than 1.0 Å in any direction.

Figure 8 shows how the potential energy surface changes with $R_2$ for HIS complexes. The lowest (most negative) energy region (yellow) is typically centered on or close to the center of mass (purple circles). The preference for small $R_2$ shifts and the resulting small strengthening in the T-shaped interaction energy arise since $\theta$ and $\alpha$ already optimize secondary intramolecular interactions that lead to the strongest T-shaped interactions in most dimers. In cases where the $R_2$ shift is larger (1.0 Å), the shift better optimizes these interactions. For example, in the HIS(edge):3MeA(face) complex (Figure 8b), the $R_2$ shift moves the HIS lone pair across the face of 3-methyladenine toward the atom with the largest positive charge (C4). Similarly, for both the A or 3MeA(edge):HIS(face) dimer (Figure 8c,d,f), the $R_2$ shift moves acidic nucleobase bonds toward the basic N atom of the HIS ring.

In summary, among the four geometrical variables considered, the T-shaped interactions are most dependent on the monomer edge ($\theta$). Therefore, this key structural feature will be discussed in more detail in the following section.

*3.1.2. Dependence of Optimal T-Shaped Structure on Nucleobase Methylation.* The first important conclusion about the geometries of T-shaped complexes is that, regardless of the nucleobase methylation (charged) state, a stronger interaction is generally observed for dimers with the monomer edge bridging the $\pi$-system ($\theta$ = letter) compared to edges involving a single atom ($\theta$ = number) due to greater overlap. Although early studies on the benzene dimer have also identified bridged structures to be the most stable T-shaped orientations,[26a] the majority of recent studies on T-shaped complexes have only considered interactions that involve a particular atom directed at the aromatic $\pi$-system.[34,38,46] In cases where bridged structures involving small molecules directed toward an aromatic ring are considered,[45] only select

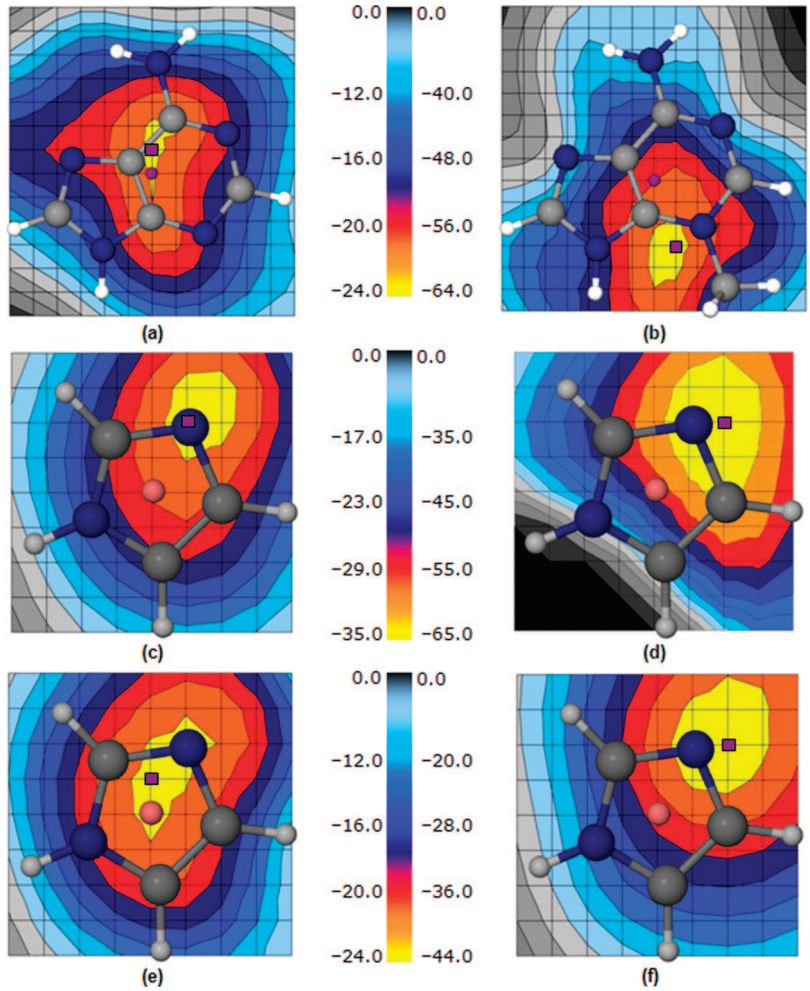



***Figure 8.*** Interaction energy (kJ mol$^{-1}$) as a function of $R_2$ shift for optimal $R_1$, $\alpha$ and $\theta$ orientations of HIS T-shaped dimers with (a) adenine (face), (b) 3-methyladenine (face), (c) adenine (edge involving model glycosidic bond), (d) 3-methyladenine (edge involving model glycosidic bond), (e) adenine ('enzymatic' edge, not involving model glycosidic bond), and (f) 3-methyladenine ('enzymatic' edge, not involving model glycosidic bond). Purple circles indicate the origin (center of mass) and purple squares indicate the point with the strongest interaction.

combinations of molecules are examined, and no systematic investigation has been done. Furthermore, studies of T-shaped interactions involving amino acids have not considered bridged structures.[34] Our results clearly indicate that to identify global minima, and fully understand these interactions, the bridged structures must be examined.

The second important conclusion regarding the T-shaped structures is that the favored bridged orientation depends on both the properties of the monomer edge and the monomer face. For dimers with an amino acid edge interacting with the face of A, Figure 6 reveals that the most acidic (or positive) edge of the amino acid prefers to be directed toward the electron-rich face of adenine. For example, in the HIS edge complex, the optimal structure directs the acidic N−H bond and a C−H bond toward adenine ($\theta$=B). This is consistent with T-shaped structures reported by Tsuzuki et al. for benzene–pyridine complexes.[47] Similarly, TYR bridges an acidic O−H bond and a C−H bond ($\theta$=A) about the center of mass of adenine. This result is consistent with the preferred structure of the freely optimized phenol–benzene T-shaped dimer,[26m] which also validates our potential energy surface scans.

For dimers with an amino acid edge interacting with the face of 3MeA, we find that the most basic (or negative) edge of the amino acid prefers to be directed toward the cationic nucleobase. The best example is the HIS complex, where directing the HIS lone pair toward the cationic face ($\theta$=3) results in the largest (most negative) interaction energy (−61.6 kJ mol$^{-1}$). Alternatively, when an acidic edge of HIS ($\theta$=1) is directed toward 3MeA, the interaction is extremely repulsive (+13.8 kJ mol$^{-1}$). This is opposite to the trend discussed for adenine, where the $\theta$=1 acidic edge leads to a stronger interaction (−18.4 kJ mol$^{-1}$) than the $\theta$=3 basic edge (−8.8 kJ mol$^{-1}$). Indeed, the optimal orientation for HIS edge interacting with adenine ($\theta$=B) leads to a very small interaction with 3-methyladenine (−0.3 kJ mol$^{-1}$). Similarly, for the TYR edge, the most stable complexes involve the hydroxyl lone pair directed toward 3MeA ($\theta$=F) but the acidic hydroxyl hydrogen directed toward A ($\theta$=A). These results show that the cationic charge of the damaged nucleobase dictates the relative orientation of the amino acid and base and therefore plays a large role in the nature of T-shaped interactions.

Due to the differences in charge between adenine and 3-methyladenine, it is not surprising that the interaction energies have a different dependence on the amino acid edge. Indeed, the strongest interactions for adenine face dimers increase as PHE < TYR < HIS < TRP, which is due to the relative dipole moments and size of the $\pi$-system of the various aromatic amino acids, while the strengths of dimers involving 3-methyladenine face increase as PHE < TRP < TYR < HIS, which is due to the increasing basicity of the amino acid. Since the favored monomer orientation in nucleobase face dimers depends on the relative acidity and basicity of the amino acid edge, a larger amino acid dipole moment causes a larger variation in the interaction energies as a function of $\theta$,[54] which suggests that electrostatics play a very important role in these T-shaped contacts.

For complexes involving adenine or 3-methyladenine edges, the optimal interaction occurs with the most acidic edge of the nucleobase, the model glycosidic bond ($\theta$=4 (adenine and 3MeA(edge):HIS(face) dimer) or $\theta$=D (3-methyladenine)), directed toward the electron-rich amino acid face. Our observations for 3-methyladenine edge dimers are consistent with previous research on the benzene (face)–pyridinium cation (edge) interactions,[47] where the strongest T-shaped complex directs the N−H bond of pyridinium toward the $\pi$-system of benzene.[55] For the 'enzymatic' interactions, the strongest binding occurs between the amino group of A ($\theta$=8) and the amino acid $\pi$-system. Although the amino group of 3MeA ($\theta$=G) interacts with PHE in the most stable complex, two C−H$\cdots\pi$ interactions ($\theta$=B) yield the strongest binding with the remaining amino acid faces. Due to the similarity in the structures of adenine and 3-methyladenine edges, the optimal interactions for amino acid face dimers increase with the dipole moments as well as the size of the $\pi$-system of the aromatic amino acids (PHE < TYR < HIS < TRP) in the 'strongest' and 'enzymatic' complexes for both nucleobases.

*3.1.3. Dependence of T-Shaped Interaction Energies on Nucleobase Methylation.* The discussion in the previous section reveals that the optimal T-shaped structures of dimers involving an amino acid edge are highly dependent upon the nature of the nucleobase. These differences in key structural features have large implications for the relative strengths of nucleobase–amino acid dimers. Indeed, adenine interactions range between −14 and −35 kJ mol$^{-1}$, while 3-methyladenine interactions are even larger, ranging between −17 and −72 kJ mol$^{-1}$.

To further examine the magnitude of the nucleobase–amino acid T-shaped interactions and validate our MP2/6−31G*(0.25) results, the T-shaped interactions for structures with the strongest (most negative) interactions as found in the MP2/6−31G*(0.25) potential energy surface scans were estimated at the CCSD(T)/CBS limit. Table 1 displays the CCSD(T)/CBS estimates as well as the binding strengths calculated at all levels of theory required to perform the extrapolation.

Previous studies on stacking interactions show that MP2 binding strengths increase with the basis set size,[26,27] and we see the same trend for T-shaped interactions. Thus, in comparison to 6−31G*(0.25), MP2/CBS leads to a 8−32% increase in the T-shaped interaction energies for adenine and 3−30% for 3-methyladenine. Nevertheless, MP2 is known to overestimate the stacking interaction energies of the natural nucleobases,[29e,f,30a,b,31c,32d] and MP2/6−31G*(0.25) overestimates the CCSD(T)/6−31G*(0.25) correlation energy of the T-shaped interactions examined in this study by 1−8 kJ mol$^{-1}$. However, CCSD(T)/CBS and MP2/6−31G*(0.25) results deviate by only 0−2 kJ mol$^{-1}$ for all T-shaped dimers. Therefore, the MP2/6−31G*(0.25) calculations account for 91−105% of the CCSD(T)/CBS T-shaped interaction energies. This is an even better agreement between MP2/6−31G*(0.25) and CCSD(T)/CBS T-shaped energies than previously reported for the stacking interactions between the natural nucleobases, where only 80% of the correct stacking interaction is recovered at the MP2/6−31G*(0.25) level.[29] This justifies our choice of MP2/6−31G*(0.25) for the potential energy surface scans as a balance between cost and accuracy. Furthermore, our MP2/6−31G*(0.25) results are reliable for understanding the relative magnitude and importance of a wide range of nucleobase–amino acid interactions that do not necessarily correspond to global minima but may be imposed in biological systems.

For adenine complexes, the strongest CCSD(T)/CBS amino acid edge interactions range between −15 and −23 kJ mol$^{-1}$. When the edges of adenine are considered, the strongest T-shaped interactions range between −26 and −35 kJ mol$^{-1}$, while the 'enzymatic' interactions range between −17 and −24 kJ mol$^{-1}$. Most importantly, the magnitude of the T-shaped interactions between adenine and the amino acids are up to −35 kJ mol$^{-1}$, which is much larger than anticipated.

As mentioned for the MP2/6−31G*(0.25) results, methylation has a large effect on T-shaped interactions at the CCSD(T)/CBS level. Specifically, the strongest CCSD(T)/CBS amino acid edge interactions with 3-methyladenine range between −17 and −62 kJ mol$^{-1}$. Any given amino acid edge complex increases in strength by 17−176% upon methylation of adenine at N3. The largest methylation effect occurs for TYR and HIS complexes (57% and 176%, respectively) due to a strong interaction between an amino acid lone pair and the 3-methyladenine face. Indeed, the HIS lone pair–$\pi$(cation) interaction is the strongest T-shaped interaction involving HIS and is consistent with a previously published pyridinium (edge)–benzene (face) interaction (−61.7 kJ mol$^{-1}$), which was described as a hydrogen bond since the interaction energy is stronger than that of the stacked structure and the origin of the attraction is mainly electrostatic.[47] The magnitude of this HIS lone pair–$\pi$(cation) interaction is also consistent with the magnitude of similar interactions calculated by Egli et al. for various biological systems.[56]

The strongest (most negative) CCSD(T)/CBS T-shaped interactions involving a 3-methyladenine edge range between −45 and −70 kJ mol$^{-1}$, while the 'enzymatic' interactions range between −35 and −54 kJ mol$^{-1}$. Although there are not large changes in the preferred nucleobase orientation upon methylation as discussed for amino acid edge dimers, the cationic charge increases the acidity of the nucleobase bonds interacting with the amino acid face, which results in

stronger interaction energies. Indeed, the effect of methylation in these complexes corresponds to a 63−125% increase, which is larger than discussed for amino acid edge complexes. Thus, although the adenine interactions are strong, methylation has a large effect on the T-shaped interactions.

**3.2. Comparison of the Magnitude of T-Shaped and Stacking Interactions between Adenine or 3-Methyladenine and the Aromatic Amino Acids.** Due to the remarkable magnitude of T-shaped interactions between the amino acids and natural or damaged nucleobases, a direct comparison of these interactions to the previously calculated stacking interactions between the same molecules is necessary to understand their relative impact on biological processes. However, we must first extrapolate the previously reported MP2/6−31G*(0.25) stacking energies[24,25] to the CCSD(T)/CBS level of theory, where the extrapolated binding strengths and results from all intermediate calculations are provided in Table 1. We find that MP2/6−31G*(0.25) generally overestimates the stacking interaction energy by 0.5−6 kJ mol$^{-1}$ (or up to 10%) when compared to CCSD(T)/CBS, where only 2 dimers decrease in strength by 0.1 and 0.5 kJ mol$^{-1}$. Therefore, MP2/6−31G*(0.25) is useful for scanning the potential energy surfaces of stacked complexes between the amino acids and natural or damaged nucleobases. However, the Δ(CCSD(T)−MP2) difference is larger for stacking (0−6 kJ mol$^{-1}$) than T-shaped (0−2 kJ mol$^{-1}$) interactions. Furthermore, CCSD(T)/CBS strengthens the T-shaped interactions (by up to 5%) and weakens the stacking (by up to 10%). These differences indicate that the strength of the stacking and T-shaped interactions are even more similar at the CCSD(T)/CBS level of theory compared with MP2/6−31G*(0.25), and therefore it is crucial to compare the T-shaped and stacking interactions of nucleobase–amino acid dimers at the CCSD(T)/CBS level.

CCSD(T)/CBS interaction energies for adenine show that the largest (most negative) T-shaped interactions are generally stronger than stacking interactions by 1.4 (TRP) − 3.5 (HIS) kJ mol$^{-1}$. In the case of TYR, the stacking interaction is only slightly larger (more negative by 1.7 kJ mol$^{-1}$) than the strongest T-shaped interaction. Furthermore, adenine 'enzymatic' interactions are only 21% (HIS) − 34% (TYR) weaker than stacking interactions. Similarly, adenine face T-shaped interactions are 25% (HIS) − 35% (PHE) smaller than the stacking interactions. These results emphasize that T-shaped interactions between the aromatic amino acids and the natural nucleobases are equivalent in magnitude or only slightly weaker than stacking interactions.

As discussed for adenine, the strongest 3-methyladenine T-shaped interactions are larger (more negative) than stacking interactions by 1.4 (PHE) − 5.0 (HIS) kJ mol$^{-1}$. 'Enzymatic' interactions involving a 3MeA edge are slightly weaker than stacking interactions, where the largest decrease is 20% (for PHE and TYR). Since 3-methyladenine amino acid edge interactions are weaker than the corresponding nucleobase edge interactions, they are also weaker than the corresponding stacking interactions (by up to 60% for PHE). However, the HIS edge interaction in the HIS:3MeA complex is 12 kJ mol$^{-1}$ stronger than the corresponding stacking interaction. The implications and general importance of the relative magnitude of T-shaped and stacking interactions will be discussed in the following section.

**3.3. Summary and Importance of Protein–DNA Noncovalent Interactions.** By systematically examining the potential energy surfaces for protein–DNA noncovalent stacking and T-shaped interactions at very high levels of theory, we have found that although T-shaped interactions are often believed to be weak and relatively insignificant, these interactions can be very close in magnitude to $\pi$-stacking interactions between the same monomers. This statement is further justified by surveys of the protein data bank. For example, Rooman et al. identified several different HIS:A contacts in a range of X-ray crystal structures, and approximately 40% of these correspond to T-shaped arrangements.[34] Furthermore, although only two crystal structures of DNA repair enzymes that remove alkylated nucleobases are available with bound substrates, both crystal structures show a range of active site stacking and T-shaped interactions.[14,15]

Both stacking and T-shaped interactions are close in magnitude to biologically relevant hydrogen bonds. Indeed, lone pair–$\pi$ interactions involving a positively charged nucleobase can exceed the strength of a single hydrogen bond, which has been noted previously for a range of biological systems.[56] For example, the adenine-thymine Watson-Crick hydrogen-bond strength, which involves at least two strong hydrogen bonds, is estimated to be −70 kJ mol$^{-1}$ at the CCSD(T)/CBS level.[32d] Even T-shaped structures that do not correspond to the strongest interaction for a given amino acid–nucleobase dimer can be quite strong, and these structures may sometimes be more relevant to biological processes where dynamics or geometrical constraints prohibit optimal monomer orientations. Thus, our results indicate that the T-shaped interactions between nucleobases and aromatic amino acids can provide stability to many different enzymatic systems, including those involved in DNA transcription, replication, and repair, and these interactions cannot be ignored when studying biological processes.

In addition to gaining a better understanding of the magnitude of T-shaped interactions, our study reveals how the relative monomer orientation governs these interactions and thereby leads to extremely attractive forces. Specifically, we find that the strongest interactions between neutral monomers occur when the most acidic bond is directed toward the electron-rich $\pi$-system, while the strongest interactions for cationic monomers occur when a basic monomer edge is directed toward the positively charged $\pi$-system. These results, in conjunction with crystallographic data, can help clarify the interactions observed in the active sites of enzymes. Since we are ultimately interested in the role of noncovalent interactions in the DNA repair process facilitated by DNA glycosylases, we will illustrate this point by considering the active site of AAG.[15] Specifically, the protonation state of an active site HIS is not clear from crystallographic data, and our study suggests that the preferred HIS edge interaction will direct the lone pair of the basic N atom in HIS toward the cationic damaged nucleobase. Furthermore, our calculations indicate that this orientation is less favorable for the undamaged base and

therefore may help the enzyme differentiate between damaged and natural bases. Similarly, our calculations suggest that the hydroxyl group hydrogen of a TYR in the AAG active site is directed away from the face of the damaged nucleobase to maximize a lone pair–$\pi$ interaction.[15] Thus, our calculations provide clues about how DNA repair enzymes can use T-shaped interactions to govern their activity by, for example, selectively removing cationic damaged nucleobases. These examples confirm the hypothesis of Egli et al. that lone pair–$\pi$ interactions in particular may serve as a reporter of unusual protonation states of nucleobases[56] and also suggest that other T-shaped interactions involving cationic nucleobases can behave in a similar way.

Our calculations show that both stacking and T-shaped interactions are dependent on the amino acid. We previously reported that the stacking interactions of adenine and 3-methyladenine increase with the dipole moment of the amino acid, which induces an electrostatic interaction and suggests the importance of long-range dispersion interactions.[19,24,25] Similarly, this study reveals that T-shaped adenine and 3-methyladenine dimers also have strong electrostatic interactions. Therefore, PHE was found to have the weakest stacking and T-shaped interactions among all amino acids. Furthermore, whether a stacking or T-shaped interaction is favored depends on the amino acid. Therefore, we can speculate about the most beneficial interactions in, for example, the active sites of DNA glycosylases for substrate identification. Specifically, we can compare the magnitude of stacking, amino acid edge, and nucleobase 'enzymatic' edge interactions. Our calculations suggest that stacking interactions will be slightly more favorable in the case of PHE, TRP, and TYR, while T-shaped interactions with the HIS edge are stronger than the corresponding stacking interactions. Based on these global conclusions, it is interesting to note that close examination of the crystal structure of AlkA[14] with neutral hypoxanthine bound in the active site reveals that TRP and TYR are stacked with the substrate. Furthermore, the crystal structure of AAG[15] shows the substrate held in the active site by a T-shaped interaction with the HIS edge and stacking interactions with TYR, as well as T-shaped interactions with the TYR edge (which our calculations predict are only slightly weaker than the TYR stacking interactions).

One of the main goals of our study was to understand the effects of methylation on the stacking and T-shaped interactions. We find that stacking interactions increase by 60−115% upon methylation, while T-shaped interactions increase by 17−176% for amino acid edge dimers and 84−125% for nucleobase edge dimers. Indeed, since the differences between T-shaped and stacking interactions are generally small, our results suggest that alkylation has a greater effect on the interaction energies than the type of noncovalent interaction occurring between the substrate and active site residues or the amino acid involved. This general conclusion is especially true for nucleobase edge and stacking interactions. Although the strength of amino acid edge interactions is more dependent on the amino acid, the nucleobase face alkylation state still drastically affects the interaction energy. Thus, this paper shows that both the geometry and the magnitude of these noncovalent interactions are greatly affected by nucleobase alkylation. Therefore, our work suggests that alkylation must play a key role in dictating the active site interactions used by DNA repair enzymes for substrate recognition and binding and validates the range of amino acids found in the active sites of these enzymes.

## 4. Conclusions

T-shaped interactions between all aromatic amino acids and adenine or 3-methyladenine were systematically investigated to gain information about the geometric constraints that govern these interactions as well as their magnitude. To our knowledge, this work represents the most detailed study of T-shaped interactions between two different aromatic systems. Our study reveals that T-shaped interactions are highly dependent on the nature of the monomer edge, where the strongest T-shaped interaction generally occurs when the monomer edge bridges the $\pi$-system using two atoms rather than directing a single atom at the aromatic system. This result is crucial for future studies that wish to determine the global minimum for T-shaped monomer orientations, where the majority of past studies neglect these orientations. Furthermore, the favored monomer edge depends on the properties of the monomer face, where the most acidic edge is directed toward the (neutral) $\pi$-systems of adenine and the amino acids, while the most basic edge is directed toward (cationic) 3-methyladenine. Thus, the methylation state governs the preferred orientation of the amino acid and base, which provides clues about interactions within the active sites of enzymes that repair DNA alkylation damage and reveals ways to identify the alkylation (or protonation) state of nucleobases in DNA−protein systems.

The most significant result of the present work revolves around the magnitude of T-shaped interactions. After extrapolation to the highest level of theory possible for these systems, we find that T-shaped interactions involving adenine are up to $-35$ kJ mol$^{-1}$, which is comparable to the strength of stacking interactions between the corresponding monomers and also comparable to biologically relevant hydrogen bonds evaluated at the same high level of theory. Furthermore, T-shaped interactions involving 3-methyladenine are up to $-70$ kJ mol$^{-1}$, which is also similar to the stacking interactions between the corresponding monomers but much larger than the corresponding adenine interactions (up to 176% increase upon methylation).

To the best of our knowledge, this is the first study to compare noncovalent stacking and T-shaped interactions of biological systems at the CCSD(T)/CBS level with an extensive potential energy search, where our results emphasize the importance of comparing these interactions at the highest level of theory possible. Due to the magnitude of these interactions, our calculations suggest that it is crucial to examine T-shaped interactions to understand biological processes, where we have specifically applied our findings to better understand DNA repair enzymes. Future work must consider environmental effects on our findings as well as a greater range of natural and damaged nucleobases.

**Acknowledgment.** We thank the Natural Sciences and Engineering Research Council (NSERC), the Canada Research Chair program, and the Canada Foundation for Innovation (CFI) for financial support. We also thank the Upscale and Robust Abacus for Chemistry in Lethbridge (URACIL) for computer resources. L.R.R. thanks NSERC and the Alberta Ingenuity Fund (AIF) for student scholarships.

**Supporting Information Available:** MP2/6−31G*-(0.25) interaction energies as well as optimal $R_1$, $\alpha$, and $R_2$ values for each $\theta$ in all adenine and 3-methyladenine complexes. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Hunter, C. A. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 1584–1586.

(2) Lehn, J.-M. *Supramolecular Chemistry: Concepts and Perspectives;* VCH: New York, 1995; pp 89−195..

(3) Meyer, E. A.; Castellano, R. K.; Diederich, F. *Angew. Chem., Int. Ed.* **2003**, *42*, 1210–1250.

(4) Dziubek, K.; Podsiadło, M.; Katrusiak, A. *J. Am. Chem. Soc.* **2007**, *129*, 12620–12621.

(5) Nishio, M. *Cryst. Eng. Comm.* **2004**, *6*, 130–158.

(6) Brana, M. F.; Cacho, M.; Gradillas, A.; Pascual-Teresa, B.; Ramos, A. *Curr. Pharm. Des.* **2001**, *7*, 1745–1780.

(7) Chen, Y. C.; Wu, C. Y.; Lim, C. *Proteins: Struct., Funct., Bioinformat.* **2007**, *67*, 671–680.

(8) Seeburg, E.; Eide, L.; Bjoras, M. *Trends Biochem. Sci.* **1995**, *20*, 391–397.

(9) Wood, R. D.; Mitchell, M.; Sgouros, J.; Lindahl, T. *Science* **2001**, *291*, 1284–1289.

(10) Berti, P. J.; McCann, J. A. B. *Chem. Rev.* **2006**, *106*, 506–555.

(11) Stivers, J. T.; Drohat, A. C. *Arch. Biochem. Biophys.* **2001**, *396*, 1–9.

(12) Stivers, J. T.; Jiang, Y. L *Chem. Rev.* **2003**, *103*, 2729–2759.

(13) David, S. S.; Williams, S. D. *Chem. Rev.* **1998**, *98*, 1221–1261.

(14) Teale, M.; Symersky, J.; DeLucas, L. *Bioconjugate Chem.* **2002**, *13*, 403–407.

(15) Lau, A. Y.; Wyatt, M. D.; Glassner, B. J.; Samson, L. D.; Ellenberger, T. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 13573–13578.

(16) Labahn, J.; Scharer, O. D.; Long, A.; EzazNikpay, K.; Verdine, G. L.; Ellenberger, T. E. *Cell* **1996**, *86*, 321–329.

(17) Fujii, T.; Saito, T.; Nakasaka, T. *Chem. Pharm. Bull.* **1989**, *37*, 2601–2609.

(18) Fujii, T.; Itaya, T. *Heterocycles* **1988**, *48*, 1673–1724.

(19) Rutledge, L. R.; Campbell-Verduyn, L. S.; Hunter, K. C.; Wetmore, S. D. *J. Phys. Chem. B* **2006**, *110*, 19652–19663.

(20) Robertson, K. D.; Jones, P. A. *Carcinogenesis* **2000**, *21*, 461–467.

(21) Drabløs, F.; Feyzi, E.; Aas, P. A.; Vaagbø; C. B.; Kavli, B.; Bratlie, M. S.; Pena-Diaz, J.; Otterlei, M.; Slupphaug, G.; Krokan, H. E. *DNA Repair* **2004**, *3*, 1389–1407.

(22) Mishina, Y.; Duguid, E. M.; He, C. *Chem. Rev.* **2006**, *106*, 215–232.

(23) Wyatt, M. D.; Allan, J. M.; Lau, A. Y.; Ellenberger, T. E.; Samson, L. D. *Bioessays* **1999**, *21*, 668–676.

(24) Rutledge, L. R.; Campbell-Verduyn, L. S.; Wetmore, S. D. *Chem. Phys. Lett.* **2007**, *444*, 167–175.

(25) Rutledge, L. R.; Durst, H. F.; Wetmore, S. D. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2801–2812.

(26) (a) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Am. Chem. Soc.* **1994**, *116*, 3500–3506. (b) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Phys. Chem.* **1996**, *100*, 18790–18794. (c) Williams, V. E.; Lemieux, R. P.; Thatcher, G. R. J. *J. Org. Chem.* **1996**, *61*, 1927–1933. (d) Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 10887–10893. (e) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104–112. (f) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2003**, *107*, 8377–8379. (g) Sinnokrot, M. O.; Sherrill, C. D. *J. Am. Chem. Soc.* **2004**, *126*, 7690–7697. (h) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200–10207. (i) Riley, K. E.; Merz, K. M., Jr. *J. Phys. Chem. B* **2005**, *109*, 17752–17756. (j) Lee, E. C.; Hong, B. H.; Lee, J. Y.; Kim, J. C.; Kim, D.; Kim, Y.; Tarakeshwar, P.; Kim, K. S. *J. Am. Chem. Soc.* **2005**, *127*, 4530–4537. (k) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2006**, *110*, 10656–10668. (l) Waller, M. P.; Robertazzi, A.; Platts, J. A.; Hibbs, D. E.; Williams, P. A. *J. Comput. Chem.* **2006**, *27*, 491–504. (m) Lee, E. C.; Kim, D.; Jureèka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. *J. Phys. Chem. A* **2007**, *111*, 3446–3457. (n) Arnstein, S. A.; Sherrill, C. D. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2646–2655. (o) Quiñonero, D.; Frontera, A.; Escudero, D.; Ballester, P.; Costa, A.; Deyà, P. M. *Theor. Chem. Acc.* **2008**, *120*, 385–393.

(27) (a) Hobza, P.; Šponer, J. *Chem. Rev.* **1999**, *99*, 3247–3276. (b) Šponer, J.; Leszczynski, J.; Hobza, P. *J. Mol. Struct. (Theochem)* **2001**, *573*, 43–53. (c) Šponer, J.; Leszczynski, J.; Hobza, P. *Biopolymers (Nucleic Acid Sci.)* **2002**, *61*, 3–31. (d) Hobza, P. *Annu. Rep. Prog. Chem., Sec. C: Phys. Chem.* **2004**, *100*, 3–27.

(28) (a) Hobza, P.; Šponer, J.; Polasek, M. *J. Am. Chem. Soc.* **1995**, *117*, 792–798. (b) Šponer, J.; Leszczynski, J.; Hobza, P. *J. Phys. Chem. A* **1997**, *101*, 9489–9495. (c) Kratochvil, M.; Engkvist, O.; Šponer, J.; Jungwirth, P.; Hobza, P. *J. Phys. Chem. A* **1998**, *102*, 6921–6926.

(29) (a) Šponer, J.; Leszczynski, J.; Hobza, P. *J. Phys. Chem.* **1996**, *100*, 5590–5596. (b) Šponer, J.; Hobza, P. *Chem. Phys. Lett.* **1997**, *267*, 263–270. (c) Jureèka, P.; Nachtigall, P.; Hobza, P. *Phys. Chem. Chem. Phys.* **2001**, *3*, 4578–4582. (d) Leininger, M. L.; Nielsen, I. M. B.; Colvin, M. E.; Janssen, C. L. *J. Phys. Chem. A* **2002**, *106*, 3850–3854. (e) Jureèka, P.; Šponer, J.; Hobza, P. *J. Phys. Chem. B* **2004**, *108*, 5466–5471. (f) Šponer, J.; Jureèka, P.; Marchan, I.; Javier Luque, F.; Orozco, M.; Hobza, P. *Chem. Eur. J.* **2006**, *12*, 2854–2865.

(30) (a) Hobza, P.; Šponer, J. *J. Am. Chem. Soc.* **2002**, *124*, 11802–11808. (b) Jureèka, P.; Hobza, P. *J. Am. Chem. Soc.* **2003**, *125*, 15608–15613. (c) Šponer, J.; Jureèka, P.; Hobza, P. *J. Am. Chem. Soc.* **2004**, *126*, 10142–10151.

(31) (a) Šponer, J.; Florian, J.; Ng, H.-L.; Šponer, J. E.; Spadkova, N. A. *Nucleic Acids Res.* **2000**, *28*, 4893–4902. (b) Hill, G.; Forde, G.; Hill, N.; Lester, W. A.; Sokalski, W. A.; Leszczynski, J. *Chem. Phys. Lett.* **2003**, *381*, 729–732. (c) Dabkowska, I.; Gonzalez, H. V.; Jureèka, P.; Hobza, P. *J. Phys. Chem. A* **2005**, *109*, 1131–1136. (d) Cysewski, P.;

Czyznikowska-Balcerak, Z. *J. Mol. Struct. (Theochem)* **2005**, *757*, 29–36. (e) Matta, C. F.; Castillo, N.; Boyd, R. J. *J. Phys. Chem. B* **2006**, *110*, 563–578. (f) Kabelac, M.; Sherer, E. C.; Cramer, C. J.; Hobza, P. *Chem.-Eur. J.* **2007**, *13*, 2067–2077.

(32) (a) Mitchell, J. B.; Nandi, C. L.; McDonald, I. K.; Thornton, J. M.; Price, S. L. *J. Mol. Biol.* **1994**, *239*, 315–331. (b) Chelli, R.; Gervasio, F. L.; Procacci, P.; Schettino, V. *J. Am. Chem. Soc.* **2002**, *124*, 6133–6143. (c) Morozov, A. V.; Misura, K. M. S.; Tsemekhman, K.; Baker, D. *J. Phys. Chem. B* **2004**, *108*, 8489–8496. (d) Jureèka, P.; Šponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(33) (a) Rooman, M.; Lievin, J.; Buisine, E.; Wintjens, R. *J. Mol. Biol.* **2002**, *319*, 67–76. (b) Biot, C.; Buisine, E.; Kwasigroch, J. M.; Wintjens, R.; Rooman, M. *J. Biol. Chem.* **2002**, *277*, 40816–40822. (c) Biot, C.; Buisine, E.; Rooman, M. *J. Am. Chem. Soc.* **2003**, *125*, 13988–13994. (d) Wintjens, R.; Biot, C.; Rooman, M.; Lievin, J. *J. Phys. Chem. A* **2003**, *107*, 6249–6258. (e) Biot, C.; Wintjens, R.; Rooman, M. *J. Am. Chem. Soc.* **2004**, *126*, 6220–6221. (f) Marsili, S.; Chelli, R.; Schettino, V.; Procacci, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2673–2685.

(34) Cauët, E.; Rooman, M.; Wintjens, R.; Lievin, J.; Biot, C. *J. Chem. Theory Comput.* **2005**, *1*, 472–483.

(35) Cysewski, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2636–2645.

(36) Zhang, R. B.; Somers, K. R. F.; Kryachko, E. S.; Nguyen, M. T.; Zeegers-Huyskens, T.; Ceulemans, A. *J. Phys. Chem. A* **2005**, *109*, 8028–8034.

(37) Vaupel, S.; Brutschy, B.; Tarakeshwar, P.; Kim, K. S. *J. Am. Chem. Soc.* **2006**, *128*, 5416–5426.

(38) Shibasaki, K.; Fujii, A.; Mikami, N.; Tsuzuki, S. *J. Phys. Chem. A* **2006**, *110*, 4397–4404.

(39) Bendová, L.; Jureèka, P.; Hobza, P.; Vondrášek, J. *J. Phys. Chem. B* **2007**, *111*, 9975–9979.

(40) Mishra, B. K.; Sathyamurthy, N. *J. Phys. Chem. A* **2007**, *111*, 2139–2147.

(41) Cheney, B. V.; Schulz, M. W.; Cheney, J.; Richards, W. G. *J. Am. Chem. Soc.* **1988**, *110*, 4195–4198.

(42) Tsuzuki, S.; Honda, K.; Fujii, A.; Uchimaru, T.; Mikami, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2860–2865.

(43) Gervasio, F. L.; Chelli, R.; Marchi, M.; Procacci, P.; Schettino, V. *J. Phys. Chem. B* **2001**, *105*, 7835–7846.

(44) Scheiner, S.; Kar, T.; Pattanayak, J. *J. Am. Chem. Soc.* **2002**, *124*, 13257–13264.

(45) Ringer, A. L.; Figgs, M. S.; Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2006**, *110*, 10822–10828.

(46) Gil, A.; Branchadell, V.; Bertran, J.; Oliva, A. *J. Phys. Chem. A* **2007**, *111*, 9372–9379.

(47) Tsuzuki, S.; Mikami, M.; Yamada, S. *J. Am. Chem. Soc.* **2007**, *129*, 8656–8662.

(48) We note that initial calculations were completed where the PHE edge was rotated about its center of mass every 5° to give 5 dimers between $\theta$=1 and $\theta$=A. Since maximum and minimum structures were found at $\theta$=1 and A, respectively, only these two extremes were modeled for the remaining dimers.

(49) (a) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243–252. (b) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Olsen, J. *Chem. Phys. Lett.* **1999**, *302*, 437–446.

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

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

(52) Werner, H. J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schutz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. *MOLPRO, Version 2006.1*; University College Cardiff Consultants Ltd.: Cardiff, U.K., 2006.

(53) McConnell, T. L.; Wheaton, C. A.; Hunter, K. C.; Wetmore, S. D. *J. Phys. Chem. A* **2005**, *109*, 6351–6362.

(54) The MP2/6−31G(d) dipole moments of the amino acids decrease as HIS (3.949 D) > TRP (1.926 D) > TYR (1.480 D).

(55) We note that Tsuzuki et al. (ref 47) did not find a bridged structure similar to our $\theta$=D edge despite free optimizations. However, their input geometries also did not consider a bridged structure. Our results suggest that geometry optimizations starting from the bridged structure would lead to a minimum on the potential energy surface that is lower in energy.

(56) Egli, M.; Sarkhel, S. *Acc. Chem. Res.* **2007**, *40*, 197–205.

CT8002332

JCTC Journal of Chemical Theory and Computation

# Influence of Nitroxide Spin Labels on RNA Structure: A Molecular Dynamics Simulation Study

Hang Yu,[†] Yuguang Mu,*[,†] Lars Nordenskiöld,[†] and Gerhard Stock*[,‡]

*School of Biological Sciences, Nanyang Technological University, Singapore 637551, and Institute of Physical and Theoretical Chemistry, J. W. Goethe University, D-60438 Frankfurt, Germany*

Received July 8, 2008

**Abstract:** Pulsed electron double resonance (PELDOR) experiments on oligonucleotides provide a distance ruler that allows the measurement of nanometer distances accurately. The technique requires attachment of nitroxide spin labels to the nucleotides, which may possibly perturb its conformation. To study to what extent nitroxide spin labels may affect RNA structure, all-atom molecular dynamics simulations in explicit solvent are performed for six double-labeled RNA duplexes. A new parametrization of the force field for the nitroxide spin label is developed, which leads to intramolecular distances that are in good agreement with experimental results. Comparison of the results for spin-labeled and unlabeled RNA reveals that the conformational effect of the spin label depends significantly on whether the spin label is attached to the major or the minor groove of RNA. While major-groove spin labeling may to some extent affect the conformation of nearby base pairs, minor-groove spin labeling has the advantage of mostly preserving the RNA conformation.

## Introduction

Nucleic acid conformations have been extensively studied since the discovery of central dogma in biochemistry.[1,2] As one of the key molecules in biological processes, RNA has gained major interest in its structure and folding, especially after the discovery of gene regulation function in RNA interference.[3] Together with an increasing number of resolved structures, this may facilitate the use of RNA as a potential target in drug design.[4]

As crystallography may only partly represent the structures and dynamics in the cellular environment, alternative spectroscopic methods like electron paramagnetic resonance (EPR),[5] nuclear paramagnetic resonance (NMR),[6] and fluorescence resonance energy transfer (FRET)[7] can provide complementary information on structures in solution that resemble biological conditions. EPR spectroscopy has been an effective method for structural characterization of RNA.[8−10] In particular, pulsed electron double resonance (PELDOR) can measure intramolecular distances ranging from 14 to 50 Å with 1 Å precision.[11,12] To be detectable by EPR, labels such as nitroxide spin labels (see Figure 1) need to be attached to RNA.[13] Various approaches to spin labeling of nucleotides have been proposed.[8,12,14−16] Site-directed spin label attachment to the 2′-sugar position of nucleic acids have been found to lead to relatively broad distance distributions.[8] Using a flexible C—S linkage attached to backbone phosphorothioate allows easy and fast synthesis. A simulation study of spin-labeled DNA duplexes with a flexible C—S linkage demonstrated that such spin labels exert little perturbation to the structure of attached DNA,[16] but the flexible linkage might influence the result of nanometer ruler

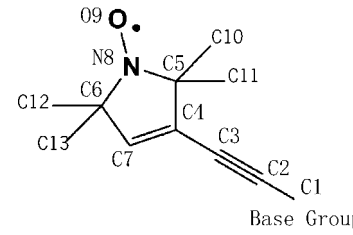



***Figure 1.*** Structure and atom labeling of nitroxide spin label group.

* Corresponding author e-mail: ygmu@ntu.edu.sg (Y.M.) and stock@theochem.uni-frankfurt.de (G.S.).

[†] Nanyang Technological University.

[‡] J. W. Goethe University.

10.1021/ct800266e CCC: $40.75 
Published on Web 09/09/2008

***Table 1.*** Sequences of slRNA and cRNA[a]

| RNA | sequence | RNA | sequence |
|---|---|---|---|
| **1** | 5′GCUGAUA**U**CAGC | **4** | 5′GCUGUAU**A**CAGC |
| | 3′CGAC**U**AUAGUCG | | 3′CGAC**A**AUAUGUCG |
| **2** | 5′GACUGAUCAG**U**C | **5** | 5′CGUGUAUGCAUA**C**ACG |
| | 3′C**U**GACUAGUCAG | | 3′GCA**C**AUACGUAUGUGC |
| **3** | 5′CGACUGAUAUCAG**U**CG | **6** | 5′CGCUACAUAG**U**GAGCG |
| | 3′GC**U**GACUAUAGUCAGC | | 3′GCG**A**UGUAUCACUCGC |

[a] Spin-labeled nucleotides are underlined. cRNA 1 and 4 are exactly the same sequence as spin-labeled RNA 1 and 4, respectively, but without spin label.

measurement.[15] Recently, a new labeling method via a rigid R−C≡C−C linkage at the nucleic acid base has been proposed, which facilitates accurate PELDOR measurements of distances between two labels in a range of 19−53 Å.[12,14] For example, distance measurements in various RNA duplexes have facilitated the discrimination of A- and B-form conformation.[14]However, it is not yet clear to what extent the rigid linker at the nucleic acid base will influence the structural and dynamical properties of the RNA attached.

To fully assess the potential of this spin labeling method and the associated PELDOR measurements, the conformational flexibility of the labels and their influence on RNA structure need to be investigated. To this end, we perform all-atom molecular dynamics (MD) simulations in explicit solvent for six double-labeled RNA duplexes. Due to improved nucleic acid force fields and appropriate description of the electrostatic, MD studies have emerged as a versatile tool to study the structure and dynamics of RNA systems in atomistic detail.[17−25] A new parametrization of force field for the nitroxide spin label is developed, which represents an improvement of our previous work.[12] We show that the new model leads to intramolecular RNA distances that are in good agreement with existing experimental results. Comparison of the results for spin-labeled and unlabeled RNA reveals that the conformational effect of the spin label depends significantly on whether the spin label is attached to the major or the minor groove of RNA. While major-groove spin labeling may affect the conformation of nearby base pairs, minor-groove spin labeling has the advantage of mostly preserving the RNA conformation.

## Methods

**1. RNA Systems.** Following the experimental investigations,[14] we studied six duplex RNAs (RNA1 to RNA6) of various lengths and sequences, see Table 1. We considered two types of spin labels, which differ in their position relative to RNA. Spin labels attached at the C5 position of pyrimidine bases are located in the *major* groove of RNA. This concerns spin labels at uracil bases in spin-labeled (sl) RNA1−RNA3 and of cytosine bases in slRNA5. On the other hand, spin labels attached to purine bases at the C2 position are located in the *minor* groove of RNA. This is the case for the spin labels on adenine bases in slRNA4. Finally, we considered a mixed case, slRNA6, which contains a labeled adenine located in the minor groove and a labeled uracil in the major groove. To compare the behavior of labeled and unlabeled RNA, we also studied unlabeled control RNA1 and RNA4, referred to as cRNA1 and cRNA4.

**2. Force Field.** In all simulations, the GROMACS (version 3.3) software package[26] combined with the Amber98 force field[27] was used. For slRNA1 and slRNA4 an additional 50 ns simulations were performed, using the recent Amber parametrization parmbsc0[28] which avoids artificial transitions along the backbone torsion angles $\alpha$ and $\gamma$.[29,30] However, the average structural properties such as the A-form percentage obtained for both force fields were quite similar. This indicates that our general results concerning the perturbation of RNA conformation by spin labels does not depend on the nucleic acid force field.

The force field parameters for the nitroxide spin label (Figure 1) were obtained via Hartree−Fock (HF) calculations with 6−31G* basis set using Gaussian03 program.[31] Adopting AMBER force field parametrization philosophy,[32] all parameters of the bonded interactions of the spin label were generated, see Table 2. Partial charges were derived from the HF/6−31G* calculations and fitted with the RESP algorithm using the R.E.D program.[33] Standard AMBER force field parameters were assumed for the remaining nonbonded interactions. The resulting force field parameters for the nitroxide spin label represent an improvement over our previous simpler parametrization,[12] which assumed a rigid artificial long bond between C1 and C4 (Figure 1). When applied to reproduce the quantitative orientation and distance distribution of model biradicals,[34] the previous force field was found to behave too rigidly because of these simplifications. The new parameters are largely consistent with the parametrization recently published by Darian and Gannett.[35] However, two bond angles on the five-member ring, C5−C4−C7 and C4−C7−C6 (in their nomenclature C3−C4−C5 and C1−C5−C4), are different. Our values are 112.3° and 112.9° and their values are both 120°, resulting in a sum of the five ring bond angles of 539.9° (our model) and 553.82° (their model). So our model better describes the planar nature of the ring.

**3. MD Simulation Details.** All RNA systems considered (labeled slRNA1−slRNA6 as well as unlabeled cRNA1 and cRNA4) were solvated in a rectangular box of TIP3P water,[36] keeping a minimum distance of 10 Å between the solute and each face of the box. Sodium counterions were added to neutralize the system, and water molecules overlapping with ions were removed. The numbers of total atoms in the simulation box are 25151, 21557, 31937, 22236, 32823, 32237 for slRNA1−slRNA6, respectively. The equation of motion was integrated by using a leapfrog algorithm[37] with a time step of 2 fs. Covalent bond lengths involving hydrogen atoms were constrained by the procedure SHAKE[38] with a relative geometric tolerance of 0.0001. We used a particle-mesh Ewald treatment for the long-range electrostatics.[39] The pair list of nonbonded interaction was updated every 10 fs. The solute and solvent were separately weakly coupled to external temperature baths at 300 K.[40] The temperature coupling constant was 0.5 ps. The total system was weakly coupled to an external pressure bath at 1 atm using a coupling constant of 5 ps.

Starting from canonical A-form, all RNAs were minimized *in vacuo* for 1000 steps to relax the initial structural constrains. After minimization, the RNAs were solvated in

***Table 2.*** Force Field Parameters for Spin Label

| atoms | C1 | C2 | C3 | C4 | C5 | C6 | C7 | N8 | O9 |
|---|---|---|---|---|---|---|---|---|---|
| charges(e) | −0.0710 | −0.1787 | −0.0443 | −0.0202 | 0.2853 | 0.2312 | −0.2147 | 0.1769 | −0.4037 |
| type | C2 | CZ | CZ | CM | CT | CT | CM | NA | O |
| atoms | HC7 | C10 | H1C10 | C11 | H1C11 | C12 | H1C12 | C13 | H1C13 |
| charges(e) | 0.1702 | −0.2023 | 0.0592 | −0.2023 | 0.0592 | −0.2948 | 0.0876 | −0.2948 | 0.0876 |
| type | H5 | CT | HT | CT | HT | CT | HT | CT | HT |
| bonds | C1−C2 | C2−C3 | C3−C4 | C4−C5 | C5−N8 | N8−O9 | C6−N8 | C6−C7 | C4−C7 |
| $l_{bond}$ (nm) | 0.144 | 0.12 | 0.144 | 0.15 | 0.146 | 0.125 | 0.146 | 0.15 | 0.1318 |
| $k_{bond}$(kJ/nm$^2$) | 251040 | 292880 | 251040 | 265265 | 355682 | 241000 | 355681 | 224262 | 265265 |
| bonds | C5−C10 | C5−C11 | C6−C12 | C6−C13 | C10−H1C10 | | | | |
| $l_{bond}$ (nm) | 0.153 | 0.153 | 0.153 | 0.153 | 0.108 | | | | |
| $k_{bond}$(kJ/nm$^2$) | 251040 | 251040 | 251040 | 251040 | 307106 | | | | |

| angles | C1−C2−C3 | C2−C3−C4 | C3−C4−C5 | C3−C4−C7 | C4−C5−N8 | C5−N8−O9 | C5−N8−C6 | C6−N8−O9 | N8−C6−C7 | C6−C7−C4 |
|---|---|---|---|---|---|---|---|---|---|---|
| $A_0$(degree) | 180 | 180 | 120 | 120 | 99.4 | 122.1 | 115.7 | 122.2 | 99.6 | 112.9 |
| $K_{angle}$(kJ/mol rad$^2$) | 556 | 556 | 644 | 644 | 544 | 503 | 563 | 503 | 527 | 527 |
| angles | N8−C5−C10 | C10−C5−C11 | C6−C7−HC7 | C4−C7−HC7 | C7−C4−C5 | | | | | |
| $A_0$(degree) | 110.4 | 110.7 | 122.1 | 125.0 | 112.3 | | | | | |
| $K_{angle}$(kJ/mol rad$^2$) | 560 | 560 | 780 | 790 | 545 | | | | | |

***Table 3.*** Molecular Dynamics Characterization of Spin-Labeled RNAs

| slRNA | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $<d_{SL}>_{PELDOR}$ [Å] | 19.3 ± 1.2 | 33.7 ± 3.9 | 38.7 ± 1.3 | 21.9 ± 0.8 | 33.6 ± 1.6 | 26.9 ± 1.3 |
| $<d_{SL}>_{MD}$ [Å] | 17.2 ± 1.8 | 31.3 ± 1.8 | 36.4 ± 2.2 | 22.8 ± 0.7 | 34.3 ± 1.8 | 24.6 ± 2.4 |
| $<d_B>$ [Å] | 10.8 ± 0.6 | 28.1 ± 1.3 | 32.3 ± 1.7 | 12.4 ± 0.5 | 28.3 ± 0.9 | 23.4 ± 1.2 |
| correlation | 0.6 | 0.6 | 0.5 | 0.6 | 0.7 | 0.7 |
| rmsd [Å] | 2.8 | 1.9 | 2.6 | 2.6 | 2.6 | 3.4 |
| A-form [%] | 78.1 | 60.1 | 58.1 | 73.0 | 75.3 | 80.9 |
| populations[%] | 55/24/8 | 59/14/11 | 77/7/6 | 97/1/1 | 89/5/2 | 70/10/7 |

Listed are the internitroxide distances of the spin labels obtained from PELDOR experiments ($<d_{SL}>_{PELDOR}$)[14] and MD calculations ($<d_{SL}>_{MD}$) and the corresponding interbase distances $<d_B>$, the correlation coefficient between $<d_{SL}>_{MD}$) and $<d_B>$, the mean RMSD with respect to ideal A-form, the percentage of A-form conformation, the number of clusters, and the population probability of the two most populated clusters.

TIP3P water, and a 100 ps simulation of water molecules and counterions was performed with fixed solute, followed by a 100 ps NPT run without constraining the solute. Subsequently, the simulation was continued for 50 ns at constant temperature (300 K) and pressure (1 atm), whereby the coordinates were saved every 0.1 ps for analysis.

**4. Analysis of Trajectories.** Root mean squared deviations (rmsd) of all RNAs were calculated with respect to initially perfect A-form configuration. Internitroxide distances were measured as the distance between the nitrogen atoms of the two spin labels. Interbase distances were measured as the distance between carbon atoms of the base to which the spin labels were attached (C5 on purine, C2 on pyrimidine). The A-form of RNA is characterized by the quantity *Zp*, representing the mean *z*-coordinate of the backbone phosphorus atoms (with respect to individual dimer reference frames) that are greater than 1.5 Å for A-type and less than 0.5 Å for B-form steps.[41] Major and minor groove widths were defined from the distance between two phosphate atoms on the RNA backbone, with phosphate atom radius included.[41] Helical parameters as well as base pair conformations (classified as A, B, and TA-form) were calculated using the programs X3DNA[42] and CURVES 5.0.[43] Structural snapshots were selected every 5 ps for clustering. The clustering algorithm[44] was based on pairwise RMSDs with a cutoff of 2 Å for all atoms. The structural snapshot at the center of each cluster was taken as its representative structure.

## Results and Discussion

**1. Intramolecular Distances.** Distance calculations from previous MD simulations of RNA using an earlier version of the spin label force field were in good agreement with experimental PELDOR distances.[14] To get a first impression on the quality of the new force field, we therefore again compare calculated and experimental internitroxide distances for all spin-labeled RNAs. As shown in Table 3, we find good overall agreement of theory and experiment, with somewhat better results for slRNAs 2, 3, and 4 than obtained by the previously used model. Although the main reason for the reparameterization of the spin label force field was a better description of the spin label dynamics,[34] it is nevertheless reassuring that the new force field also yields a somewhat improved description of the overall structure.

For the interpretation of PELDOR experiments, it is interesting to know to what extent the internitroxide distances of the spin labels monitor the corresponding interbase distances in the RNA (see Methods for definitions). Figure 2 shows the time evolution of both quantities for all spin-labeled RNAs. As a consequence of the relative orientation of the two spin labels, the absolute values of internitroxide

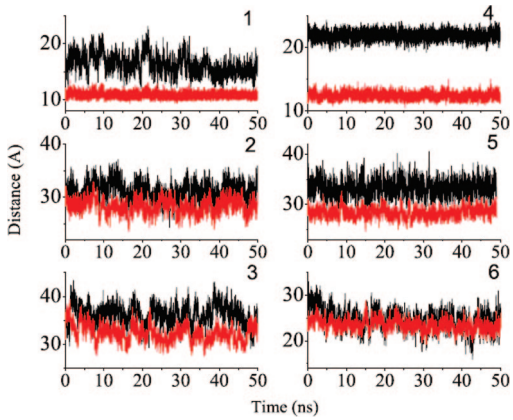



***Figure 2.*** Time evolution of the internitroxide (black) and interbase (red) distances of slRNA1−slRNA6.

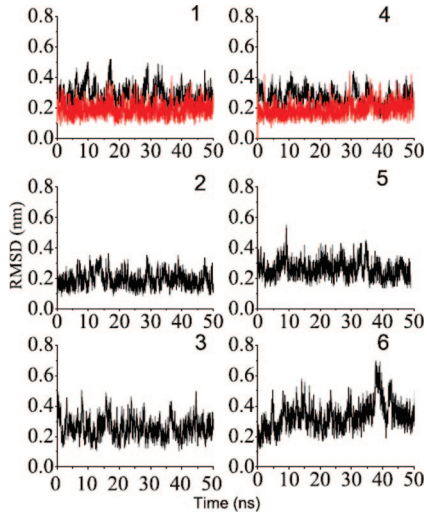



***Figure 3.*** Time evolution of the rmsd of spin-labeled slRNA1−slRNA6 (black) and unlabeled cRNA1 and cRNA4 (red).

and interbase distances may differ significantly. In particular, that is the case for slRNA4 which has the spin labels located in the minor groove.

For all systems, the patterns of the time traces are quite similar for internitroxide and interbase distances. As listed in Table 3, we typically find a correlation coefficient of ≈0.6. The similar pattern indicates that the motions of the spin labels are mainly due to the motion of RNA bases instead of the flexibility of the spin labels. Compared to the spin-labeling method using a flexible C—S linkage,[15,16] the rigidity of the spin labels in the present approach allows for a more direct observation of the base pair movement and therefore a more accurate and sensitive probing of RNA conformation.

**2. Structure of Spin-Labeled RNA.** A well-known measure for the overall stability of a MD simulation is the rmsd along a trajectory (see Methods for definition). Figure 3 shows the time evolution of the rmsd (with respect to ideal A-form) of all labeled and unlabeled RNAs considered. With the exception of slRNA1 and slRNA6, the RMSDs of the spin-labeled RNAs fluctuate steadily around 2−3 Å (see Table 3 for mean RMSDs). It is interesting to compare the RMSDs of slRNA1 (as an example for major-groove

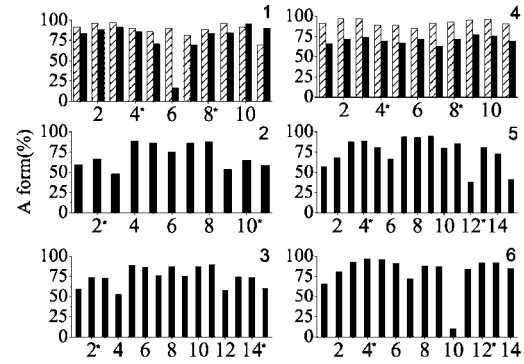



***Figure 4.*** Percentage of time in which base pair steps maintain A-form conformation. Shown are spin-labeled slRNA1−slRNA6 (black columns) and unlabeled cRNA1 and cRNA4 (shaded columns). The base pair step which contains the spin-labeled base is marked by an asterisk.

labeling) and slRNA4 (as an example for minor-groove labeling) to the RMSDs of the corresponding unlabeled cRNA1 and cRNA4. In both cases we find a significant increase of the RMSDs upon labeling. This finding is in agreement with the observation that spin-labeled RNAs exhibit lower melting temperatures than unlabeled RNAs.[14]

The significant structure in the rmsd time traces of slRNA1 and slRNA6 can be explained as conformational transitions of these systems. In fact, a clustering of the RNA structures of all trajectories reveals that only the minor-groove labeled slRNA4 remains essentially (>96%) in a single conformation, see Table 3. This is similar to the two unlabeled cRNA1 and cRNA4, which are found to occur in two clusters of 97% and 3% population, respectively. On the other hand, the major-groove labeled slRNAs show several thermally populated conformational states. This observation is in line with PELDOR measurements that show an increased damping of the PELDOR time traces for major-groove labeled slRNAs.[14]

As an alternative measure of the stability of RNA structure, we next consider the RNA base pair conformation, which can be roughly classified into A-form, B-form, and other structures (see Methods for definitions). To this end, Figure 4 shows the mean percentage of the A-form structure of all spin-labeled RNAs. Most of the base steps are found to maintain A-form conformation and intact base pairing for over 70% of time during simulation. Exceptions are again slRNA1 and slRNA6, which both show that one base step *between* the spin labels exhibits significantly low A-form content. Comparison with unlabeled cRNA1 and cRNA4 shows that spin labeling appears to somewhat reduce the A-form probability.

**3. Effects of Major-Groove Spin Labeling.** The results presented above have shown that the structural and dynamical effects of spin labeling in the major-groove are most prominent in slRNA1. In the following, we therefore adopt this system for further analysis. Although the spin labels of this RNA are attached at base steps 5 and 8, Figure 4 shows that mainly the structure of base step 6 deviates from A-form. This suggests that spin label in the major groove somehow deforms the conformation of the neighboring base pairs.

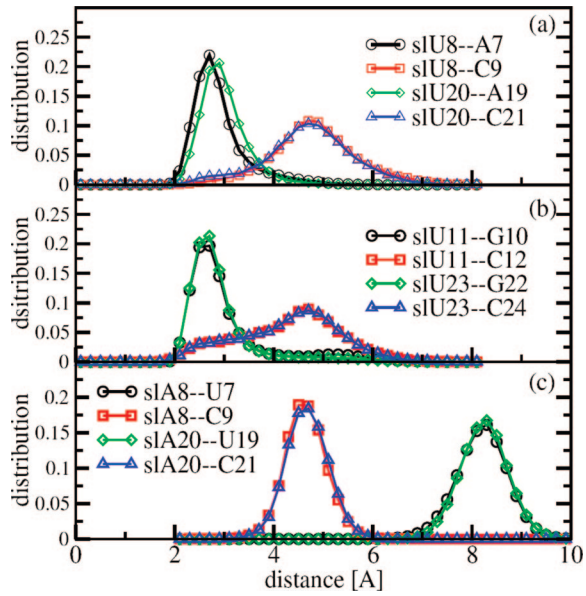



***Figure 5.*** MD distribution of minimum distances between the two spin-label groups and their neighbor base in slRNA1 (a), slRNA2 (b), and slRNA4 (c).

The steric influence of the spin label groups can be estimated by calculating the minimum distance between the atoms of the spin label (C4−C13 in Figure 1) and the neighboring base atoms of the same strand. As an example, Figure 5(a) displays the distribution of the minimum distances between the spin label group on U8 and neighbor bases A7 and C9 as well as between the spin label group on U20 and neighbor bases A19 and C21. In the case of A-$U_{SL}$, where the distance is measured between the base of adenine and the spin label of uracil, we obtain a mean distance of ≈2.4 Å in both cases. That is, for more than 73% of the simulation time the uracil spin labels are in close contact (≤2.5 Å) with the adenine bases. Such close contacts perturb the base configuration and may cause the opening of base pairs U6-A19 and A7-U18,[14] thus resulting in a low A-form content. The situation is similar in the case of G-$U_{SL}$ contacts found in slRNA2 (see Figure 5(b)) and slRNA3, which also show a mean distance of ≈2.4 Å in all cases considered. Further evidence of the effect of the spin label is gained by considering the width of the major groove. Comparing, e.g., spin-labeled slRNA1 and unlabeled cRNA1, Figure 6 shows that the label widens the major groove around the spin-labeled residues.

We note that base pairs U6-A19 and A7-U18 are the nearest neighbors to the spin-labeled bases in the 3′ direction. In contrast, the two base pairs G4-C21 and C9-G16 are the nearest neighbors in the 5′ direction. These base pairs exhibit a mean distance of ≈4.5 Å to the spin labels and are therefore rarely in close contact with the spin labels. In summary, due to the narrow and deep nature of the RNA major groove, we find close contacts (≤2.5 Å) of uracil spin labels with nearest neighboring bases in the 3′ direction, while we find sufficient space (≈4.5 Å) of uracil spin labels with nearest-neighbor bases in the 5′ direction. This finding is nicely illustrated in Figure 6, which shows representative MD snapshots of spin-labeled slRNA1 (a) and unlabeled cRNA1 (b).

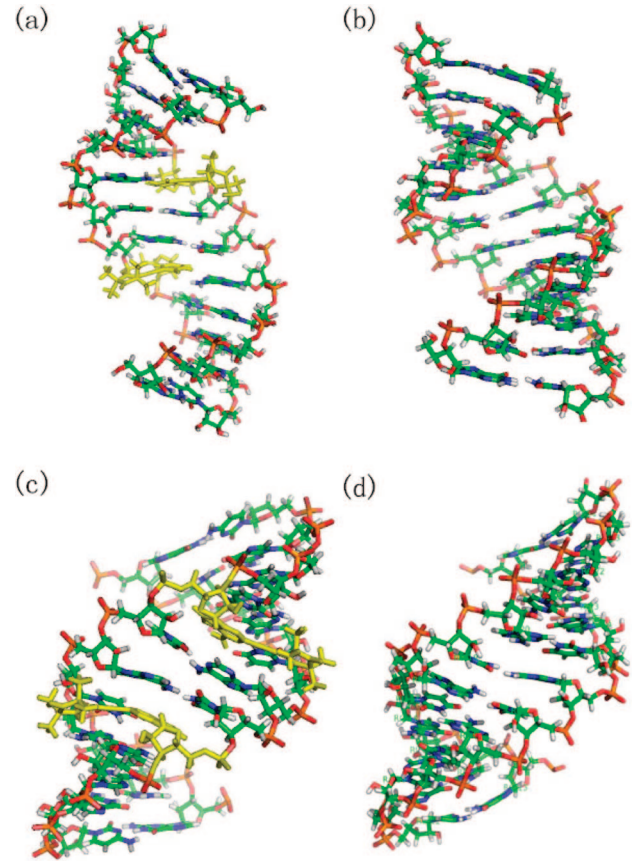



***Figure 6.*** Representative MD snapshots of spin-labeled slRNA1(a) and slRNA4(c) as well as of the corresponding unlabeled cRNA1 (b) and cRNA4 (d). Structures are viewed from a major groove in (a) and (b) and from a minor groove in (c) and (d). Spin-labeled residues are colored in yellow.

**4. Effect of Minor-Groove Spin Labeling.** The comparison of the computational results for minor groove spin-labeled slRNA4 with the results for the corresponding unlabeled cRNA4 reveals that the spin label hardly influences the structures of the RNA. Although the overall rmsd is somewhat increased (Figure 3) and the A-form content is somewhat decreased (Figure 4) upon labeling, there is no indication for local distortion of RNA as it is the case for major groove spin-labeled slRNA1 (Figure 7). Moreover, slRNA4 is found 97% of the time in a single conformational state (Table 3). These findings are in line with the distribution of the distances between the two spin label head groups and base pairs 6 and 7 (Figure 5(c)), which reveals that there is no close contact (≤2.5 Å) between spin labels and neighboring base pairs. Furthermore, Figure 6 shows that the groove widths of the labeled slRNA4 do not change evidently compared with the unlabeled cRNA4.

## Conclusions

We have developed a new parametrization of the force field for the nitroxide spin label, which was shown to lead to intramolecular distances that are in good agreement with experimental results. Employing this force field, we have studied in detail the effects of the labels on various RNA duplex structures. Nitroxide spin labels attached to pyrimidine bases such as uracil and cytosine at the C5 position are

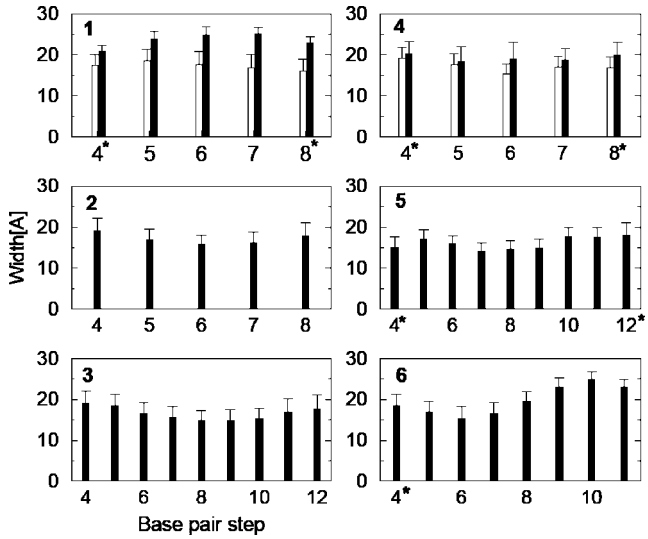



***Figure 7.*** Average major groove width of spin-labeled RNA1−6 (black columns) and unlabeled cRNA1 and cRNA4 (white columns). The base pair step which contains the spin-labeled base is marked by an asterisk.

located in the major groove, while labels attached to purine bases such as adenine at the C2 position are located in the minor groove. While there is ample space for the label in the minor groove, spin labeling in the major groove may lead to close contacts of the label and the neighbor bases, particularly in the 3′ direction. Such close contacts perturb the base configuration and may cause the opening of the neighbor base pairs, thus resulting in a low A-form content.

Interestingly, the situation is different for spin labeling of DNA duplexes,[35] where the label mostly affected the conformation of the labeled base pair rather than the adjacent base pairs. This is caused by the different structures of DNA and RNA duplexes: While RNA mainly adopts the B-form conformation, RNA is mainly found as an A-form helix which forms a deep and narrow major groove. Hence the A-form base pairs force the rigid spin label to incline in the 3′ direction, thus getting into close contact with base pairs that are one step down to the spin-labeled bases.

**Acknowledgment.** We thank T. F. Prisner, O. Schiemann, and J. W. Engels for numerous inspiring and helpful discussions and a long-standing collaboration. Furthermore, Y.M. gratefully acknowledges support from the University Research Committee (URC, RG65/06 grant) and the Ministry of Education AcRF Tier 2 grant (T206B3210RS), L.N. gratefully acknowledges support from the Ministry of Education AcRF Tier 2 grant (T206B3207), and G.S. gratefully acknowledges support from the Deutsche Forschungsgemeinschaft (via SFB 579 RNA-ligand interactions), the Fonds der Chemischen Industrie, and the Frankfurt Center for Scientific Computing.

## References

(1) Crick, F. H. C. *What Mad Pursuit*; Basic Books: New York, 1988.

(2) Judson, H. F. *The Eighth Day of Creation*; Simon & Schuster: New York, 1980.

(3) Denli, A. M.; Hannon, G. J. RNAi: an ever-growing puzzle. *Trends Biochem. Sci.* **2003**, *28*, 196.

(4) Delihas, N.; Rokita, S. E.; Zheng, P. Natural antisense RNA/target RNA interactions: Possible models for antisense oligonucleotide drug design. *Nat. Biotechnol.* **1997**, *15*, 751.

(5) Eaton, S. S. *Biomedical EPR*; Kluwer Academic/Plenum Publishers: New York, 2005.

(6) Vliegenthart, J. F. G. *NMR spectroscopy and computer modeling of carbohydrates: recent advances*; American Chemical Society: New York, 2006.

(7) Periasamy, A. *Molecular imaging: FRET microscopy and spectroscopy*; Published for the American Physiological Society by Oxford University Press: New York, 2005.

(8) Schiemann, O.; Fritscher, J.; Kisseleva, N.; Sigurdsson, S. T.; Prisner, T. F. Structural investigation of a high-affinity MnII binding site in the hammerhead ribozyme by EPR spectroscopy and DFT calculations. Effects of neomycin B on metal-ion binding. *Chembiochem* **2003**, *4*, 1057.

(9) Horton, T. E.; DeRose, V. J. Cobalt hexammine inhibition of the hammerhead ribozyme. *Biochemistry* **2000**, *39*, 11408.

(10) Qin, P. Z.; Dieckmann, T. Application of NMR and EPR methods to the study of RNA. *Curr. Opin. Struct. Biol.* **2004**, *14*, 350.

(11) Jeschke, G. Distance measurements in the nanometer range by pulse EPR. *Chemphyschem* **2002**, *3*, 927.

(12) Schiemann, O.; Piton, N.; Mu, Y.; Stock, G.; Engels, J. W.; Prisner, T. F. A PELDOR-based nanometer distance ruler for oligonucleotides. *J. Am. Chem. Soc.* **2004**, *126*, 5722.

(13) Werner, H. J.; Schulten, K.; Weller, A. Electron transfer and spin exchange contributing to the magnetic field dependence of the primary photochemical reaction of bacterial photosynthesis. *Biochim. Biophys. Acta* **1978**, *502*, 255.

(14) Piton, N.; Mu, Y.; Stock, G.; Prisner, T. F.; Schiemann, O.; Engels, J. W. Base-specific spin-labeling of RNA for structure determination. *Nucleic Acids Res.* **2007**, *35*, 3128.

(15) Cai, Q.; Kusnetzow, A. K.; Hideg, K.; Price, E. A.; Haworth, I. S.; Qin, P. Z. Nanometer distance measurements in RNA using site-directed spin labeling. *Biophys. J.* **2007**, *93*, 2110.

(16) Price, E. A.; Sutch, B. T.; Cai, Q.; Qin, P. Z.; Haworth, I. S. Computation of nitroxide-nitroxide distances in spin-labeled DNA duplexes. *Biopolymers* **2007**, *87*, 40.

(17) Cheatham, T. E., III. Simulation and modeling of nucleic acid structure, dynamics and interactions. *Curr. Opin. Struct. Biol.* **2004**, *14*, 360.

(18) Sanbonmatsu, K. Y.; Joseph, S.; Tung, C.-S. Simulating movement of tRNA into the ribosome during decoding. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15854.

(19) Auffinger, P.; Westhof, E. RNA hydration: three nanoseconds of multiple molecular dynamics simulations of the solvated tRNAAsp anticodon hairpin. *J. Mol. Biol.* **1997**, *269*, 326.

(20) Pan, Y.; Priyakumar, U. D.; MacKerell, A. D. Conformational Determinants of Tandem GU Mismatches in RNA: Insights from Molecular Dynamics Simulations and Quantum Mechanical Calculations. *Biochemistry* **2005**, *44*, 1433.

(21) Mu, Y. G.; Stock, G. Conformational dynamics of RNA-peptide binding; A molecular dynamics simulation study. *Biophys. J.* **2006**, *90*, 391.

(22) Koplin, J.; Mu, Y. G.; Richter, C.; Schwalbe, H.; Stock, G. Structure and dynamics of an RNA tetraloop: A joint

molecular dynamics and NMR study. *Structure* **2005**, *13*, 1255.

(23) Clerte, C.; Hall, K. B. Characterization of multimeric complexes formed by the human PTB1 protein on RNA. *RNA* **2006**, *12*, 457.

(24) Sorin, E. J.; Engelhardt, M. A.; Herschlag, D.; Pande, V. S. RNA simulations: probing hairpin unfolding and the dynamics of a GNRA tetraloop. *J. Mol. Biol.* **2002**, *317*, 493.

(25) Barthel, A.; Zacharias, M. Conformational transitions in RNA single uridine and adenosine bulge structures: A molecular dynamics free energy simulation study. *Biophys. J.* **2006**, *90*, 2450.

(26) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701.

(27) Cheatham, T. E., III.; Cieplak, P.; Kollman, P. A. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845.

(28) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **2007**, *92*, 3817.

(29) Beveridge, D. L.; Barreiro, G.; Byun, K. S.; Case, D. A.; Cheatham, T. E., III.; Dixit, S. B.; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Seibert, E.; Sklenar, H.; Stoll, G.; Thayer, K. M.; Varnai, P.; Young, M. A. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps. *Biophys. J.* **2004**, *87*, 3799.

(30) Varnai, P.; Zakrzewska, K. DNA and its counterions: a molecular dynamics study. *Nucleic Acids Res.* **2004**, *32*, 4269.

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

(32) Case, D. A.; Pearlman, D. A.; Caldwell, J. C.; Cheatham, T. E., III.; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Sibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *AMBER 6*; University of California: SF, 1999.

(33) Pigache, A.; Cieplak, P.; Dupradeau, F. Y. Automatic and highly reproducible RESP and ESP charge derivation: Application to the development of programs RED and X RED. In *227th ACS National Meeting,* Anaheim, CA, 2004.

(34) Marko, A.; Margraf, D.; Mu, Y.; Stock, G.; Prisner, T. ; Quantification of oreintation selection in PELDOR experiments. Manuscript in preparation for J. Chem. Phys. **2008**.

(35) Darian, E.; Gannett, P. M. Application of molecular dynamics simulations to spin-labeled oligonucleotides. *J. Biomol. Struct. Dyn.* **2005**, *22*, 579.

(36) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. K. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926.

(37) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159*, 98.

(38) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327.

(39) Tom, D.; Darrin, Y.; Lee, P. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089.

(40) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular-Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684.

(41) El Hassan, M. A.; Calladine, C. R. Two distinct modes of protein-induced bending in DNA. *J. Mol. Biol.* **1998**, *282*, 331.

(42) Lu, X.-J.; Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **2003**, *31*, 5108.

(43) Lavery, R.; Sklenar, H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.* **1988**, *6*, 63.

(44) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide folding: When simulation meets experiment. *Angew. Chem., Int. Ed.* **1999**, *38*, 236.

CT800266E

JCTC Journal of Chemical Theory and Computation

# Nature of Glycine and Its α-Carbon Radical in Aqueous Solution: A Theoretical Investigation

Geoffrey P. F. Wood,[†,‡,§] Mark S. Gordon,[||] Leo Radom,[†,§] and David M. Smith*[,⊥]

*School of Chemistry and ARC Centre of Excellence for Free Radical Chemistry and Biotechnology, University of Sydney, Sydney, New South Wales 2006, Australia, Department of Chemistry, Iowa State University, Ames, Iowa 50011, and Centre for Computational Solutions in the Life Sciences, Rudjer Boskovic Institute, 10002 Zagreb, Croatia*

Received July 23, 2008

**Abstract:** Quantum chemistry calculations and classical molecular dynamics simulations have been used to examine the equilibria in solution between the neutral and zwitterionic forms of glycine and also of the glycyl radical. The established preference (by 30 kJ mol$^{-1}$) for the zwitterion of glycine was confirmed by both the quantum chemical calculations and the classical molecular dynamics simulations. The best agreement with experiment was derived from thermodynamic integration calculations of explicitly solvated systems, which gives a free energy difference of 36.6 ± 0.6 kJ mol$^{-1}$. In contrast, for the glycyl radical in solution, the neutral form is preferred, with a calculated free energy difference of 54.8 ± 0.6 kJ mol$^{-1}$. A detailed analysis of the microsolvation environments of each species was carried out by evaluating radial distribution functions and hydrogen bonding patterns. This analysis provides evidence that the change in preference between glycine and glycyl radical is due to the inherent gas-phase stability of the neutral α-carbon radical rather than to any significant difference in the solvation behavior of the constituent species.

## 1. Introduction

Free radicals derived from α-amino acids are known to be important species in many biological processes. For example, the oxygen-centered tyrosyl radical is thought to be involved in photosynthesis as well as in the vital reduction of RNA to DNA.[1] Other peptide radicals have been implicated in a range of areas relevant to human health such as Alzheimer's disease, atherosclerosis, and diabetes as well as aging.[2−4]

A class of peptide radicals that arises frequently in biological systems are those derived from the homolytic cleavage of the $C_\alpha$-H bond. Constructive examples of this type may be found in the glycyl radical subclass of the radical-SAM superfamily of enzymes, which are important in various metabolic pathways of anaerobic bacteria.[5] Harmful examples are known to occur frequently as part of the degradation of proteins through fragmentation and rearrangement reactions initiated by reactive oxygen species.[2] Indeed, mechanisms of this latter type may well be intimately involved in the diseases noted above.

The prevalence of the $C_\alpha$-centered radicals derived from amino acids is thought to be associated with the unusually low $C_\alpha$-H bond dissociation energies (BDEs) of the relevant closed-shell parent species. This is normally attributed to the captodative stabilizing effect of having both electron-withdrawing and electron-donating substituents acting on a single radical center.[6] While this special stabilizing effect is relatively straightforward in peptide-based $C_\alpha$-radicals, it becomes more complicated in the amino acid building blocks themselves. The complication arises because, in the absence

* Corresponding author e-mail: david.smith@irb.hr.

† School of Chemistry, University of Sydney.

‡ Present address: Laboratory of Computational Chemistry and Biochemistry, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland.

§ ARC Centre of Excellence for Free Radical Chemistry and Biotechnology.

|| Iowa State University.

⊥ Rudjer Boskovic Institute.

10.1021/ct8002942 CCC: $40.75 
Published on Web 09/11/2008

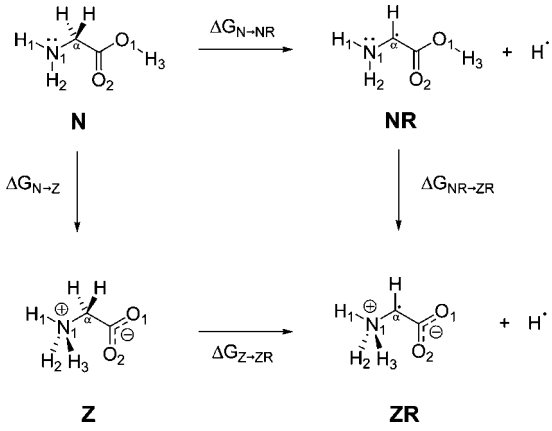



***Figure 1.*** Free energy cycle showing the conversion of neutral glycine (**N**) to zwitterionic glycine (**Z**) ($\Delta G_{N\rightarrow Z}$) and neutral glycyl radical (**NR**) to zwitterionic glycyl radical (**ZR**) ($\Delta G_{NR\rightarrow ZR}$). Alternatively, the cycle may be viewed as showing the $C_\alpha$-H bond dissociations of the closed-shell species **N** ($\Delta G_{N\rightarrow NR}$) and **Z** ($\Delta G_{Z\rightarrow ZR}$).

of surrounding peptide bonds, both the closed-shell parent and the $C_\alpha$-radical of an amino acid can potentially exist in either a neutral or a zwitterionic form. In the gas phase, the neutral form is preferred. However, in aqueous solution, the differences between the interaction of the neutral and zwitterionic forms with the solvent can be expected to have a large and possibly even dominating effect on the relative stabilities.

In order to investigate the effect of $C_\alpha$-radical generation on the neutral-zwitterion equilibrium in amino acids, we have chosen to characterize the simplest possible system in which it is present, namely glycine. Using a range of theoretical techniques, we have calculated the relative free energies, both in the gas phase (g) and in aqueous solution (aq), of the neutral and zwitterionic forms of glycine (**N** and **Z**) and the glycyl radical (**NR** and **ZR**). Figure 1 shows the four relevant species arranged in the form of a free energy cycle to facilitate both discussion and calculation.[7]

The branch of Figure 1 connecting neutral glycine with its zwitterionic counterpart ($\Delta G_{N\rightarrow Z}$) has received by far the most attention in the literature.[8,9] Ab initio studies find that the zwitterion of glycine in the gas phase ($\mathbf{Z}_{(g)}$) is not a local minimum on the potential energy surface but rather collapses to neutral glycine ($\mathbf{N}_{(g)}$).[9]

In contrast to the situation in the gas phase, both the neutral and zwitterionic forms of glycine are stable entities in aqueous solution ($\mathbf{N}_{(aq)}$ and $\mathbf{Z}_{(aq)}$), with an experimentally known energy difference in favor of the zwitterion of $\Delta G_{N\rightarrow Z(aq)} = -30$ kJ mol$^{-1}$.[8] Because of the fundamental nature of the problem, numerous groups have previously used theory to study the neutral-zwitterion equilibrium of glycine.[9] However, depending on the level of sophistication employed, the results can show significant variation. Some calculations indicate that the neutral form ($\mathbf{N}_{(aq)}$) is more stable by approximately 5 kJ mol$^{-1}$, while others favor the zwitterion ($\mathbf{Z}_{(aq)}$) by approximately 50 kJ mol$^{-1}$.[9]

Although no systematic comparison between the solution-phase energetics of the neutral and zwitterionic forms of the glycyl radical has been carried out, Barone and co-workers[10] have undertaken extensive work on their magnetic properties. Their general conclusion, in agreement with indirect experimental evidence,[11] is that the neutral form dominates in solution for pH values below 10. In particular, the zwitterion was discounted[10a,d] because the calculated hyperfine coupling-constant (*hfcc*) values did not agree with the solution-phase experiments but did show agreement with *hfcc* values derived from solid-state glycyl radical experiments,[12] in which the radical is known to exist in a zwitterionic form. On the other hand, the calculated *hfcc* values[10c,d,g] and g-tensors[10g] for the neutral form agree well with the observed spectroscopic features, provided a sufficiently sophisticated description of the solvent[10h] is used.

In the present study, we are particularly interested in understanding the details of why the preference for the zwitterionic form of glycine in aqueous solution[8] changes to a preference for the neutral form in the case of the glycyl radical. We approach this question from a thermodynamic point of view, which we believe is complementary to the magnetic approach that has been comprehensively applied to the glycyl-radical system in recent years.[10]

## 2. Theoretical Methodology

Gas-phase quantum-mechanical energies were obtained with the high-level CBS-QB3 procedure[13] using Gaussian 03.[14] As mentioned in the Introduction, the zwitterion of glycine in the gas phase ($\mathbf{Z}_{(g)}$) is not a local minimum on the potential energy surface but rather collapses to neutral glycine ($\mathbf{N}_{(g)}$).[9] In order to investigate the magnitude of $\Delta G_{N\rightarrow Z(g)}$, we have therefore chosen to use a $C_s$-symmetry-constrained geometry to approximate the zwitterionic structure ($\mathbf{Z}_{(g)}$). Implicit solvation calculations were performed using a polarizable continuum model (IEF-PCM)[15] with Bondi's all-atom radii and all other parameters appropriate for the solvent water. The geometric contribution to solvation was calculated by re-evaluating the gas-phase CBS-QB3 energies using B3-LYP/6−311G(d,p) geometry optimizations in conjunction with the IEF-PCM methodology. Finally, total free energies in solution and p$K_a$s were obtained by adding implicit solvation energies derived from IEF-PCM B3-LYP/cc-pVTZ//B3-LYP/6−31+G(d,p) single-point calculations to the re-evaluated CBS-QB3 energies.

Explicit solvation energies were obtained using classical molecular dynamics simulations. Classical valence and van der Waals parameters were assigned to each of the four solutes with the assistance of the antechamber[16] module of the AMBER 8[17] software package. Partial solute charges were obtained by restrained fitting to the electrostatic potentials (RESP)[18] derived from the IEF-PCM B3-LYP/cc-pVTZ//B3-LYP/6−31+G(d,p) calculations mentioned above.[19] Each of the four solutes was placed in a box of 793 TIP3P waters. Following energy minimization of the resultant system in order to remove close contacts, NPT molecular dynamics simulations were run using a 2 fs time step with a coupling constant of 0.2 ps to the constant target temperature of 300 K and a coupling constant of 1.0 ps to the constant target pressure of 1 bar. A 9.0 Å cutoff for nonbonded interactions was used in combination with the particle mesh Ewald procedure for long-range electrostatics,

while bond lengths were constrained using the SHAKE algorithm.[20] After an equilibration period of 10 ps, structural data were accumulated over 2 ns for the purpose of radial distribution function (RDF) and H-bond analysis. Following this, the energetic contribution of the solvent to each branch of Figure 1 was obtained by performing the four corresponding alchemical mutations, both in the forward and reverse directions (resulting in eight mutations in total).[21] The free energy differences associated with these mutations were evaluated using a thermodynamic integration protocol in which only the interactions between the solutes and the solvent were considered. The resulting solvent contributions were then combined with the CBS-QB3 energies to yield the final free energy differences in solution. While the idea of obtaining the solvent contribution to chemical equilibria from classical free energy calculations is not new,[22] it has been shown to be an accurate means to access this quantity that still finds widespread applicability in the modern era.[23]

In our calculations, the thermodynamic integration was performed using the Gibbs module from the Amber 6 program suite,[24] employing electrostatic decoupling. Simulations were run using a 1 fs time step with 20 discrete ($\lambda$) windows between each physical state. At each value of $\lambda$, 100 ps of equilibration was performed prior to 1 ns of data collection. Coupling constants of 1.0 ps to the target temperature and pressure were employed for these simulations. Further details can be found in the Supporting Information.

## 3. Results and Discussion

Our approach begins with a gas-phase treatment using the high-level CBS-QB3 procedure.[13] Application of this methodology yields the gas-phase free energies denoted as $\Delta G_{(g)}$. To supplement these results and arrive at the relevant free energies in aqueous solution ($\Delta G_{(aq)}$), we have used two alternative procedures. The first (implicit) approach involves the polarizable continuum model (PCM) of Tomasi and coworkers.[15] In this approach, the free energy of solvation of each species is calculated and added to the gas-phase free energy. In the second (explicit) approach, suitably parametrized models of the four compounds are placed in a box of 793 TIP3P water molecules and "alchemically" transformed according to the four branches of Figure 1.[21] The free energies associated with these transformations are then evaluated using thermodynamic integration.[25] Due to the fact that free energy is a state function, we may arrange the two differences of differences into an equality

$$\Delta\Delta G = \Delta G_{N\rightarrow NR} - \Delta G_{Z\rightarrow ZR} = \Delta G_{N\rightarrow Z} - \Delta G_{NR\rightarrow ZR} \quad (1)$$

which holds both in the gas phase and in aqueous solution and to which the free energy of the hydrogen atom does not contribute.[21] Traditionally such a cycle is employed to circumvent the calculation of more "difficult" free energy differences (such as $\Delta G_{N\rightarrow Z} - \Delta G_{NR\rightarrow ZR}$) through their substitution by more "straightforward" evaluations (like $\Delta G_{N\rightarrow NR} - \Delta G_{Z\rightarrow ZR}$). In the current work, however, we have explicitly calculated all four differences.

In the gas phase, the energy difference between the $C_s$-constrained zwitterionic form of glycine and its neutral isomer ($\Delta G_{N\rightarrow Z(g)}$) is 112 kJ mol$^{-1}$ (CBS-QB3, Table 1). On the other hand, using an explicit representation of the solvent yields a value for $\Delta G_{N\rightarrow Z(aq)}$ of $-36.6 \pm 0.6$ kJ mol$^{-1}$ (Table 1), in good agreement with the experimental result ($-30$ kJ mol$^{-1}$). The implicit approach to solvation also predicts $\mathbf{Z}_{(aq)}$to be more stable than $\mathbf{N}_{(aq)}$ (in this case by 50.2 kJ mol$^{-1}$), but the comparison with experiment is less satisfactory.

***Table 1.*** Free Energy Differences Relevant to Figure 1 (298 K, kJ mol$^{-1}$)

| free energy | CBS-QB3$_{(g)}$ | CBS-QB3$_{(aq)}$ implicit[a] | CBS-QB3$_{(aq)}$ explicit[b] |
|---|---|---|---|
| $\Delta G_{N\rightarrow Z}$ | 112.0 | −50.2 | $-36.6 \pm 0.6$[c] |
| $\Delta G_{NR\rightarrow ZR}$ | 200.8 | 42.3 | $54.8 \pm 0.6$[c] |
| $\Delta G_{N\rightarrow NR}$ | 303.9 | 297.4 | $300.8 \pm 0.2$[c] |
| $\Delta G_{Z\rightarrow ZR}$ | 392.8 | 389.9 | $400.5 \pm 0.1$[c] |
| $\Delta G_{(N\rightarrow Z\ -\ NR\rightarrow ZR)}$ | −88.8 | −92.5 | $-91.4 \pm 1.2$[d] |
| $\Delta G_{(N\rightarrow NR\ -\ Z\rightarrow ZR)}$ | −88.8 | −92.5 | $-99.7 \pm 0.3$[d] |

[a] Solvent effects calculated using the PCM model. [b] Solvent effects calculated using a box of 793 TIP3P water molecules. [c] The tabulated figure represents an average of the results from the simulations run in the forward and reverse directions. The uncertainty reflects half of the difference between these results. [d] The uncertainty is the sum of uncertainties in each branch contributing to the difference.

While it is not the aim of this study to simply reproduce the equilibrium behavior between the neutral ($\mathbf{N}_{(aq)}$) and zwitterionic ($\mathbf{Z}_{(aq)}$) forms of glycine, the good agreement with experiment obtained by combining CBS-QB3 gas-phase energies with an explicit classical representation of the solvent, for this equilibrium, is important and encouraging. In particular, this result can be considered as a calibration of the approach, indicating that the chosen methodology can be considered to be reasonably reliable for the closed-shell equilibrium (between $\mathbf{N}_{(aq)}$ and $\mathbf{Z}_{(aq)}$) and by implication can be expected to be similarly reliable for treating the closely related equilibrium between the neutral and zwitterionic forms of glycyl radical ($\mathbf{NR}_{(aq)}$ and $\mathbf{ZR}_{(aq)}$). Indeed, an approach involving calibration of a methodology with a known result, followed by informed application to a closely related model, has been advocated for some time.[26]

The CBS-QB3 results predict that the neutral form of the glycyl radical $\mathbf{NR}_{(g)}$ is significantly more stable (in the gas phase) than the (constrained) zwitterionic form $\mathbf{ZR}_{(g)}$, which can be seen by the value of $\Delta G_{NR\rightarrow ZR(g)} = 200.8$ kJ mol$^{-1}$ in Table 1. This large energy difference comes about because, in addition to the contribution arising from charge separation (such as that which dominates $\Delta G_{N\rightarrow Z(g)}$), there is captodative stabilization in the neutral form of the radical ($\mathbf{NR}_{(g)}$) that is absent in the zwitterionic counterpart ($\mathbf{ZR}_{(g)}$). However, due to the anticipated preferential solvation of the zwitterion ($\mathbf{ZR}_{(g)}$), the extent of this difference might be expected to be considerably reduced in aqueous solution, as in the case of glycine itself (**N** vs **Z**). Indeed, the value of $\Delta G_{NR\rightarrow ZR(aq)}$ calculated with explicit solvent indicates that the impact of aqueous solvation on this difference is some 146.0 kJ mol$^{-1}$. However, even this large differential solvation effect is not sufficient to overcome the inherent preference (by 200.8 kJ mol$^{-1}$) for the neutral radical ($\mathbf{N}_{(g)}$), which is predicted to remain the more stable radical species in solution by 54.8

$\pm$ 0.6 kJ mol$^{-1}$ (using the explicit model). The implicit solvation model also supports this conclusion but, as was the case for $\Delta G_{N\rightarrow Z(aq)}$, this approach probably overstabilizes the zwitterionic form, leading to a prediction of $\Delta G_{NR\rightarrow ZR(aq)} = 42.3$ kJ mol$^{-1}$.

Intrigued by the earlier suggestion that a protonated form of the glycyl radical ($NH_2C(\bullet)HCO_2H_2\oplus$) could be necessary to account for the observed magnetic properties in acidic solution,[10b] we have also calculated the p$K_a$ values of the two relevant radical-cation species using the implicit approach outlined above. We find that the calculated p$K_a$ of the species ($NH_2C(\bullet)HCO_2H_2\oplus$) that would result from protonation of the oxygen atom of the neutral glycyl radical is $-5.4$. Similarly, the p$K_a$ of the less stable species, resulting from protonation of the nitrogen ($\oplus NH_3C(\bullet)HCO_2H$), is calculated to be $-4.1$. Both of these values are low, and the true values are likely to be even lower, given that the same methodology overestimates the experimental p$K_a$ value for the $\oplus NH_3CH_2COOH \rightarrow \oplus NH_3CH_2COO\ominus + H\oplus$ reaction by 0.5 p$K_a$ units (2.8 as opposed to 2.3, see Table S3 of the Supporting Information).[27] We therefore conclude that it is unlikely for the glycyl radical to become protonated, even under strongly acidic conditions. Such a conclusion is compatible with the spectral parameters derived from vibrational-averaging[10c,d] and molecular dynamics simulations[10g] presented by Barone and co-workers, which are in turn in good agreement with those measured at a pH of 1.[11]

Clearly, the presence of a radical center at $C_\alpha$ of glycine drastically alters the equilibrium between, and the acidity of, the respective neutral and zwitterionic forms. The extent to which the presence of the radical alters the equilibrium, when compared with the same situation in the closed-shell counterparts, is provided by the quantity $\Delta\Delta G = \Delta G_{N\rightarrow Z} - \Delta G_{NR\rightarrow ZR}$. As mentioned previously, this is equivalent to the difference in the two bond dissociation energies $\Delta\Delta G = \Delta G_{N\rightarrow NR} - \Delta G_{Z\rightarrow ZR}$, which in turn can be called a radical stabilization energy.[28] Regardless of which branches of the thermodynamic cycle are used, the gas-phase results in Table 1 show that the neutral form of the radical is favored over the zwitterionic form in the gas phase by an additional 88.8 kJ mol$^{-1}$ when compared with the same situation for the closed-shell parent species.

Interestingly, despite the potential for large differences in the solvation energies of the various species in Figure 1, the value of $\Delta\Delta G_{(aq)}$ is not significantly different from $\Delta\Delta G_{(g)}$. In the case of the implicit model, the effect of solvation is predicted to cause $\Delta\Delta G$ to become more negative than the gas-phase value by just 3.7 kJ mol$^{-1}$. The explicit solvation model also suggests that $\Delta\Delta G_{(aq)}$ is only marginally more negative than $\Delta\Delta G_{(g)}$. In this case, however, the effect is expressed as a range (2.6$-$10.7 kJ mol$^{-1}$) rather than as a single value.[29] In both cases, the minor difference between $\Delta\Delta G_{(g)}$ and $\Delta\Delta G_{(aq)}$ indicates that the reason why there is a qualitative shift accompanying solvation in the equilibrium between the neutral and zwitterionic forms of glycine, but not for the glycyl radical, is largely associated with the underlying gas-phase stabilities rather than any drastically different solvation behavior.

***Table 2.*** Charge Values (e) Obtained by the RESP Procedure[a] for Neutral Glycine (**N**), the Neutral Glycyl Radical (**NR**), Zwitterionic Glycine (**Z**), and the Zwitterionic Glycyl Radical (**ZR**)

| atom | N | NR | Z | ZR |
|---|---|---|---|---|
| $N_1$ | $-1.071$ | $-0.644$ | $-0.073$ | $-0.136$ |
| $H_1$ | 0.398 | 0.407 | 0.276 | 0.321 |
| $H_2$ | 0.398 | 0.407 | 0.276 | 0.321 |
| $H_3$ | 0.520 | 0.494 | 0.276 | 0.321 |
| $C_\alpha$ | 0.328 | $-0.117$ | 0.001 | $-0.101$ |
| $H_\alpha$ | 0.040 | 0.199 | 0.069 | 0.152 |
| $H_{\alpha'}$ | 0.040 | | 0.069 | |
| C | 0.736 | 0.632 | 0.758 | 0.732 |
| $O_1$ | $-0.720$ | $-0.690$ | $-0.826$ | $-0.804$ |
| $O_2$ | $-0.670$ | $-0.688$ | $-0.826$ | $-0.804$ |

[a] Using electrostatic potentials derived from the IEF-PCM B3-LYP/cc-pVTZ//B3-LYP/6$-$31+G(d,p) calculations.

The minimal net impact of solvation on $\Delta\Delta G$ shown in Table 1 could be taken to imply that the relative solvation environments of the closed-shell species are quite similar to those of the relevant radical counterparts. In addition to this circumstantial thermodynamic argument, it is possible to probe such phenomena more directly through a detailed structural analysis of the simulations carried out in the present study using explicit water.

Prior to embarking on such a structural analysis however, it is informative to briefly examine the RESP charges obtained for each of the four solutes examined in this study (Table 2). It is through these charge values that the differing electronic distributions, reflected in the differing electrostatic potentials, enter the classical molecular dynamics simulations. Several important factors can be seen by inspection of Table 2. First, the atomic charges for the two zwitterionic species (**Z** and **ZR**) are relatively similar to one another, as demonstrated by the largest difference between them (for $C_\alpha$) of just 0.1 e. On the other hand, there are more significant differences between the two neutral species (**N** and **NR**). In particular, the charge on the nitrogen is 0.4 e less negative in the radical than in the closed-shell species. This is compensated for by the charge on $C_\alpha$ becoming negative (partially offset by a more positive charge for $H_\alpha$) and a more negative total charge on the carboxylic acid substituent (from $-0.1$ to $-0.3$ e). These results are consistent with the concept of the captodative effect, which sees the radical center receive an increased donation of density from the amino substituent, combined with an increased acceptance by the carboxylic acid substituent.

One means of obtaining an informative overview of the microsolvation of the various species examined in this study is through the inspection of selected radial distribution functions (RDFs). For example, Figure 2 shows the RDFs of water oxygens ($O_w$) around $N_1$ (solid lines) and $O_2$ (dashed lines) of all four species (see Figure 1). The major difference between the neutral (**N** and **NR**) and zwitterionic (**Z** and **ZR**) RDFs is the significantly larger peak heights associated with the latter. This is simply a reflection of the stronger interaction of the more polar species with the aqueous medium, as is also quantitatively evident from the free energy changes shown in Table 1. In accordance with the thermodynamic expectations, the RDFs for the two zwitterionic

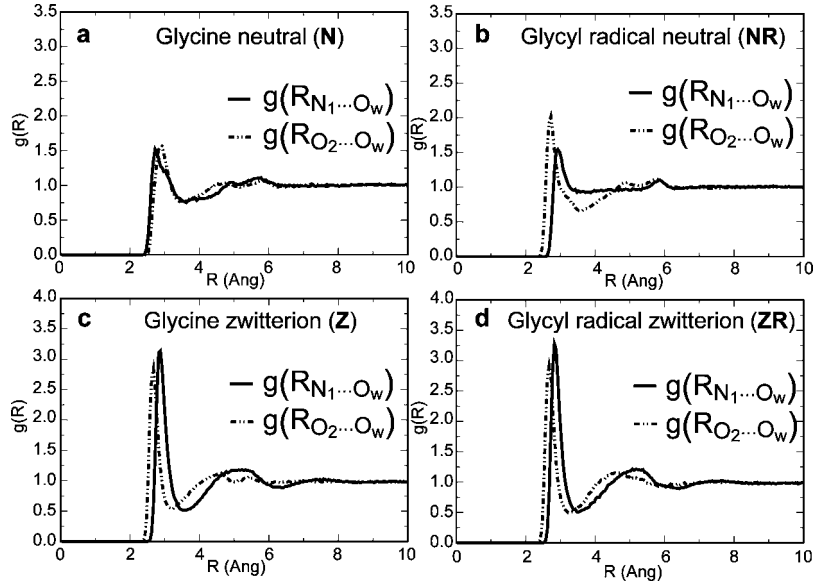



***Figure 2.*** Radial distribution functions (g(R)) of water oxygens ($O_w$) around (**a**) $N_1$ (solid lines) and $O_2$ (dashed lines) of neutral glycine (**N**), (**b**) $N_1$ and $O_2$ of neutral glycyl radical (**NR**), (**c**) $N_1$ and $O_2$ of glycine zwitterion (**Z**), and (**d**) $N_1$ and $O_2$ of glycyl radical zwitterion (**ZR**).

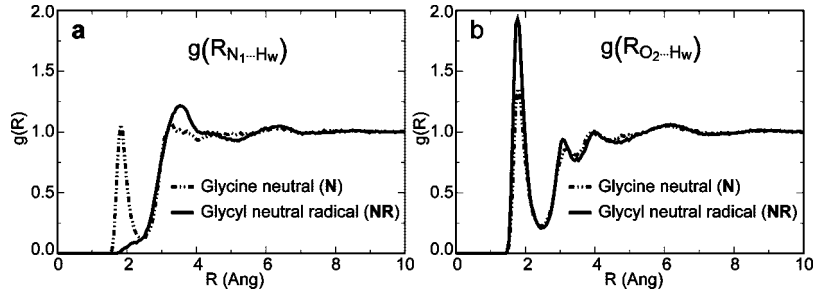



***Figure 3.*** Radial distribution functions of water protons ($H_w$) around (**a**) $N_1$ of neutral glycine (**N**, dashed line) and of neutral glycyl radical (**NR**, solid line) and (**b**) $O_2$ of neutral glycine (**N**, dashed line) and of neutral glycyl radical (**NR**, solid line).

systems **Z** and **ZR** can be seen to be very similar to one another (Figure 2c,d). A comparable observation is qualitatively valid for the two neutral systems (**N** and **NR**, Figure 2a,b), although some minor differences are apparent. In particular, the N-atom of the radical (**NR**) appears to be less well solvated than the analogous N-atom in the closed-shell system (**N**). The opposite trend is apparent for the corresponding carbonyl oxygen atom.

Additional information pertaining to the nature of the microsolvation of the $N_1$ and $O_2$ atoms of neutral glycine (**N**) and its $C_\alpha$-derived radical (**NR**) is provided by the RDFs of the water protons ($H_w$) surrounding these atoms. For example, the RDFs of solvent protons surrounding the nitrogen atom shown in Figure 3a are markedly different for species **N** and **NR**. In particular, the absence of the peak centered at ~2 Å in the solid curve gives a strong indication that the lone pair on the N-atom acts as a substantially weaker hydrogen-bond acceptor in the $C_\alpha$-radical (**NR**) than in neutral glycine (**N**) itself. Such behavior can be rationalized in terms of the enhanced (captodative) delocalization in the glycyl radical, for example by making the nitrogen lone pair less accessible for H-bonding. Insofar as hydrogen-bonding tendencies are related to proton affinity,[30] this result can also be connected to the reduced basicity at the nitrogen atom in the glycyl radical. The loss of H-bond-accepting ability at this atom appears to be partially compensated for by a concomitant increase in the H-bond-accepting ability at the carbonyl oxygen of the same species (**NR**). This is manifested in the enhanced peak height associated with the solid curve in Figure 3b.

***Table 3.*** Average Number of Hydrogen Bonds for the Duration of the Simulations between Water and Various Sites on Neutral Glycine (**N**), the Neutral Glycyl Radical (**NR**), Zwitterionic Glycine (**Z**), and the Zwitterionic Glycyl Radical (**ZR**)

| site | $N_{(aq)}$ | $NR_{(aq)}$ | site | $Z_{(aq)}$ | $ZR_{(aq)}$ |
|---|---|---|---|---|---|
| $N_1-H_1$ | 0.9 | 1.0 | $N_1-H_1$ | 1.1 | 1.1 |
| $N_1-H_2$ | 0.9 | 1.0 | $N_1-H_2$ | 1.1 | 1.1 |
| $N_1$ | 1.4 | 0.3 | $N_1-H_3$ | 1.1 | 1.1 |
| $O_1-H_3$ | 1.0 | 1.0 | $O_1$ | 3.5 | 3.3 |
| $O_2$ | 2.0 | 2.5 | $O_2$ | 3.5 | 3.7 |

In addition to the RDF analysis presented above, we have also probed the microsolvation of all four species by monitoring H-bonds throughout the relevant trajectories, where an H-bond X-H•••Y is defined to exist if the X•••Y length is less than 3.5 Å and the angle defined by the three centers comprising the bond is between 120.0° and 180.0°. The results of this analysis, which are shown in Table 3, serve to provide quantitative confirmation of the conclusions

suggested by the RDFs. Specifically, the effect of H-atom loss on the solvation of glycine zwitterion is rather minimal (recall the similarity between parts c and d of Figure 2). The average H-bonding behavior of the $NH_3\oplus$ group can be seen to be virtually identical for the closed-shell (**Z**) and open-shell (**ZR**) species. The carboxylate oxygens exhibit a minor loss of mutual equivalence in the radical (**ZR**), but the average H-bond-accepting capacity of the $CO_2^-$ group appears to be unaffected by $C_\alpha$-radical formation.

The H-bond-donating abilities of the $N_1-H$ and $O_1-H_3$ groups of the neutral systems also remain virtually unaltered in response to radical generation. On the other hand, the degree of H-bond acceptance by both $N_1$ and $O_2$ does seem to differ significantly between the closed- and open-shell neutral systems. In quantitative support of the graphical interpretation presented in Figure 3, the average number of H-bonds accepted by $N_1$ in the neutral radical (**NR**) (0.3) is reduced by 1.1 (from 1.4) with respect to the closed-shell parent (**N**). At the same time, the H-bond-accepting capacity of $O_2$ (2.5) is enhanced in the radical (**NR**) by 25% compared with that observed (2.0) for the closed-shell species (**N**). Again, the result is readily rationalized in terms of the captodative delocalization in the glycyl radical.

## 4. Concluding Remarks

In summary, we have used a variety of theoretical means to investigate the comparative equilibria between the neutral and zwitterionic forms of glycine and its $C_\alpha$-radical. Our calculations show that an explicit classical representation of the solvent is able to satisfactorily reproduce the known magnitude of the free energy preference for zwitterionic glycine in aqueous solution. An analogous application of the same methodology to the glycyl radical shows that, in contrast to the closed-shell parent system, the neutral form is preferred in solution by approximately 55 kJ $mol^{-1}$. Examination of the components of this difference reveals that this preference can be almost entirely attributed to the captodative stabilization of the neutral radical in the gas-phase reference state. In other words, even though the zwitterionic radical is significantly better solvated than the neutral radical, the extent of this interaction is not sufficient to overcome the underlying preferential gas-phase stabilization in the neutral glycyl radical. In a related finding, our calculations indicate that it is very unlikely for the neutral glycyl radical to become protonated, even at very low pH values.

A convenient quantitative measure of the relevant differential stabilization is provided by the quantity denoted in the present study as $\Delta\Delta G$, the difference between the zwitterionic and neutral energies for glycine on the one hand and the glycyl radical on the other. Our calculations predict that $\Delta\Delta G$ adopts a value in the gas phase of 88.8 kJ $mol^{-1}$. Despite the potential for solvation to have a substantial effect on this quantity, our best prediction for the value of $\Delta\Delta G_{(aq)}$ lies between 90 and 100 kJ $mol^{-1}$, indicating that the impact of the aqueous medium in this case is, in actual fact, quite minor. Analysis of the microsolvation patterns of the four species investigated in the present study supports this conclusion. While large differences are found when collectively comparing the neutral (**N** and **NR**) with the zwitterionic (**Z** and **ZR**) systems, each open-shell and closed-shell pair is found to exhibit relatively similar general solvation patterns. An important difference arises for the neutral pair (**N** and **NR**), for which the N-atom of the $C_\alpha$ radical (**NR**) is found to exhibit a markedly reduced propensity for H-bond acceptance (compared with **N**), whereas the carbonyl oxygen of the radical experiences an apparently compensatory effect. This phenomenon again appears to be a consequence of the simultaneous and synergistic action of $\pi$-electron-donating ($NH_2$) and $\pi$-electron-accepting ($CO_2H$) substituents adjacent to a radical center.

**Acknowledgment.** We gratefully acknowledge the award (to L.R.) of an Australian Research Council Discovery grant, funding (to L.R and G.P.F.W.) through the ARC Centre of Excellence for Free Radical Chemistry and Biotechnology, and generous allocations of computer time from the Australian Partnership for Advanced Computing (APAC) and the Australian Centre for Advanced Computing and Communications (AC3). We also acknowledge support (for M.S.G) from the U.S. Air Force Office of Scientific Research and the Australia Fulbright Association. Finally, the support (of D.M.S) by the Croatian Ministry of Science (project 098-0982933-2937) and the E.C. (FP6 contract 043749) is gratefully acknowledged.

**Supporting Information Available:** Gaussian archive entries (Table S1), extracts from Amber prep files (Table S2), details of p$K_a$ evaluations (Table S3), and details of the convergence of the RDFs and thermodynamic integration calculations (Table S4). This material is available free of charge via the Internet at http://pubs.acs.org.

### References

(1) Stubbe, J.; van der Donk, W. *Chem. Rev.* **1998**, *98*, 705, and references therein.

(2) Davies, M. J.; Dean, R. T. *Radical-Mediated Protein Oxidation: from Chemistry to Medicine*; Oxford University Press: Oxford; New York, 1997; pp 203−237, and references therein.

(3) Dean, R. T.; Fu, S.; Stocker, R.; Davies, M. J. *Biochem. J.* **1997**, *324*, 1.

(4) (a) Rauk, A.; Armstrong, D. A.; Fairlie, D. P. *J. Am. Chem. Soc.* **2000**, *122*, 9761. (b) Rauk, A. *Can. Chem. News* **2001**, *53*, 20. (c) Brunelle, P.; Rauk, A. *J. Alzheimer's Dis.* **2002**, *4*, 283.

(5) (a) Frey, P. A.; Magnusson, O. T. *Chem. Rev.* **2003**, *103*, 2129. (b) Wang, S. C.; Frey, P. A. *Trends Biochem. Sci.* **2007**, *32*, 101.

(6) (a) Viehe, H. G.; Janouesk, Z.; Merenyi, R.; Stella, L. *Acc. Chem. Res.* **1985**, *18*, 148. (b) Easton, C. J. *Chem. Rev.* **1997**, *97*, 53. (c) Rauk, A.; Yu, D.; Taylor, J.; Shustov, G. V.; Block, D. A.; Armstrong, D. A. *Biochemistry* **1999**, *38*, 9089. (d) Wood, G. P. F.; Moran, D.; Jacob, R.; Radom, L. *J. Phys. Chem. A* **2005**, *109*, 6318.

(7) Throughout this paper, the quantity $\Delta G_{m\rightarrow n}$ refers to the free energy difference $G_n - G_m$.

(8) Wada, G.; Tamura, E.; Okina, M.; Nakamura, M. *Bull. Chem. Soc. Jpn.* **1982**, *55*, 3064.

(9) See for example: (a) Gaffney, J. S.; Pierce, R. C.; Friedman, L. *J. Am. Chem. Soc.* **1977**, *99*, 4293. (b) Clementi, E.; Cavallone, F.; Scordamaglia, R. *J. Am. Chem. Soc.* **1977**, *99*, 5531. (c) Tse, Y. C.; Newton, M. D.; Vishveshwara, S.; Pople, J. A. *J. Am. Chem. Soc.* **1978**, *100*, 4329. (d) Jensen, J. H.; Gordon, M. S. *J. Am. Chem. Soc.* **1995**, *117*, 8159. (e) Nagaoka, M.; Okuyama-Yoshida, N.; Yamabe, T. *J. Phys. Chem. A* **1998**, *102*, 8202. (f) Bandyopadhyay, P.; Gordon, M. S. *J. Chem. Phys.* **2000**, *113*, 1104. (g) Shoeib, T.; Ruggiero, G. D.; Siu, M. K. W.; Hopkinson, A. C.; Williams, I. H. *J. Chem. Phys.* **2002**, *117*, 2762. (h) Leung, K.; Rempe, S. B. *J. Chem. Phys.* **2005**, *122*, 184506. (i) Aikens, C. M.; Gordon, M. S. *J. Am. Chem. Soc.* **2006**, *128*, 12835. (j) Chang, J.; Lenhoff, A. M.; Sandler, S. I. *J. Phys. Chem. B* **2007**, *111*, 2098. (k) Bachrach, S. M. *J. Phys. Chem. A* **2008**, *112*, 3722.

(10) (a) Barone, V.; Adamo, C.; Grand, A.; Subra, R. *Chem. Phys. Lett.* **1995**, *242*, 351. (b) Barone, V.; Adamo, C.; Grand, A.; Jolibois, F.; Brunel, Y.; Subra, R. *J. Am. Chem. Soc.* **1995**, *117*, 12618. (c) Rega, N.; Cossi, M.; Barone, V. *J. Am. Chem. Soc.* **1997**, *119*, 12962. (d) Rega, N.; Cossi, M.; Barone, V. *J. Am. Chem. Soc.* **1998**, *120*, 5723. (e) Ciofini, I.; Adamo, C.; Barone, V. *J. Chem. Phys.* **2004**, *121*, 6710. (f) Brancato, G.; Barone, V.; Rega, N. *Theor. Chem. Acc.* **2007**, *117*, 1001. (g) Brancato, G.; Rega, N.; Barone, V. *J. Am. Chem. Soc.* **2007**, *129*, 15380. (h) Brancato, G.; Rega, N.; Barone, V. *J. Chem. Phys.* **2008**, *128*, 144501.

(11) (a) Paul, H.; Fischer, H. *Helv. Chim. Acta* **1971**, *54*, 485. (b) Neta, P.; Fessenden, R. W. *J. Phys. Chem.* **1971**, *75*, 738.

(12) Nunone, K.; Muto, H.; Torivama, K.; Iwasaki, M. *J. Chem. Phys.* **1976**, *65*, 3805.

(13) (a) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1999**, *110*, 2822. (b) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **2000**, *112*, 6532.

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

(15) (a) Munnucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *106*, 5151. (b) Cancès, M. T.; Munnucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032. (c) Cossi, M.; Munnucci, B.; Tomasi, J. *Chem. Phys. Lett.* **1998**, *286*, 253.

(16) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, *25*, 247–260.

(17) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 8*; University of California: San Francisco, CA, 2004.

(18) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269.

(19) This procedure is similar to that employed in the following: Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J. W.; Wang, J.; Kollman, P. A. *J. Comput. Chem.* **2003**, *24*, 1999.

(20) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 2272.

(21) The simulations of the explicit solvation contributions to $\Delta G_{N\rightarrow NR(aq)}$ and $\Delta G_{Z\rightarrow ZR(aq)}$ involved the direct mutation of **N** (or **Z**) to **NR** (or **ZR**), and no attempt was made to simulate the bare hydrogen atom in a box of water. For the purpose of reporting BDEs, the solvation free energy used for the H-atom (5.4 kJ $mol^{-1}$) was taken from an implicit PCM calculation. It is important to note that this quantity cancels entirely from $\Delta\Delta G_{(N\rightarrow NR-Z\rightarrow ZR)}$ (Table 1) as well as from $\Delta G_{NR\rightarrow ZR(aq)}$.

(22) Jorgensen, W. L.; Ravimohan, C. *J. Chem. Phys.* **1985**, *83*, 3050.

(23) See, for example Jorgensen, W. L.; Thomas, L. L. *J. Chem. Theory Comput.* **2008**, *4*, 869.

(24) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Radmer, R. J.; Duan, J.; Pitera, J.; Massova, I. G.; Seibel, L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *AMBER 6*; University of California: San Francisco, CA, 1999.

(25) Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1991**, *95*, 1175.

(26) (a) Pople, J. A. In *Theoretical Models for Chemistry, Proceedings of the Summer Research Conference on Theoretical Chemistry, Energy Structure and Reactivity*; Smith, D. W., Ed.; John Wiley & Sons: New York, 1973; pp 51−61. (b) See also Pople, J. A. *J. Chem. Phys.* **1965**, *48*, 229.

(27) See for example: Nelson, D. L.; Cox, M. M. *Lehninger Principles of Biochemistry*, 5th ed.; W. H. Freeman: New York, 2008; p 73.

(28) (a) See for example Henry, D. J.; Parkinson, C. J.; Mayer, P. M.; Radom, L. *J. Phys. Chem. A* **2001**, *105*, 6750, and references therein. (b) For a recent comprehensive review on radical stability, see Zipse, H. *Top. Curr. Chem.* **2006**, *263*, 163.

(29) The range in values comes about because of inadequacies in the numerical integration scheme.

(30) See for example Chan, B.; Del Bene, J. E.; Elguero, J.; Radom, L. *J. Phys. Chem. A* **2005**, *109*, 550.

CT8002942

JCTC Journal of Chemical Theory and Computation

# Molecular Polarization Effects on the Relative Energies of the Real and Putative Crystal Structures of Valine

Timothy G. Cooper, Katarzyna E. Hejczyk, William Jones, and Graeme M. Day*

*The Pfizer Institute for Pharmaceutical Materials Science, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom*

Received May 28, 2008

**Abstract:** The computer-generation of the crystal structures of the α-amino acid valine is used as a challenging test of lattice energy modeling methods for crystal structure prediction of flexible polar organic molecules and, specifically, to examine the importance of molecular polarization on calculated relative energies. Total calculated crystal energies, which combine atom-atom model potential calculations of intermolecular interactions with density functional theory intramolecular energies, do not effectively distinguish the real (known) crystal structures from the rest of the low energy computer-generated alternatives when the molecular electrostatic models are derived from isolated molecule calculations. However, we find that introducing a simple model for the bulk crystalline environment when calculating the molecular energy and electron density distribution leads to important changes in relative total crystal energies and correctly distinguishes the observed crystal structures from the set of computer-generated possibilities. This study highlights the importance of polarization of the molecular charge distribution in crystal structure prediction calculations, especially for polar flexible molecules, and suggests a computationally inexpensive approach to include its effect in lattice energy calculations.

## Introduction

The identification or prediction of the most stable crystalline form of an organic molecule is of considerable scientific interest but remains a challenge for both experimental and computational methods. In particular, to avoid any unanticipated polymorphic change during manufacturing it is vitally important in the pharmaceutical industry to know that the crystalline form being produced is the thermodynamically most stable. Such a change in form could have unpredictable effects on processing and tabletting of the drug as well as its final bioavailability, since polymorphs can have markedly different dissolution rates.[1] For this reason, great amounts of time and money are spent screening for polymorphs to maximize the likelihood that all crystal forms of the drug molecule have been found.[2] If theoretical studies could identify all possible crystal structures of a particular organic molecule and reliably predict their relative thermodynamic stabilities, such calculations could help inform and possibly direct the current experimental approaches used to produce the different crystal forms.

A variety of computational methods have been developed for crystal structure prediction (CSP), with the aim of producing the possible crystal structures of a particular molecule using only the chemical formula as input.[3] The difficulty in CSP is not necessarily in generating all crystal packing possibilities, including the observed crystal structure (or structures, where the molecule displays polymorphism), but in identifying those structures that will be experimentally observed from the many putative structures; often there are tens or even hundreds of distinct crystal structures (local minima on the lattice energy surface) within a small (e.g., 5 kJ $mol^{-1}$) range in lattice energy,[4] and the assumption is that the lowest energy structure is the most likely structure to be observed experimentally. CSP of rigid organic molecules is well developed, and, using a sufficiently high quality model intermolecular potential, the observed crystal structure(s) are quite reliably found among a small set of the lowest energy computationally generated structures.[5] How-

* Corresponding author e-mail: gmd27@cam.ac.uk.

10.1021/ct800195g CCC: $40.75 
Published on Web 09/11/2008

ever, predicting the crystal structures of molecules with conformation freedom is an entirely more difficult proposition. There are two causes of difficulties associated with molecular flexibility: a) the conformational search space that needs to be sampled when searching for all crystal packing possibilities, with each degree of intramolecular freedom adding an extra dimension to the already large and complex search space, and b) comparison of the energies of putative crystal structures, which involves calculation of a lattice (*inter*molecular) and a conformational (*intra*molecular) component for the generated crystal structures, with the two components needing to be both accurate and balanced with respect to each other.

We have chosen the series of α-amino acid crystal structures as a test set for the development of CSP methods for conformationally flexible molecules. This family of molecules was chosen for several reasons: their biological importance; the importance of amino acid functional groups in pharmaceutical molecules; and because the series of similar molecules allows stepwise progression from systems with very limited molecular flexibility to larger, more flexible molecules. The striking differences in conformations between the gas phase and the solid state make amino acids very challenging systems to study, especially using the most common approach to CSP, where molecular geometries are derived from isolated molecule calculations. A study by Görbitz[6] of the chiral hydrophobic α-amino acids, using hydrogen bonding donor and acceptor sites to simulate the surrounding crystal, has shown that the conformations found in the crystals can be reproduced by quantum mechanics methods. However, there is flexibility about both the carboxylate and amino groups in the α amino acids, and the observed conformations in crystal structures vary considerably, e.g. the $H_\alpha-C_\alpha-C-O$ dihedral angle in the four published polymorphs of glycine in the Cambridge Structural Database[7] (CSD) varies between 35° and 84°.

One approach to CSP for moderately flexible molecules is to consider a set of fixed molecular geometries, normally obtained from quantum mechanical (QM) calculations *via* a scan of the flexible intramolecular degrees of freedom. Alternatively, conformations can be chosen by comparison to known crystal structures of similar molecules. In both approaches, crystal structures are generated with each conformation, and the total energies used to evaluate the relative stabilities of the resulting crystal structures are calculated as

$$U_{total} = \sum_{i\in M, k\in N} [A_{\iota\kappa}\exp(-B_{\iota\kappa}R_{ik}) - C_{\iota\kappa}R_{ik}^{-6} + U_{electrostatic}(\rho_M, \rho_N)] + U_{molecular}(\mathrm{QM}, \rho) \quad (1)$$

where the terms in the summation represent intermolecular interactions in the crystal, and $U_{molecular}$ is the energy of the molecular conformation in that crystal structure. Here, $U_{molecular}$ is taken from the QM evaluation of the molecular energy. $A_{\iota\kappa}$, $B_{\iota\kappa}$, and $C_{\iota\kappa}$ are empirically derived parameters describing the repulsion-dispersion interactions between atoms $i$ and $k$ in molecules M and N; $\iota$ and $\kappa$ are the atom types of atoms $i$ and $k$, respectively. $U_{electrostatic}$ is calculated from a representation of the molecular charge density, $\rho$, either by a set of partial charges distributed around the molecule (typically at atomic positions) or by a multipole expansion (i.e., charge, dipole, quadrupole, etc.) at each atomic site.

In preliminary studies of the crystal packings of valine and other flexible molecules, we have observed that the above approach, combining atomistic (*inter*molecular) with quantum mechanical (*intra*molecular) energies, resulted in an imbalance between the *inter*- and *intra*molecular contributions to the relative energies of putative crystal structures. In particular, the intramolecular energies were often found to dominate the ranking of structures, and, while the true crystal structures stand out as having among the best intermolecular energies, those with the lowest *total* energies all had favorable *intra*molecular energies but poor *inter*molecular energies. The energy model is dominated by the QM calculated conformational energy. We therefore sought to correct the imbalance between the two energy contributions in the computational method, which we propose is largely due to molecular polarization being ignored in the intermolecular model, i.e. the electrostatic interactions in eq 1 are based on the charge density ($\rho$) of the isolated molecule.

Deriving the electrostatic model from isolated molecule charge densities ignores the rearrangement of the molecular electron density due to the crystalline environment, which is known to be important in molecular crystals.[8] This polarization serves to lower the total crystal energy by strengthening the intermolecular interactions between molecular electron densities. Therefore, models derived from isolated molecule (i.e., nonpolarized) charge densities will underestimate intermolecular electrostatic stabilization energies, which are dominant for systems with significant charge separation, such as salts or zwitterionic molecules (such as the α-amino acids). One strategy used to model induction effects is to include molecular or atomic polarizabilities in the atom-atom model, as is done in polarizable force fields.[9,10] The theory for calculating the resulting induction energy has been presented elsewhere,[11,12] and methods are continually developing for the derivation of such atomic polarizabilities.[13,14] An alternative is to perform the molecular charge density calculation in an environment representative of the crystal, so that the atomic partial charge or multipole analysis is performed on a molecular electron density that is a better representation of the molecule in the crystal. The two approaches have recently been compared in an investigation of the magnitude of induction energy contributions to the energies of molecular organic crystal structures.[15] We have followed the latter approach here and investigate the influence of a very simple description of the bulk crystal environment during the molecular QM calculation: the environment of the molecule in the crystal is modeled as a polarizable continuum, in the same way that solvation effects on molecular properties are often modeled, using Tomasi and co-workers' polarizable continuum model (PCM)[16−18] with dielectric constants typical of molecular organic crystals.

We test this approach for valine, the α-amino acid with an isopropyl side group, and, as such, one more flexible torsion angle than alanine, whose crystal structures we

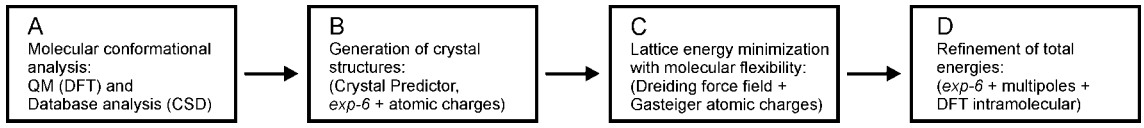



***Figure 1.*** Outline of the crystal structure prediction methodology applied here to valine.

studied previously.[19] The two known racemic polymorphs (monoclinic,[20] CSD refcode VALIDL and triclinic[21] VALIDL02, both with $Z' = 1$) have almost identical molecular conformations and crystal packings and only differ by the relative orientation of pairs of hydrogen bonded layers. Enantiopure valine crystallizes with two molecules in the asymmetric unit[22] ($Z' = 2$, refcode LVALIN01) and, while we have not performed the computationally expensive task of generating and energy minimizing all possible $Z' = 2$ structures, we compare the calculated energy of the known structure with the computer-generated $Z' = 1$ alternative crystal structures to confirm that the observed $Z' = 2$ form is lower in energy than the $Z' = 1$ possibilities.

## Methods

**Overall Approach.** We combine our general strategy for treating molecular flexibility in CSP[23] with database guided sampling of conformational space,[19] which is necessary because of the extreme differences in molecular structure between the gas and crystalline phases for the α-amino acids.

The procedure comprises 4 steps (Figure 1):

A) analysis of the molecular conformations using both QM electronic structure calculations and crystal structures of similar molecules from the CSD; (B) generation of possible crystal packings with rigid molecular conformations from (A) *via* a sampling of unit cell parameters, molecular positions, and molecular orientations, within the most likely space groups; C) energy minimization of the computer-generated crystal structures with a molecular mechanics description of angle bending and torsion angles, allowing selected intramolecular degrees of freedom to adjust to the crystal environment, and D) final energy refinement using an accurate description of intermolecular interactions, in particular using a detailed description of the molecular electrostatic distribution.

The emphasis of the current study is on the final step in the procedure, where we aim to develop methods for evaluating the relative energies of the computer-generated crystal structures as accurately as possible, bearing in mind the associated computational expense and the need for methods of evaluating thousands of crystal structures in a reasonable time. As a result of the strong electrostatic interactions in the crystal structures of zwitterionic molecules, we chose this system to test a simple method of accounting for polarization effects on the calculated relative energies.

**Molecular Conformational Analysis.** We treated valine as having three flexible torsion angles (Figure 2), assuming that the methyl group orientations are unimportant to the crystal packing. For geometry optimizations and the torsion angle scan of the isopropyl group (Step A), we performed density functional theory (VWN/DNP) calculations on the isolated molecule using the Dmol3 module[24] within the

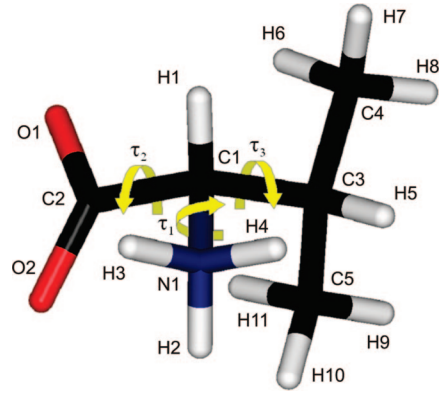



***Figure 2.*** Molecular model of valine showing the definition of the three flexible torsion angles ($\tau_1 = H_2-N_1-C_1-H_1$, $\tau_2 = O_1-C_2-C_1-H_1$, $\tau_3 = H_5-C_3-C_1-H_1$). The amino torsion angle ($\tau_1$) is shown in the staggered conformation.

Accelrys Materials Studio package.[25] Full geometry optimization of the valine molecule leads to the nonzwitterionic form, so it was necessary to constrain the three N−H bond lengths to maintain the zwitterionic form; we used a N−H bond length of 1.035 Å, the mean value from neutron diffraction α-amino acid crystal structures in the CSD. The isopropyl group of valine was scanned for the full range of torsion angles ($0° \le \tau_3 \le 360°$) in 20° intervals, with the angle measured between the α-hydrogen atom and the β-hydrogen atom of the isopropyl group ($\tau_3 = H-C-C-H$, Figure 2). The amino group was kept fixed in the staggered conformation ($\tau_1 = H-N-C-H = 180°$), while all other degrees of freedom were allowed to relax at each point in the scan.

**Generation of Crystal Structures.** We generated trial crystal structures (Step B) using the Crystal Predictor code, which employs a low-discrepancy sequence to search the crystal packing space with quasi-random values for unit cell parameters, molecular orientations, and positions followed by rigid molecule lattice energy minimization.[26] Searches were performed with a set of 15 rigid molecular geometries (see Results and Discussion), chosen in a similar way to our crystal packing study of alanine,[19] to sample the relevant region of the conformational energy surface. Crystal structures were generated in 12 space groups: the 5 most commonly observed chiral space groups for organic molecular crystals ($P2_12_12_1$, $P2_1$, $P1$, $P2_12_12$, and $C2$) and the 7 most common space groups with mirror or inversion symmetry ($P2_1/c$, $Pna2_1$, $Pnma$, $C2/c$, $P\bar{1}$, $Pbca$, and $Pbcn$). These account for over 90% of all the entries found in chiral and racemic space groups in the CSD.[27] Searches were continued until 50000 lattice energy minimizations had been performed for each conformation. We monitored convergence of the set of low energy structures, and 50000 minimizations were sufficient to give an apparently complete sampling in all cases. We then merged the set of predicted crystal structures

from each molecular conformation (keeping racemic and chiral crystal structures separate), and the structures were clustered to remove duplicates (the clustering algorithm is described below).

**Model Potentials.** For lattice energy minimizations during the initial crystal structure search (step B), we used an *exp-6* model potential with Williams and co-workers' empirically derived parameters to describe C, N, O, and $H_C$ (hydrogen bonded to carbon)[28,29] and polar hydrogen atom ($H_N$) parameters taken from Coombes et al.[30] Electrostatic interactions were modeled by molecular electrostatic potential (ESP) fitted atomic partial charges obtained from a single point DFT calculation of the gas phase molecule (DMol3, VWN/DNP). For the final energy minimizations (Figure 1, step D), we applied the same *exp-6* model potential but with a more elaborate description of electrostatic interactions, using a distributed multipole model. Atomic multipoles, $Q^i{}_u$, up to hexadecapole on each atom $i$ (the subscript $u$ refers to the multipole component), were taken from a distributed multipole analysis (DMA) of a B3LYP/6−31G** electron density, calculated for the specific molecular conformation under consideration using the Gaussian03 program.[31] The DMA was performed using the program GDMA,[32] using the original DMA algorithm.[33,34] Final lattice energy minimizations were performed using the DMAREL crystal structure modeling program.[35] All *exp-6* interactions were evaluated up to a 15 Å cutoff, Ewald summation was employed for charge−charge, charge−dipole, and dipole−dipole electrostatic interactions, and all higher order electrostatic interactions (up to $R^{-5}$) were summed to a 15 Å cutoff on whole molecules.

Crystal structures were also lattice energy minimized using the same *exp-6* model potential with distributed multipoles derived from polarized molecular electron densities. These polarized electron densities were calculated using the polarizable continuum model (PCM),[16−18] as implemented in Gaussian03, to model the environment of the molecule. The united atom topological model (UA0) was used for atomic radii, with several choices of dielectric constant (ranging from $\epsilon = 3$ to 11). As with the unpolarized electrostatic model, the GDMA program was used to perform the distributed multipole analysis of the resulting charge density, and lattice energy minimizations were performed within the program DMAREL. The resulting atomic multipoles (from the PCM molecular calculations) differ from the unpolarized multipoles (the isolated molecule) by $\Delta Q^i{}_u$, and the resulting induction energy is a sum of two terms: the amplified intermolecular electrostatic interactions and the partly counteracting increase in internal molecular energy ($\Delta U_{mol}$) upon distortion of the molecular charge distribution. In the purely classical approach, a classical expression is used for the internal molecular energy contribution (i.e., assuming that the molecular energy increases bilinearly with the $\Delta Q$: $\Delta U_{mol} = {}^1/_2 \sum_{u,u',i,i'} \Delta Q^i{}_u \xi^{ii'}{}_{uu'} \Delta Q^{i'}{}_{u'}$).[11,12] The resulting expression for the induction energy requires the interaction of $\Delta Q^i{}_u$ on a reference molecule with $Q^i{}_u$ on all surrounding molecules. Using standard lattice energy modeling software, where all molecules possess the same set of atomic multipoles, two calculations are required to obtain this energy: one where all molecules are assigned atomic multipoles ($Q^i{}_u + \Delta Q^i{}_u/2$) and a second where all molecules have atomic multipoles ($\Delta Q^i{}_u/2$).[15] However, as we perform a molecular QM calculation for each crystal structure in this work, we can avoid the classical approximation to $\Delta U_{mol}$. Instead, we take the increase in molecular energy directly from the density functional theory calculations on the isolated molecule and the molecule in the PCM environment, e.g. at $\epsilon = 3$ we take $\Delta U_{mol} = U_{mol}[\text{DFT}, \rho(\epsilon{=}3)] - U_{mol}[\text{DFT}, \rho(\text{vacuum})]$, where the interaction energy between the molecule and the dielectric continuum is excluded from $U_{mol}[\text{DFT}, \rho(\epsilon{=}3)]$. We then use the total polarized atomic multipoles ($Q^i{}_u + \Delta Q^i{}_u$) in our intermolecular energy calculations, which give the sum of electrostatic and the intermolecular part of the induction energy.

The final total crystal energies (in step D) were calculated as the sum of inter- and intramolecular energy terms, the *inter*molecular contribution taken from the *exp-6* + DMA model, and the *intra*molecular contribution from the DFT molecular energy calculation. The magnitude of the induction energy can be evaluated from the difference in the calculated intermolecular electrostatic energies from the $\rho$(PCM,$\epsilon$) and $\rho$(vacuum) models + $\Delta U_{mol}$:

$$U_{induction}(\varepsilon) = (U_{electr}[\rho(\varepsilon)] - U_{electr}[\rho(\text{vacuum})]) + (U_{mol}[\text{DFT},\rho(\varepsilon)] - U_{mol}[\text{DFT},\rho(\text{vacuum})]) \quad (2)$$

**Flexible Force Fields.** We tested four flexible molecule force fields (CVFF,[36] COMPASS,[37] Dreiding,[38] and UFF[39]) for the intermediate lattice energy calculation where molecular flexibility is allowed (Step C). The purpose of this step in the procedure is to allow the molecular geometry, which was held rigid during the initial generation of trial crystal structures, to adjust to its crystal environment. Force field partial charges were used for the CVFF and COMPASS force fields, while Gasteiger partial charges[40] were used with UFF and the Dreiding force field. Using a similar method in our study of phenobarbital crystal structures,[23] the computer-generated crystal structures were energy minimized with the molecules treated as comprising four rigid units: the amine ($NH_3$), carboxylate ($CO_2$), and isopropyl ($C_3H_7$) groups and the central CH to which they are all bonded, with reorientation of rigid units allowed relative to one another using the force field energy terms for the conformational energy. The internal structure of the rigid units was constrained at the DFT optimized geometries.

**Comparison and Clustering of Structures.** To compare the predicted and observed crystal structures and for removal of duplicate crystal structures ("clustering"), we used the Compack algorithm[41] with a tolerance of 20% on interatomic distances (excluding hydrogen atoms) within a cluster of 15 molecules (i.e., a central molecule and a coordination sphere of its 14 nearest neighbors). Clustering was used to remove duplicate crystal structures between steps B and C (Figure 1) and after the final energy minimizations in step D. After step B some crystal packings are found in searches using more than one molecular model (i.e., different values of $\tau_2$ and $\tau_3$), with small differences in packing caused by the differences in molecular geometry. The 20% tolerance on interatomic distances was chosen as sufficiently relaxed to

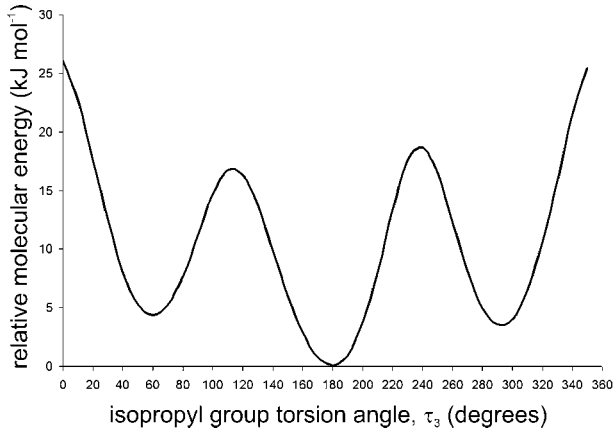



***Figure 3.*** Conformational energy profile for rotation of the isopropyl group orientation $\tau_3$ (with $\tau_1 = 180°$ and $\tau_2 = 0°$), calculated at the VWN/DNP level of theory (DMol3).

cluster such structures (*i.e.* treating them as identical) but not remove structures with distinct packing arrangements. The crystal structure with the lowest total energy from each cluster is retained, and, in this way, the version of each crystal structure with its most favorable conformation is retained and moved forward to the structural and energetic refinements in the latter steps.

## Results and Discussion

**A. Molecular Conformational Analysis.** Conformations for the crystal structure searches were chosen using a similar approach to our previous study of alanine:[19] the amino group was kept fixed in the staggered conformation ($\tau_1 = 180°$), and conformations were taken with the carboxylate torsion angle ($\tau_2$) incremented in 15° steps from 15° to 75°, a range chosen from an analysis of similar molecules in the CSD. 10° increments in $\tau_2$ were used in our alanine calculations, but postanalysis found that a coarser sampling (20° steps) would have sampled conformational space sufficiently.[19] Here, we use the intermediate step size of 15°, which should generate as close to the full set of low energy crystal structures as possible while lowering the number of crystal structure generation calculations. Starting points for $\tau_3$ were chosen from a DFT torsional energy scan of the isopropyl group (Figure 3), which showed three local minima, corresponding to staggered configurations of the $\beta$-carbon to α-carbon bond. The five orientations of the carboxylate group were combined with the three isopropyl group minimum energy orientations, giving fifteen conformations with which trial crystal structures were generated.

**B. Generation of Crystal Structures.** To check that the search located the known racemic valine polymorphs, the crystal structures generated with the rigid molecule conformations (Figure 1, step B) were compared with the true DL-valine crystal structures extracted from the CSD; good representations of both polymorphs were located with several of the molecular conformations (Table 1). However, the energetic rankings of the known structures among the predictions are poor, even among the separate lists from each molecular conformation, both in terms of the number of lower energy computer-generated crystal structures than the known structure, $N_{lower}$[4] (a perfect prediction would give $N_{lower} = 0$), and the energy difference between the known structure and the lowest energy computer-generated structure, $\Delta E$ (we are aiming for low values, with $\Delta E < 0$ if no unobserved computer-generated crystal structures have better calculated energies than the known polymorphs). For small rigid molecules, $\Delta E$ is usually less than 2−3 kJ mol$^{-1}$ and $N_{lower}$ is normally less than 5, when a high quality intermolecular model potential is used.[5]

***Table 1.*** Rigid Conformations That Led to a Match to the Known Crystal Structures of DL-Valine

| torsion angles | | | VALIDL ($P2_1/c$) | | VALIDL02 ($P\bar{1}$) | |
|---|---|---|---|---|---|---|
| $\tau_1$ | $\tau_2$ | $\tau_3$ | $N_{lower}$[a] | $\Delta E$[b]/kJ mol$^{-1}$ | $N_{lower}$[a] | $\Delta E$[b]/kJ mol$^{-1}$ |
| 180° | 15° | 60° | 28 | 10.3 | 16 | 8.2 |
| 180° | 30° | 60° | 26 | 7.8 | 16 | 6.5 |
| 180° | 45° | 60° | 36 | 6.4 | 30 | 5.3 |
| 180° | 60° | 60° | 105 | 13.9 | 88 | 11.9 |

[a] $N_{lower}$ is the number of lower energy predicted crystal structures among the set of structures generated for that conformation. [b] $\Delta E$ is the energy difference between the computer-generated version of the observed crystal structure and the lowest energy predicted crystal structure for that molecular conformation.

These poor rankings can partly be attributed to the use of rigid molecular geometries when generating the trial crystal structures, as $\tau_3$ in the observed DL-valine crystal structures ($\tau_3 \sim 80°$) is approximately 20° away from the closest minimum on the conformational energy surface (Figure 3). These results highlight a limitation of the method previously applied to the crystal structure prediction of alanine, in that the true crystal structures might be poorly ranked if the starting molecular geometries differ significantly from their optimum conformations in the crystal structures. The intermediate energy minimization step (Figure 1, step C) in which the molecular conformation can adjust to the crystal structure is included here to alleviate this problem.

**C. Lattice Energy Minimization with Molecular Flexibility.** The results using rigid molecular models fixed at the set of initial molecular geometries (Table 1) are clearly unsatisfactory, so all crystal structures were energy minimized again but without the rigid molecule constraints: the important torsion angles ($\tau_1$, $\tau_2$, $\tau_3$, Figure 2) were given freedom to adjust to their crystal packing environments. We use force field descriptions of intramolecular degrees of freedom in this step and are mainly concerned with the force field's ability to provide a reliable description of the molecular structure in the crystal. Therefore, we tested a set of force fields by energy minimizing the known crystal structures of two molecules: L- and DL-alanine and leucine, the amino acids with a slightly smaller (alanine) and larger (leucine) hydrophobic side group than valine. We then compared a selection of torsion angles in the optimized crystal structures with those in the experimentally determined structures. Based on these tests (detailed results are deposited as Supporting Information), we chose the Dreiding force field as providing the best results. Therefore, we used the Dreiding force field for the flexible molecule energy minimization of the computer-generated crystal structures of valine.

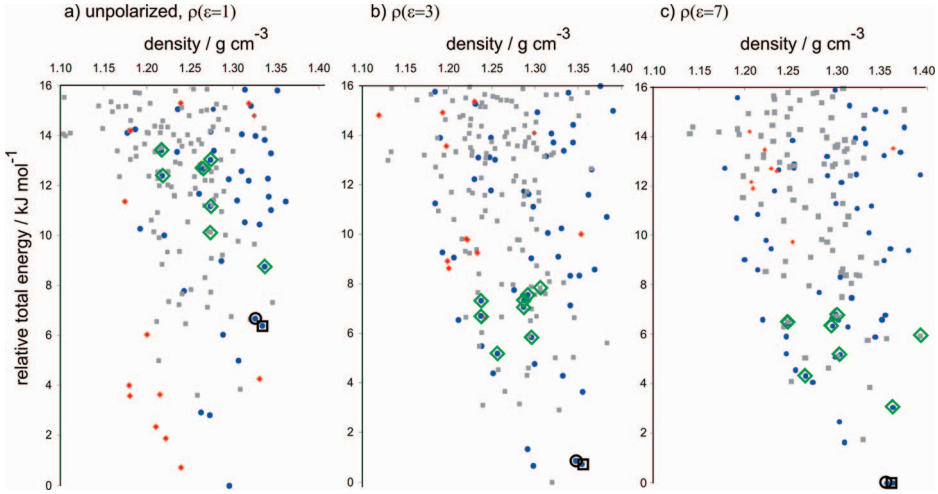



***Figure 4.*** Plot of density vs relative energy for the computer-generated crystal structures of DL-valine after minimization with the Dreiding flexible force field followed by calculation of the total energies using the *exp-6* model potential, atomic multipoles, and DFT for the molecular energies. Electrostatic models and relative molecular energies were taken from a) the isolated molecule, b) PCM calculations with $\epsilon = 3$, and c) PCM calculations with $\epsilon = 7$. Structures are color coded by their lattice (intermolecular) energies: blue circles have intermolecular energies within 20 kJ mol$^{-1}$ of the lowest lattice energy structure; gray squares have intermolecular energies between 20 and 40 kJ mol$^{-1}$ above the lowest lattice energy structure; and red diamonds have intermolecular energies more than 40 kJ mol$^{-1}$ above the lowest lattice energy structure. The known crystal structures of DL-valine are indicated by the open circle (monoclinic, CSD refcode VALIDL) and open square (triclinic, CSD refcode VALIDL02). Structures highlighted by open green diamonds show high structural similarities to the known crystal structures.

During this step, valine molecules in the computer-generated crystal structures were treated as comprising four rigid units; their relative orientations were allowed to relax at this stage while constraining the internal geometries of these rigid units at their DFT optimized geometries.

**D. Refinement of the Lattice Energies.** Clustering of the structures after energy minimization with the Dreiding force field resulted in 1130 distinct crystal structures. Each of these was then subjected to the final energy minimization step, where the molecular geometry was constrained at the geometry resulting from the Dreiding force field lattice energy minimization. The molecular geometry was extracted from *each* crystal structure, and a single point DFT molecular energy calculation was performed to obtain the molecular energy and electron density distribution. A distributed multipole analysis was performed on the resulting wave function, and then the crystal structure was lattice energy minimized using the *exp-6* + DMA model intermolecular potential. Total crystal energies were calculated as a sum of the atom-atom intermolecular energy and the DFT molecular energy.

We first examined the resulting set of crystal structures when the molecular properties (energy and atomic multipoles) were taken from isolated molecule DFT calculations. The results are summarized *via* a relative total energy vs density plot (Figure 4a), on which each point represents a distinct crystal structure. Furthermore, we color coded the points according to the intermolecular contribution to the total crystal energy.

Many of the crystal structures with lowest total energies using this model have poor *inter*molecular contributions to their energies (red diamonds in Figure 4a, which cover intermolecular energies in the range −172 to −152 kJ mol$^{-1}$). These crystal structures are clearly stabilized by favorable *intra*molecular energies, at the expense of strongly stabilizing *inter*molecular interactions. The computer-generated structures that correspond to the known polymorphs of DL-valine are ranked as the 15th (triclinic, VALIDL02) and 19th (monoclinic, VALIDL) lowest energy structures, 6.4 kJ mol$^{-1}$ and 6.7 kJ mol$^{-1}$ higher in energy than the computer-generated crystal structure with the lowest energy. While this is an improvement over the ranking with rigid molecular geometries (Table 1), the results are still unsatisfactory for crystal structure prediction, where we expect to find any observed crystal structures among the lowest energy structures. For rigid molecules, observed crystal structures are rarely outside of the 10 lowest energy structures and never more than about 5 kJ mol$^{-1}$ above the lowest energy prediction.[5] Discounting the predicted crystal structures with very poor intermolecular packing energies (red diamonds in Figure 4a) improves the situation somewhat, but the known polymorphs remain farther from the global minimum than we would expect, suggesting inadequacies in the model we have used to evaluate the total relative energies.

**The Influence of Molecular Polarization.** We then explored the influence of including a polarizing environment when calculating the molecular energies and charge density distributions. The crystal structures were energy minimized using electrostatic models (distributed multipoles) derived from molecular wave functions calculated in a series of polarizing environments: the polarizable continuum model (PCM) with dielectric constants, $\epsilon$, of 3 (typical for crystals of neutral organic molecules), 7 and 11 (chosen to represent the very polar environment in crystals with extreme charge separation). The effect of the continuum dielectric environment is to stabilize a greater charge separation in the molecule, influencing both the molecular energy and atomic

***Table 2.*** Unit Cell Parameters and Relative Energies of the Observed and Predicted Crystal Structures of DL-Valine

| | a/Å | b/Å | c/Å | α/° | β/° | γ/° | $V_{mol}$[a]/Å$^3$ | $\Delta E$[b]/kJ mol$^{-1}$ |
|---|---|---|---|---|---|---|---|---|
| Monoclinic DL-Valine | | | | | | | | |
| expt (room temperature, VALIDL) | 5.21 | 22.10 | 5.41 | 90 | 109.2 | 90 | 294.2 | - |
| predicted (unpolarized ρ) | 5.259 | 22.185 | 5.284 | 90 | 107.9 | 90 | 293.4 | +6.7 |
| predicted, ρ(ϵ=3) | 5.233 | 22.208 | 5.220 | 90 | 107.8 | 90 | 288.8 | +0.9 |
| predicted, ρ(ϵ=7) | 5.225 | 22.219 | 5.195 | 90 | 107.9 | 90 | 287.0 | 0 |
| predicted, ρ(ϵ=11) | 5.222 | 22.223 | 5.187 | 90 | 107.9 | 90 | 286.3 | 0 |
| Triclinic DL-Valine | | | | | | | | |
| expt (120 K, VALIDL02) | 5.222 | 5.406 | 10.838 | 90.9 | 92.3 | 110.0 | 287.1 | - |
| predicted (unpolarized ρ) | 5.252 | 5.287 | 11.139 | 89.7 | 82.4 | 107.9 | 291.5 | +6.4 |
| predicted, ρ(ϵ=3) | 5.221 | 5.224 | 11.155 | 97.1 | 90.3 | 107.8 | 287.3 | +0.7 |
| predicted, ρ(ϵ=7) | 5.194 | 5.214 | 11.473 | 100.1 | 90.4 | 107.8 | 290.8 | +0.0 |
| predicted, ρ(ϵ=11) | 5.187 | 5.212 | 11.166 | 96.8 | 90.4 | 107.8 | 285.1 | +0.1 |
| L-Valine | | | | | | | | |
| expt (120 K) | 9.682 | 5.247 | 11.930 | 90 | 90.6 | 90 | 303.0 | - |
| minimized[c] (unpolarized ρ) | 9.787 | 5.228 | 11.735 | 90 | 90.5 | 90 | 300.3 | −9.1 |
| minimized,[c] ρ(ϵ=3) | 9.666 | 5.195 | 11.752 | 90 | 90.5 | 90 | 295.0 | −16.4 |
| minimized,[c] ρ(ϵ=7) | 9.613 | 5.182 | 11.760 | 90 | 90.5 | 90 | 292.9 | −17.3 |
| minimized,[c] ρ(ϵ=11) | 9.595 | 5.178 | 11.764 | 90 | 90.5 | 90 | 292.3 | −16.9 |

[a] $V_{mol}$ is the crystal structure volume per molecule. [b] $\Delta E$ is the sum of the intermolecular lattice energy and the molecular energy (relative to the energy of the lowest energy conformation in any of the computer-generated crystal structures). [c] The result of energy minimization of the experimentally determined crystal structure using the same molecular model and procedure as used in the prediction calculations.

multipoles (the electrostatic model in the lattice energy calculation). The results, summarized as energy vs density plots for $\epsilon = 3$ and $\epsilon = 7$ in Figure 4b,c, show a strong dependence of the energetic ordering of the predicted crystal structures with the choice of $\epsilon$. The crystal structures with poor intermolecular energies (red diamonds) are disfavored when the polarized molecular charge distributions are used, such that these structures are all outside of the lowest 8 kJ mol$^{-1}$ range on total energy when using even a modestly polarized model ($\epsilon = 3$). As the computer-generated crystal structures with poor intermolecular energies are disfavored, those with the best intermolecular energies (blue circles, Figure 4) dominate the low energy region, and the ranking of the two known DL-valine polymorphs improves dramatically. The monoclinic (VALIDL) and triclinic (VALIDL02) polymorphs are the fourth and third lowest energy structures overall when using molecular energies and electrostatics derived with $\epsilon = 3$ and are first and second (i.e., the two lowest energy structures) with both $\epsilon = 7$ and $\epsilon = 11$. By calculating the molecular energies and charge density distributions in a polarizing environment, the ranking on total energy predicts the two known polymorphs perfectly, a dramatic improvement over results using unpolarized molecular models.

Table 2 summarizes structural information for the observed and predicted crystal structures of valine. Unit cell dimensions are in line with what should be expected. Cell dimensions and angles are all within 3% of the observed values, while the crystal structure volumes per molecule ($V_{mol}$) are all within 4%. There is a trend of compaction of the predicted structures with increased polarization (higher $\epsilon$), because of the enhanced attractive interactions between polarized molecular charge densities. While the volumes of the predicted crystal structures minimized with multipoles calculated in a vacuum are close to the observed volumes, those using the polarized molecular models are contracted slightly. The change in volume with the polarized models is consistent with the expansion of amino acid structures from low temperature to room temperature[42] and the temperature-free (i.e., classical 0 K) nature of the predictions.

**Analysis of the Components of the Total Energy.** To determine the origin of the striking changes in the relative total crystal energies upon molecular polarization, the individual components of the total crystal energies (see eq 1) were further examined. The two energy contributions that are directly a function of the calculated electron density, $\rho$, are the DFT molecular energy and the calculated electrostatic component of the intermolecular energy; both include contributions from the induction energy when calculated from the polarized molecular charge densities. The variation of these two contributions with the value of $\epsilon$ used in the molecular PCM calculation is shown for the 20 lowest energy crystal structures (taken from the original set, Figure 4a) in Figure 5a,b. The isolated molecule calculations are included as $\epsilon = 1$.

The *absolute* values of the molecular energies in the 20 crystal structures increase by between 18.6 and 23.8 kJ mol$^{-1}$ between isolated molecule calculations and PCM calculations with $\epsilon = 11$. However, the subsequent effect on the ranking of the crystal structures is small because the *relative* molecular energies among the conformations found in the low energy crystal structures only change by a few kJ mol$^{-1}$; the total range of molecular energies among these 20 crystal structures is 51 kJ mol$^{-1}$ from the isolated molecule calculations and 55 kJ mol$^{-1}$ in the most polarizing environment ($\epsilon = 11$), Figure 5a. This change in relative intramolecular energies has some influence on the final crystal structure ranking but is not the main cause of the important changes in total relative crystal energies when the polarized molecular models are used. The molecular energies in both observed DL-valine crystal structures, relative to the most stable conformations found in any of the predicted crystal structures, *increase* from about 46 kJ mol$^{-1}$ (isolated molecules) to 51 kJ mol$^{-1}$ (PCM, $\epsilon = 11$), i.e. they have gained no stability relative to the other putative crystal

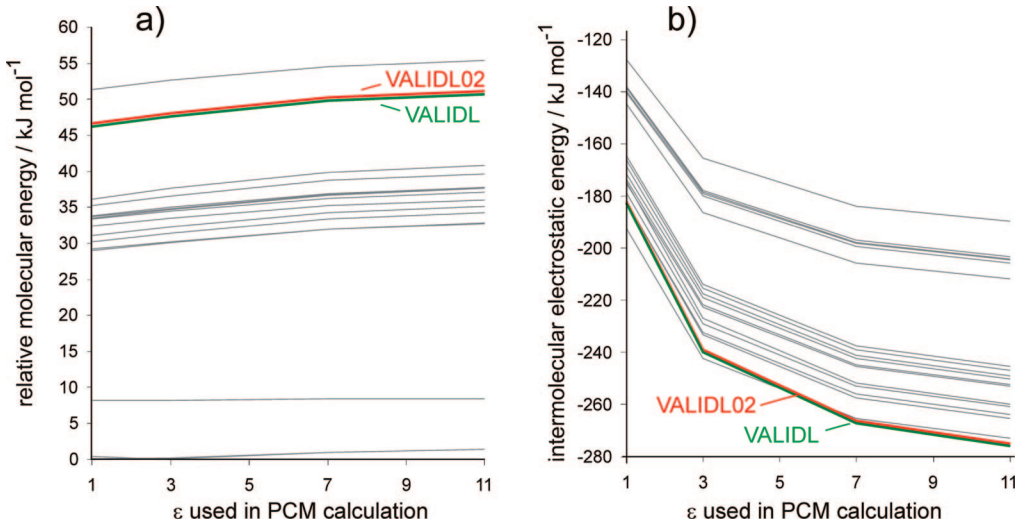



***Figure 5.*** Variation in a) relative molecular energy (relative to the energy of the lowest energy conformation in any of the crystal structures) and b) intermolecular electrostatic energy as a function of the dielectric constant, $\epsilon$, used in the molecular PCM calculation for the 20 lowest energy computer-generated crystal structures of DL-valine (from the $\rho(\epsilon=1)$ ranking).

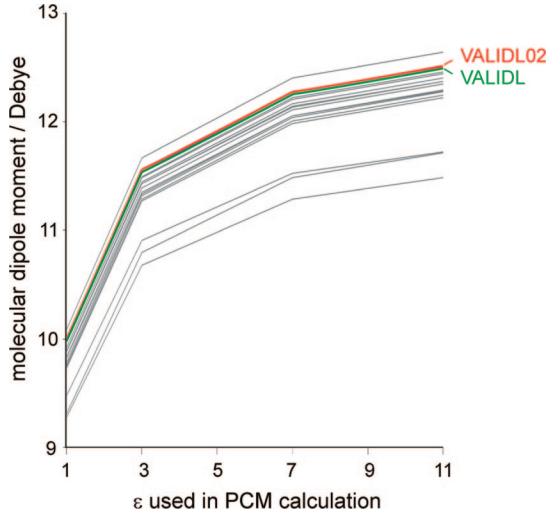



***Figure 6.*** Variation in the molecular dipole moment as a function of the dielectric constant, $\epsilon$, used in the molecular PCM calculation for the molecular geometries from the 20 lowest energy computer-generated crystal structures of DL-valine (from the $\rho(\epsilon=1)$ ranking).

structures in terms of intramolecular energy by accounting for the polarizing environment.

**Electrostatic Energy.** A much more pronounced effect is seen in the electrostatic contribution to the intermolecular energies in the crystal structures (Figure 5b). The cause of the large changes in calculated electrostatic energy is clear from the changes in molecular dipole moments between the vacuum calculations and those in a dielectric medium (Figure 6). The molecular dipole moments calculated for isolated molecules, which range from 9.3 to 10.1 Debye for the conformations in these 20 crystal structures, are far too low compared to the experimentally observed dipole moment from analysis of the X-ray determined charge density of triclinic DL-valine, which is 14.3(4) Debye.[43] The greatest change in dipole moments comes between the vacuum calculations and PCM ($\epsilon = 3$) calculations, where the molecular dipole enhancement ranges from 15.1 to 15.9% for the conformations in these 20 crystal structures. The total increase up to $\epsilon = 11$ ranges from 23.7 to 25.8%, which is in good agreement with the 23% dipole moment enhancement of valine estimated from X-ray determined charge density analysis.[8] The molecular electron densities from the PCM calculations are much closer to the real charge distribution in the crystal than those calculated for molecules in a vacuum.

The plot of calculated electrostatic energy vs dielectric constant (Figure 5b) shows why the two observed structures have improved in ranking so drastically from the model based on vacuum calculations ($\epsilon=1$) to those taking into account the strongly polarizing crystal environment. If we take $\epsilon=3$ to be a realistic description of the bulk crystalline environment, we can estimate the induction energy for each computer-generated crystal structure from eq 2, as $[U_{electrostatic}(\epsilon=3) - U_{electrostatic}(\text{vacuum}, \epsilon=1) + \Delta U_{molecular}(\epsilon=3 - \text{vacuum})]$, a quantity that varies from $-30.1$ to $-47.8$ kJ mol$^{-1}$ among these 20 lowest energy crystal structures. This 17.7 kJ mol$^{-1}$ variation in induction energy between crystal structures is more than double the entire range in total energies for this set of putative crystal structures and demonstrates why the induction energy should not be ignored in crystal structure prediction, especially for such polar molecules, where the electrostatics dominate the lattice energies. The induction energy is greatest in the two crystal structures corresponding to the known DL-valine polymorphs (Figure 5b), and this energy contribution moves these structures from being poorly ranked with the unpolarized model to being among the best few structures in the list of crystal structures with polarization taken into account.

**L-Valine.** Enantiopure L-valine crystallizes in only one known crystal structure, in the space group $P2_1$ with two molecules in the asymmetric unit; the two molecules differ in the orientation of the isopropyl group, having $\tau_3 = 77°$ and 180° in the two independent molecules. While we did not generate crystal structures with two independent molecules, we can compare the observed crystal structure with the predicted $Z`=1$ alternatives. To do this, the observed

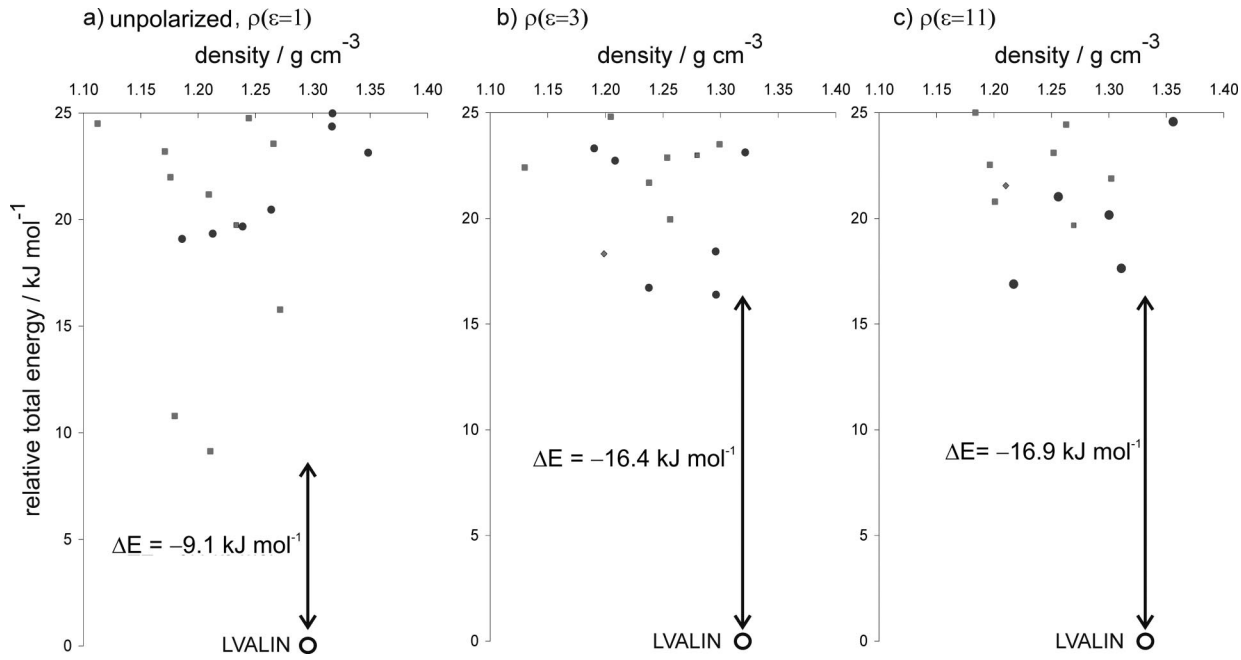



***Figure 7.*** Plot of relative energy against density for the computer-generated $Z`=1$ crystal structures of L-valine after minimization with the Dreiding flexible force field and then calculation of the total energies using the *exp-6* model potential, atomic multipoles, and DFT for the molecular energies. Electrostatic models and relative molecular energies were taken from a) the isolated molecule, b) a PCM calculation with $\epsilon = 3$, and c) a PCM calculation with $\epsilon = 7$. The observed $Z`=2$ crystal structure of L-valine (LVALIN01) is indicated by an open circle, and the energy difference between it and the lowest energy predicted $Z`=1$ structure is indicated.

structure (CSD refcode LVALIN01) was energy minimized using the same method as the computer-generated crystal structures: bond lengths and bond angles were adjusted to those in the DFT molecular structure used for the crystal structure prediction calculations, and the structure was relaxed allowing flexibility of the torsion angles with the Dreiding force field, followed by a final rigid-molecule lattice energy minimization with the *exp-6* + DMA intermolecular model potential. The total energy was calculated as a sum of this intermolecular energy and the DFT molecular energy of the final molecular conformations. The final step was repeated with molecular calculations performed with and without the PCM polarizing environment.

With all models (based on vacuum and PCM molecular charge distributions) the observed $Z`=2$ crystal structure is considerably more stable than any of the putative $Z`=1$ structures (Figure 7). The total energy difference between the observed structure and lowest energy $Z`=1$ alternative is 9.1 kJ mol$^{-1}$ before including the effects of molecular polarization, and this gap increases to over 16 kJ mol$^{-1}$ when including the polarization of the molecules. The large difference between the observed structure and any of the predicted structures suggests that a $Z`=1$ polymorph of enantiopure valine is unlikely to ever be observed, at least within the space groups considered here.

In light of the results here, the success of our previous crystal structure prediction study of alanine[19] seems fortuitous. The results for alanine, where the true crystal structures (of both L-alanine and DL-alanine) were predicted without considering molecular polarization, do not demonstrate that the polarization contribution to the lattice energy is less important for that molecule but perhaps that the induction energy varies less among the possible low energy crystal structures of L- and DL-alanine than those of DL-valine. This may be a result of less variation in molecular conformations among the low energy crystal structures of alanine. We are currently studying a wider set of amino acid crystal structures to test the performance of our computational methods on a large set of similar molecules.

## Conclusions

We have presented developments of our method for exploring the crystal packing landscape of flexible organic molecules, with the aim of crystal structure prediction. In this study, we have used computer-generated crystal structures of valine as a test of our approach and to evaluate a simple method for including the effects of polarization of the molecular charge density on the total relative energies of the known and putative polymorphs.

The results indicate that there is an imbalance in the intra- and intermolecular energy contributions to the total crystal energies if the molecular properties are taken from calculations on completely isolated molecules. To correct for this, molecular properties (energy and electron density distribution) have been calculated in a polarizing environment, described by the polarizable continuum model with dielectric constants chosen to be representative of molecular organic crystals (here, we tested values of $\epsilon$ ranging from 3 to 11). These calculations lead to molecular dipole moments that are more in line with X-ray charge density studies, and the changes in electrostatic and molecular energies give an estimate of the induction energy, which is very large for these crystal structures because valine crystallizes in zwitterionic

form. The PCM model contains no explicit information on the specific environment in each crystal structure, and the molecule in each crystal structure is immersed in the same structureless polarizing environment to polarize the molecular charge density. Therefore, the resulting polarization only distinguishes between crystal structures based on i) the conformational dependence of the molecular polarizability and ii) the different impact of molecular charge density polarization on intermolecular interactions in the various crystal structures. We view this as a first approximation to including induction energies in the assessment of relative stabilities of crystal structures, whereas a full treatment would need to model the specific, structured crystalline environment for each crystal structure.

Despite its apparent oversimplicity, the model leads to promising results here. The polarization favors those crystal structures with good intermolecular interactions over the other putative structures and, for both DL-valine and L-valine, preferentially stabilizes the true crystal structures over the many other low energy local minima. This leads to the two known polymorphs being ranked as the two lowest energy computer-generated crystal structures or as two of a small set of the lowest energy structures, depending on the value of $\epsilon$ when calculating molecular energies and charge distributions. The best choice of dielectric constant is uncertain, although the greatest effect on the molecular charge distribution is at low values and $2.5 \leq \epsilon \leq 4$ seems typical for crystals of neutral organic molecules.[44] Indeed, refractive index measurements[45,46] on L-amino acid crystal structures suggest values between about 2.25 and 2.5, which are slightly lower than static permittivities calculated by periodic DFT based lattice dynamics,[47] that suggest an average value of $\epsilon = 3.35$ for L-valine, with substantial anisotropy in the dielectric tensor. Empirically, we find that higher values of the dielectric constant perform better here: $\epsilon = 7$ results in a slightly better ranking of the known DL-valine crystal structures than $\epsilon = 3$, and higher values improve results even further. Higher dielectric constants also yield molecular dipole moments in better agreement with those from X-ray charge density studies. Our rationalization for these observations is that the higher values of $\epsilon$ might be compensating for the lack of specific interactions (such as hydrogen bonds) in the model; the calculations we use here introduce the polarizing effect of an average bulk environment but cannot model the strong polarizing effect of specific, directional intermolecular interactions. Side-by-side comparisons to polarization models that include the structure of the crystalline environment would be needed to investigate this further.

The results indicate that the effects of polarization in molecular crystals, which are usually ignored when modeling their structures and properties, can have a significant influence in calculations aimed at predicting the likely crystal structures of polar molecules. The method we have presented for including polarization in the calculations is computationally inexpensive and is being further tested and extended to the crystal structure prediction of more complex systems, in particular those of pharmaceutical interest such as more flexible molecules and multicomponent crystals (salts, cocrystals, and solvates).

**Acknowledgment.** We thank the Pfizer Institute for Pharmaceutical Materials Science for funding and Dr. Neil Feeder for helpful discussions. G.M.D. thanks the Royal Society for a University Research Fellowship.

**Supporting Information Available:** Structure files (CIF) of the 20 lowest energy predicted crystal structures of DL-valine with each of the four intermolecular electrostatic models (unpolarized molecules, $\epsilon = 3$, $\epsilon = 7$, and $\epsilon = 11$), structure files (CIF) of the 10 lowest energy predicted crystal structures of L-valine with each of the four intermolecular electrostatic models (unpolarized molecules, $\epsilon = 3$, $\epsilon = 7$, and $\epsilon = 11$), and tables summarizing the results of the testing of flexible molecule force fields for the crystal structures of alanine and leucine. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Chemburkar, S. R.; Bauer, J.; Deming, K.; Spiwek, H.; Patel, K.; Morris, J.; Henry, R.; Spanton, S.; Dziki, W.; Porter, W.; Quick, J.; Bauer, P.; Donaubauer, J.; Narayanan, B. A.; Soldani, M.; Riley, D.; McFarland, K. *Org. Process Res. Dev.* **2000**, *4*, 413–417.

(2) Morissette, S. L.; Almarsson, Ö.; Peterson, M. L.; Remenar, J. F.; Read, M. J.; Lemmo, A. V.; Ellis, S.; Cima, M. J.; Gardner, C. R. *Adv. Drug Delivery Rev.* **2004**, *56*, 275–300.

(3) Day, G. M.; Motherwell, W. D. S.; Ammon, H. L.; Boerrigter, S. X. M.; Della Valle, R. G.; Venuti, E.; Dzyabchenko, A.; Dunitz, J. D.; Schweizer, B.; van Eijck, B. P.; Erk, P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Leusen, F. J. J.; Liang, C.; Pantelides, C. C.; Karamertzanis, P. G.; Price, S. L.; Lewis, T. C.; Nowell, H.; Torrisi, A.; Scheraga, H. A.; Arnautova, Y. A.; Schmidt, M. U.; Verwer, P. *Acta Crystallogr. B* **2005**, *61*, 511–527.

(4) Day, G. M.; Chisholm, J.; Shan, N.; Motherwell, W. D. S.; Jones, W. *Cryst. Growth Des.* **2004**, *4*, 1327–1340.

(5) Day, G. M.; Motherwell, W. D. S.; Jones, W. *Cryst. Growth Des.* **2005**, *5*, 1023–1033.

(6) Görbitz, C. H. *J. Mol. Struct.-THEOCHEM* **2006**, *775*, 9–17.

(7) Allen, F. H. *Acta Crystallogr. B* **2002**, *58*, 380–388.

(8) Spackman, M. A.; Munshi, P.; Dittrich, B. *ChemPhysChem* **2007**, *8*, 2051–2063.

(9) Banks, J. L.; Kaminski, G. A.; Zhou, R. H.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **1999**, *110*, 741–754.

(10) Patel, S.; Brooks, C. L., III *J. Comput. Chem.* **2004**, *25*, 1–15.

(11) Applequist, J. *J. Chem. Phys.* **1985**, *83*, 809–826.

(12) Stone, A. J. *Chem. Phys. Lett.* **1989**, *155*, 102–110.

(13) Misquitta, A. J.; Stone, A. J.; Price, S. L. *J. Chem. Theory Comput.* **2008**, *4*, 19–32.

(14) Misquitta, A. J.; Stone, A. J. *J. Chem. Theory Comput.* **2008**, *4*, 7–18.

(15) Welch, G. W. A.; Karamertzanis, P. G.; Misquitta, A. J.; Stone,

A. J.; Price, S. L. *J. Chem. Theory Comput.* **2008**, *4*, 522–532.

(16) Cossi, M.; Barone, V.; Mennucci, B.; Tomasi, J. *Chem. Phys. Lett.* **1998**, *286*, 253.

(17) Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *106*, 5151.

(18) Cossi, M.; Scalmani, G.; Rega, N.; Barone, V. *J. Chem. Phys.* **2002**, *117*, 43–45.

(19) Cooper, T. G.; Jones, W.; Motherwell, W. D. S.; Day, G. M. *CrystEngComm* **2007**, *9*, 595–602.

(20) Mallikarjunan, M.; Rao, S. T. *Acta Crystallogr. B* **1969**, *25*, 296.

(21) Dalhus, B.; Görbitz, C. H. *Acta Crystallogr. C* **1996**, *52*, 1759.

(22) Dalhus, B.; Görbitz, C. H. *Acta Chem. Scand.* **1996**, *50*, 544.

(23) Day, G. M.; Motherwell, W. D. S.; Jones, W. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1693–1704.

(24) Delley, B. *J. Chem. Phys.* **1990**, *92*, 508–517.

(25) *MS Modelling*, *release 3.0.1*; Accelrys Inc.: San Diego, U.S.A., 2004.

(26) Karamertzanis, P. G.; Pantelides, C. C. *J. Comput. Chem.* **2004**, *26*, 304–324.

(27) Cambridge Structural. Database - Space Group Statistics Web page. http://www.ccdc.cam.ac.uk/products/csd/statistics/space_group_stats.php4 (accessed January 3, 2008).

(28) Williams, D. E.; Cox, S. R. *Acta Crystallogr. B* **1984**, *40*, 404–417.

(29) Cox, S. R.; Hsu, L.-Y.; Williams, D. E. *Acta Crystallogr. A* **1981**, *37*, 293–301.

(30) Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M. *J. Phys. Chem.* **1996**, *100*, 7352–7360.

(31) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitar, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. W.; 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.; Ayalla, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian03*; Gaussian Inc.: Wallingford, CT, 2004.

(32) Stone, A. J. GDMA: Distributed Multipole Analysis of Gaussian Wavefunctions, version 2.2.; University of Cambridge: 2005.

(33) Stone, A. J. *Chem. Phys. Lett.* **1981**, *83*, 233–239.

(34) Stone, A. J.; Alderton, M. *Mol. Phys.* **1985**, *56*, 1047–1064.

(35) Price, S. L.; Willock, D. J.; Leslie, M.; Day, G. M. DMAREL, version 4.1.1; 2001.

(36) Dauber-Osguthorpe, P.; Roberts, V. A.; Osguthrope, D. J.; Wolff, J.; Genest, M.; Hagler, A. T. *Proteins* **1988**, *4*, 31–47.

(37) Sun, H. *J. Phys. Chem.* **1998**, *102*, 7338.

(38) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. *J. Phys. Chem.* **1990**, *94*, 8897–8909.

(39) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

(40) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219.

(41) Chisholm, J.; Motherwell, S. *J. Appl. Crystallogr.* **2005**, *38*, 228–231.

(42) Kozhin, V. *Kristallografiya* **1978**, *23*, 1211–1215.

(43) Flaig, R. Ph.D. Thesis, Freie Universität Berlin, Berlin, 2000.

(44) *Handbook of Chemistry and Physics*, 72nd ed.; CRC Press: Boston, MA, 1992; pp 12−38.

(45) Misoguti, L.; Varela, A. T.; Nunes, F. D.; Bagnato, V. S.; Melo, F. E. A.; Mendes Filho, J.; Zilio, S. C. *Opt. Mater.* **1996**, *6*, 147–152.

(46) Rodrigues Jr., J. J.; Misoguti, L.; Nunes, F. D.; Mendonça, C. R.; Zilio, S. C. *Opt. Mater* **2003**, *22*, 235–240.

(47) Tulip, P. R.; Clark, S. J. *Phys. Rev. B* **2006**, *74*, 064301.

CT800195G